PROCEEDINGS – VOLUME I

Bringing together minds, ideas, and innovation!



ICCE 2025

33rd International Conference
on Computers in Education
01-05 December 2025

IIT Madras

© 2025 Asia-Pacific Society for Computers in Education. All rights reserved.

ORGANIZED BY:

APSCE

HOSTED BY:

INDIAN INSTITUTE OF TECHNOLOGY MADRAS
सिद्धिर्भवति कर्मजा

SPONSORED BY:

NPTEL

Persistent

CODE
IIT MADRAS

© 2025 Asia-Pacific Societ for Computers in Education. All rights reserved.
ISSN: 3078-436
ISBN: 9786269689064

Editors:

Bo JIANG, Yanjie SONG, Maria Mercedes RODRIGO, Jayakrishnan WARRIEM

Publisher:

Asia-Pacific Society for Computers in Education (APSCE)
Research Center for Science and Technology for Learning,
National Central University, No. 300, Zhongda Rd., Zhongli District,
Taoyuan City 320317, Taiwan (R.O.C.)
Telephone: +886-3-4227151 ext. 35407 | Fax: +886-3-4227151 ext. 35407
Email: service@apsce.net | Website: http://www.apsce.net

Date of Publication
December 2025

Copyright Notice
All rights reserved. No part of this book may be reproduced, stored in a retrieval system, transmitted, in any forms or any means, without the prior permission of the Asia-Pacific Society for Computers in Education. Individual papers may be uploaded on to institutional repositories or other academic sites for self-archival purposes.

Editors:

Bo JIANG, East China Normal University, China

Yanjie SONG, The Education University Hong Kong, Hong Kong SAR

Maria Mercedes RODRIGO, Ateneo de Manila University, Philippines

Jayakrishnan WARRIEM, Indian Institute of Technology Madras, India

Associate Editors: (in alphabetical order)

Bo ZHANG, Xi'an Jiaotong-Liverpool University, China

Lenka SCHNAUBERT, University of Nottingham, UK

Ashwin Tudur SADASHIVA, Vanderbilt University, USA

Ivica BOTIČKI, University of Zagreb, Croatia

Chih-Pu DAI, University of Hawaii, Manoa, USA

Barry Lee REYNOLDS, University of Macau, Macau SAR, China

Navneet KAUR, UNESCO, MGIP, India



Message from the *Conference Chair*



Maria Mercedes T. RODRIGO
Conference Chair
Ateneo de Manila University, Philippines

On behalf of the organizing committee, I welcome all participants of the 33rd International Conference on Computers in Education (ICCE) 2025, the flagship conference of the Asia-Pacific Society for Computers in Education (APSCE)!

As part of our program, we welcome four excellent scholars as our keynote speakers: ***Minhong (Maggie) Wang*** is a Professor at the University of Hong Kong and Director of the Laboratory for Knowledge Management and E-Learning. She challenges us to rethink how people learn in the age of AI. ***Manu Kapur,*** Director of the Singapore ETH Center and Professor of Learning Sciences and Higher Education at ETH Zurich will speak about the designing for productive failure leading to deep learning. John Stamper, Associate Professor at the Human-Computer Interaction Institute of Carnegie Mellon University will discuss the growth and use of data over the past 20 years and explore the use of data in today's world of LLMs. Finally, ***Andrew Thangaraj***, Professor of the Department of Electrical Engineering of IIT Madras will focus on the interplay among complex systems such as technology, academic institutions, students, government and others, and how the tensions between practice and research can be managed.

Joining our keynote speakers are three theme-based speakers. ***Martina Holenko Dlab*** is an Associate Professor and Head of the Chair of Multimedia Systems and E-Learning at the Faculty of Informatics and Digital Technologies, University of Rijeka, Croatia. She will speak about the aligning educational technologies with learning objectives through learner-centered design. ***Jie Chi Yang*** is Distinguished Professor in the Graduate Institute of Network Learning Technology at National Central University (NCU), Taiwan and a former program chair of ICCE. He will discuss digital game-based learning and its relationship with learner differences. Our third theme-based speaker is ***Wen Yun***, an Assistant Professor in the Learning Sciences and Assessment Department at the National Institute of Education (NIE), Nanyang Technological University, Singapore (NTU). She will be sharing her work about learner agency and second language learning.

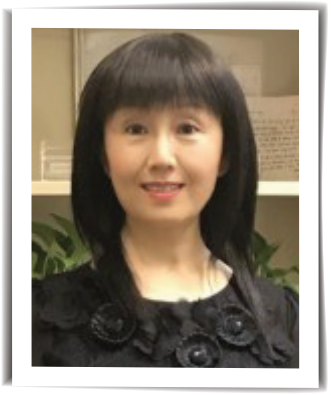
For making ICCE 2025 possible, I thank the organizing committee for all their hard work. I thank the International Program Coordination Chair and Co-Chair, Bo Jiang and Yanjie Song, respectively, our Local Organizing Committee Chair, Jayakrishnan Warriem and his team, and our consultants, Akihiro Kashihara, Jessica Sugay, and Lung-Hsiang Wong. I also thank the many sub-conference chairs, program committee members, organizers of the Workshops, Tutorials, Work-In-Progress Posters (WIPP), Doctoral Student Consortium (DSC), Posters, Early Career Workshops (ECW), and Extended Summaries (ES). I am grateful to all the paper authors and registered participants for joining us this year.

I must, of course, also thank our Managing Secretary of APSCE Pham-Duc Tho for his constant support, the standing committee, and the Executive Committee of APSCE.

I wish all the participants have fruitful and engaging conference! Thank you and have a great conference!

Massage from the *International Program Coordination Chairs*





Bo JIANG
East China Normal University, China

Yanjie SONG
The Education University of HongKong

Welcome to the 33rd International Conference on Computers in Education (ICCE 2025) organized by the Asia-Pacific Society for Computers in Education (APSCE)! ICCE is an annual conference series addressing a broad range of issues related to using Information and Communication Technology (ICT) for education and learning. ICCE2025 takes place in Chennai, India from December 01-05, 2025. The main purpose is to provide researchers from all over the world with the opportunities to share research and new ideas for building the future of the field of Computers in Education.

ICCE 2025 continues the meta-conference tradition of the previous ICCEs. The conference is organized into seven Sub-Conferences specializing specific themes. Each Sub-Conference is organized by a program committee appointed by the respective

Special Interest Group (SIG: refer to https://apsce.net/special-interest-groups).
These Sub-Conferences are:

- C1: ICCE Sub-Conference on Artificial Intelligence in Education/Intelligent Tutoring System (AIED/ITS) and Adaptive Learning
- C2: ICCE Sub-Conference on Computer Supported Collaborative Learning and Learning Sciences (CSCL/LS)
- C3: ICCE Sub-Conference on Advanced Learning Technologies (ALT), Learning Analytics and Digital Infrastructure
- C4: ICCE Sub-Conference on Technology Enhanced Learning for Mobility of Learners and Learning Experiences (TEML)
- C5: ICCE Sub-Conference on Educational Gamification and Game-based Learning (EGG)
- C6: ICCE Sub-Conference on Technology Enhanced Language Learning (TELL)
- C7: ICCE Sub-Conference on Practice-driven Research, Teacher Professional Development and Policy of ICT in Education (PTP)

The International Program Committee is led by a strong and dedicated team, which includes the Conference Chair, the Program Coordination Chair and Co-Chair, Sub-Conference Chairs and Co-Chairs and experts in the field of Computers in Education from many different countries or economies. Former ICCE local organizing and program coordination chairs have played important roles as consultants in overseeing the organization process of this conference. ICCE2025 main conference received a total of 165 papers (141 full, 12 short, and 12 posters) from 23 different countries or economies. Table 1 provides the submissions by the country/area of the first author in each paper. Top five countries/areas with the highest number of submissions were India, Japan, India, China Mainland, Taiwan, and United States of America.

Table 1 Paper Submission Statistics (based on the first author's country/ arear)

Countries/Economies					
India	64	Malaysia	2	Hong Kong	1
Japan	33	Singapore	1	Vietnam	1
Taiwan	16	Thailand	1	France	1
China	13	New Zealand	1	Kuwait	1
United States	12	Croatia	1	Mongolia	1
Philippines	5	Australia	1	UK	1
Indonesia	3	Germany	1	Canada	1
Poland	3	Netherlands	1		

All papers were subjected to a rigorous review process by at least three reviewers from the respective Sub-Conference program committees. After the reviews were completed, a metareview was provided for each paper. In total, 526 reviews and meta-reviews were received. After a discussion period within the individual program committees led by the Sub-Conference Executive Chairs and Co-Chairs, recommendations were made to the Program Coordination Chair and Co-Chair, who oversaw the review process and quality for all Sub-Conferences. This resulted in 36 full papers, 65 short papers, and 28 posters accepted across the seven Sub-Conferences. The overall acceptance rate for full papers was 25.53%, which reflects our efforts to continue the maintenance of the quality of presentations at ICCE 2025. The complete statistics of paper acceptance is shown in Table 2.

Table 2 Paper Acceptance Statistics

Sub-Conference	C1: AIED/ITS	C2: CSCL	C3: ALT	C4: TEML	C5: EGG	C6: TELL	C7: PTP	Total
Total submissions	57	15	48	7	9	8	21	**165**
Submitted as Full Paper	45	12	43	6	8	8	19	**141**
25% of Submitted as Full Paper	11.25	3.00	10.75	1.50	2.00	2.00	4.75	**35.25**
Accepted as Full Paper	10	3	12	1	3	2	5	**36**
Full Paper %	22.22%	25.00%	27.91%	16.67%	37.50%	25.00%	26.32%	**25.53%**
Accepted as Short Paper	23	4	15	4	3	5	11	**65**
Accepted as Poster	9	4	8	0	2	1	4	**28**
Rejected	15	4	13	2	1	0	1	**36**
Overall %	74%	73%	73%	71%	89%	100%	95%	**78.18%**

In addition to full papers, short papers and posters, ICCE 2025 includes various program components, such as Keynote Speeches, Theme-based Invited Speeches, Workshops, Tutorials, Interactive Events, Panels, Work-in-Progress Posters (WIPP), Extended Summary (ES), Showcase of Advancements in Technology-Enhanced Learning in Underrepresented Countries (SATELUC), Doctoral Student Consortia (DSC), and Early Career Workshop (ECW). All the papers in these program components are published in separate proceedings with their own ISBN numbers. Pre-conference events are held on the first two days of the conference, including 8 workshops, DSC, ECW, and APSCE Student Wing Workshop.

We are grateful to all who contributed to making ICCE 2025 a successful conference. First, we thank all the paper authors for their contributions and for choosing ICCE 2025 as a venue to present their research. We would also like to thank the IPC Executive Chairs/Co-Chairs and members, who undertook the responsibility of reviewing and selecting papers that represent research of high quality. Specially thanks to our Keynote and Invited Speakers for accepting our invitations and bringing inspiring research to the ICCE 2025 participants. The Local Organizing Committee deserves a big thank you for their hard work.

We hope that all participants will find ICCE 2025 interesting and inspiring, and that they will enjoy not only academic activities but also exciting cultural experiences in Chennai, India.



Message from the *Local Organizing Chair*



Jayakrishnan WARRIEM
Local Organizing Chair
Indian Institute of Technology Madras, India

Welcome to Chennai!

On behalf of the Indian Institute of Technology Madras (IIT Madras) and the Centre for Outreach and Digital Education (CODE), it is our privilege and great pleasure to host ICCE 2025 and welcome the wider ICCE community to our vibrant campus.

We extend our sincere gratitude to the Asia Pacific Society for Computers in Education (APSCE) for entrusting us with this responsibility. This conference is the result of tremendous collaboration and dedication. We thank our Conference Chair, Maria Mercedes RODRIGO, IPC Chair, Bo JIANG, IPC Co-Chair, Yanjie SONG, and the members of the Local Organizing Committee Core Team — Prof. Vignesh Muthivijayan (Finance), Mr. Sajjan DIXIT (Registration & Website Management), Vinoth and Alfred (Logistics), and Ramesh (Creatives & Publicity).

We are especially grateful to our energetic LOC Sub-Core and volunteer pool comprising PhD and MTech scholars from the Centre for Educational Technology at IIT Bombay and undergraduate students from the IIT Madras BS Program. Their enthusiasm and commitment have powered much of the on-ground experience you will witness over the next few days.

Our heartfelt thanks to our sponsors — Persistent Systems Limited, the National Program for Technology Enhanced Learning (NPTEL), and the IIT Madras BS Program — for their generous support in making this conference possible.

Most importantly, we thank the 150+ conference participants for your contributions, your research, and your presence here. We hope ICCE 2025 enriches you with new perspectives, collaborations, and friendships.

While you are here, we invite you to experience Chennai's unique charm — its heritage, culture, cuisine, and warm hospitality. We hope you will find time between sessions to explore the city's temples, beaches, food streets, and markets.

Once again, Chennai extends our Vanakkam and welcome to ICCE 2025 at IIT Madras. We wish you an inspiring and memorable conference experience!

Organization

—STANDING COMMITTEE—

Conference Chair

Maria Mercedes RODRIGO, Ateneo de Manila University, Philippines

International Program Coordination Chair

Bo JIANG, East China Normal University, China

International Program Coordination Co-Chair

Yanjie SONG, The Education University Hong Kong, Hong Kong SAR

Local Organizing Committee Chair

Jayakrishnan WARRIEM, Indian Institute of Technology Madras, India

Consultants

Akihiro KASHIHARA, The University of Electro-Communications, Japan
Lung-Hsiang WONG, Nanyang Technological University, Singapore
Jessica O. SUGAY, Ateneo de Manila University, Philippines

—SUB-CONFERENCES—

—STANDING COMMITTEE—

C1: ICCE Sub-Conference on Artificial Intelligence in Education/Intelligent Tutoring System (AIED/ITS) and Adaptive Learning

PC Executive Chair

Bo ZHANG, Xi'an Jiaotong Liverpool University, China

PC Co-chair

Anwar P.P. ABDUL MAJEED, Sunway University, Malaysia
Kwok Shing Gary CHENG, The Education University of Hong Kong, Hong Kong China
Tünde VARGA-ATKINS, University of Liverpool, UK
Xiaoming ZHAI, University of Georgia, USA

C2: ICCE Sub-Conference on Computer-Supported Collaborative Learning / Learning Sciences (CSCL/LS)

PC Executive Chair

Lenka SCHNAUBERT, University of Nottingham, UK

PC Co-chair

Gaoxia ZHU, Nanyang Technological University, Singapore
Daniel BODEMER, University of Duisburg-Essen, Germany
Camilla MATUK, New York University, USA

C3: ICCE Sub-Conference on Advanced Learning Technologies, Learning Analytics, Platforms and Infrastructure (ALT/LA/P/I)

PC Executive Chair

Ashwin Tudur SADASHIVA, Vanderbilt University, USA

PC Co-chair

A.Y.M. Atiquil ISLAM, East China Normal University, China
Manjunath K V, IIIT Dharwad, India
Anand Kumar M, NITK Surathkal, India

—SUB-CONFERENCES—

—STANDING COMMITTEE—

C4: ICCE Sub-Conference on Technology Enhanced Learning for Mobility of Learners and Learning Experiences (TEML)

PC Executive Chair

Ivica BOTICKI, Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia

PC Co-chair

Daner SUN, Education University of Hong Kong, Hong Kong SAR, China

Martina Holenko DLAB, Faculty of Informatics and Digital Technologies, University of Rijeka, Croatia

Changhao LIANG, Kyoto University, Japan

C5: ICCE Sub-Conference on Educational Gamification and Game-based Learning (EGG)

PC Executive Chair

Barry Lee REYNOLDS, University of Macau, Macau SAR, China

PC Co-chair

Ai-Chu Elisha DING, University of Georgia, USA

Louise Antonette DELAS PENAS, Ateneo de Manila University, the Philippines

Zhichun LIU, The University of Hong Kong, Hong Kong SAR, China

C6: ICCE Sub-Conference on Technology Enhanced Language Learning (TELL)

PC Executive Chair

Chih-Pu DAI, University of Hawaii, Manoa, USA

PC Co-chair

Boris VAZQUEZ-CAVO, Universidad de Málaga, Spain

Jing LEI, Syracuse University, USA

Ethel ONG, De La Salle University, Philippines

—SUB-CONFERENCES—

—STANDING COMMITTEE—

C7: ICCE Sub-Conference on Practice-driven Research, Teacher Professional Development and Policy of ICT in Education (PTP)

PC Executive Chair

Navneet KAUR, MGIEP, UNESCO, India

PC Co-chair

Lucian NGEZE, University of Dodoma, Tanzania

Prajakt PANDE, Southern Methodist University, USA

Shashi Kant SHANKAR, Ahmedabad University, India

—OTHER COMPONENTS—

Workshop/Tutorial/Interactive Event

Chair

Ali DEWAN, Athabasca University, Canada

Co-chair

Fuzheng ZHAO, Jilin University, China

Work-In-Progress Posters (WIPP)

Chair

Ming GAO, Shanghai Normal University, China

Co-chair

Chiu-Lin LAI, National Taipei University of Education, Taiwan

Yun WEN, Nanyang Technological University, Singapore

Doctoral Student Consortium (DSC)

Chair

Anveshna SRIVASTAVA, IIT Kanpur, India

Co-chair

Maiga CHANG, Athabasca University, Canada

Thomas KIN-FUNG CHIU, Chinese University of Hong Kong, Hong Kong SAR

Xuesong ZHAI, Zhejiang University, China

Lin FENG, Singapore University of Social Sciences, Singapore

Early Career Workshop (ECW)

Chair

Ramkumar RAJENDRAN, IIT Bombay, India

Co-chair

May Marie TALANDRON-FELIPE, University of Science and Technology of Southern Philippines

Ming GAO, Shanghai Normal University, China

Extended Summary (ES)

Chair

Emmanuel AYEDOUN, Kansai University, Japan

Co-chair

HongChao PENG, East China Normal University, China

Banni Satria ANDOKO, State Polytechnic of Malang, Indonesia

Advisor

Kae NAKAYA, Osaka University, Japan

APSCE Excellence Scholarship Award

Chair

Alwyn Vwen YEN LEE, Nanyang Technological University, Singapore

Co-chair

Mas Nida MD KHAMBARI, Universiti Putra Malaysia

Consultant

Su Luan WONG, Universiti Putra Malaysia

Technical Advisor

Nur Aira ABD RAHIM, Universiti Putra Malaysia

Showcase of Advancements in Technology-Enhanced Learning in Underrepresented Countries (SALUTEC)

Chair

Mas Nida KHAMBARI, Universiti Putra Malaysia, Malaysia

Co-chair

Nur Aira ABDUL RAHIM, Universiti Putra Malaysia, Malaysia

Ivica BOTICKI, University of Zagreb, Croatia

Saida ULFA, State University of Malang, Indonesia

—OTHER COMPONENTS—

Panels

Chair

Yanjie Song, The Education University of Hong Kong, HongKong SAR

Co-chair

Morris Jong, The Chinese University of Hong Kong, HongKong SAR

LOCAL ORGANIZING COMMITTEE

Chair

Jayakrishnan WARRIEM

Finance Chair

Vignesh MUTHUVIJAYAN, Indian Institute of Technology Madras, Chennai, India

Registration and Website Management

Sajjan DIXIT, Indian Institute of Technology Madras, Chennai, India

Logistics and Events

Vinoth Sakkarai PANDIYAN

Alfred Jeevan WILSON

Surya PRAKASH

Technology and Media

Ramesh RAMADASS

Pradeepa S

Volunteers

Antony PRAKASH

Aditya RAJMANE

Samarth KASHYAP

Nisumba SOODHANI

Spruha SATAVLEKAR

Setu MAHESHWARI

Abhishek KUMAR RAI

Pratiksha VIRENDRA PATIL

Jyoti SHAHA

Amit PAIKRAO

Volunteers

Meera PAWAR

Sunny PRAJAPATI

Jatin AMBASANA

Sandeep YADAV

Zun PHOO MO

Bathula SHARATH KUMAR

Sivaraman S

Arjun PRASAD

Vikram VINCENT

SPECIAL INTEREST GROUPS (SIG) 2024–2025

SIG 1: Artificial Intelligence in Education/Intelligent Tutoring Systems/Adaptive Learning (AIED/ITS/AL)
Bo JIANG, East China Normal University, China

SIG 2: Computer-supported Collaborative Learning and Learning Sciences (CSCL)
Lenka SCHNAUBERT, University of Nottingham, UK

SIG 3: Advanced Learning Technologies, Platforms and Infrastructure (ALT)

SIG 4: Technology Enhanced Learning for Mobility of Learners and Learning Experiences (TEML) (formerly CUMTEL: Classroom, Ubiquitous and Mobile Technologies Enhanced Learning)
Ivica BOTIČKI, University of Zagreb, Croatia

SIG 5: Educational Gamification and Game-based Learning (EGG)
Jewoong MOON, University of Alabama, USA

SIG 6: Technology Enhanced Language Learning (TELL)
Yanjie SONG, Education University of Hong Kong, Hong Kong

SIG 7: Practice-driven Research, Teacher Professional Development and Policy of ICT in Education (PTP)
Jayakrishnan Madathil WARRIEM, Indian Institute of Technology of Madras, India

SIG 8: Development of Information and Communication Technology in the Asia Pacific Neighborhood (DICTAP)
Patcharin PANJABUREE, Mahidol University, Thailand

SIG 9: Educational Use of Problems/Questions in Technology-Enhanced Learning (EUPQ)
Shitanshu MISHRA, UNESCO MGIEP, India

SIG 10: Learning Analytics and Educational Data Mining (LAEDM)
Ashwin T.S, Vanderbilt University, USA

SIG 11: Computational Thinking Education & STEM Education (CTE&STEM)
Ting-Chia HSU, National Taiwan University, Taiwan

C1; AIED/ITS PC MEMBERS

Abdul Nasir BIN ABD GHAFAR, Universiti Malaysia Pahang Al Sultan Abdullah, Malaysia

Aisha ABUELMAATTI, Queen Mary University, UK

Arne BEWERSDORFF, Technische Universität München, Germany

Chien-Sing LEE, Sunway University, Malaysia

Chih-Yueh CHOU, Yuan Ze University, Taiwan

Chun-Yen CHANG, Taiwan Normal University, Taiwan

Danish Mahmood KHAN, Sunway University, Malaysia

Ehsan LATIF, University of Georgia, United States of America

Farrukh HASSAN, Sunway University, Malaysia

Ganesh KUMAR, Sunway University, Malaysia

Hafiz Rashidi RAMLI, Universiti Putra Malaysia, Malaysia

Harriz BIN ZAINAL ABIDIN, International Islamic University Malaysia, Malaysia

Hassan ABUHASSNA, Sunway University, Malaysia

Jiawei HAI, Xi'an Jiaotong Liverpool University, China

Jiyao XUN, Xi'an Jiaotong Liverpool University, China

Jose-Luis Perez DE LA CRUZ, Universidad de Malaga, Spain

Kai-Yu TANG, National Chung Hsing University, Taiwan

Katie ATKINSON, University of Liverpool, UK

Lap Kei Keith LEE, Hong Kong Metropolitan University, Hong Kong China

Manoj SINHA, University of Delhi, India

Marcelo MILRAD, Linnaeus University, Sweden

Matt GLOWATZ,University College Dublin,Ireland

Mélina VERGER, Sorbonne Universite, France

Mohamad Hardyman BARAWI, Universiti Malaysia Sarawak, Malaysia

Muhammed Basheer JASSER, Sunway University, Malaysia

Nejla YÜRÜK, Gazi Eğitim Fakültesi, Turkey

Nurul Amelina NASHARUDDIN, Universiti Putra Malaysia, Malaysia

C1; AIED/ITS PC MEMBERS

Robbi RAHIM, Sekolah Tinggi Ilmu Manajemen Sukma, Indonesia

Ryan BAKER, University of Pennsylvania, United States of America

Saad ASLAM, Sunway University, Malaysia

Sein MINN, [affiliation tbc], France

Shawon SARKAR, University of Washington, United States of America

Shuchen GUO, Nanjing Normal University, China

Shukor Sanim MOHD FAUZI, Universiti Teknologi MARA, Malaysia

Si Na KEW, Universiti Teknologi Malaysia, Malaysia

Tanja MITROVIC, University of Canterbury, New Zealand

Tao WANG, University of North Dakota, United States of America

Teck Chai LAU, Xi'an Jiaotong Liverpool University, China

Thoo Ai CHIN,Universiti Teknologi Malaysia, Malaysia

Tong WU, Xi'an Jiaotong-Liverpool University, China

Tsz Kit Davy NG, The Education University of Hong Kong, Hong Kong China

Vimal Rau APAROW, University of Nottingham Malaysia, Malaysia

Viren AUL, Xi'an Jiaotong-Liverpool University, China

Winn CHOW, The University of Melbourne, Australia

Xieling CHEN, Guangzhou University, China

Xue ZHOU, University of Leicester, UK

Yingtuan LO, Xi'an Jiaotong Liverpool University, China

Yizhu GAO, University of Georgia, United States of America

Young Hoan CHO, Seoul National University, South Korea

Yue YIN, University of Illinois at Chicago, United States of America

C2: CSCL / LS PC MEMBERS

Aditi KOTHIYAL, IIT Gandhinagar, India

Rose LIANG, National University of Singapore, Singapore

Young Hoan CHO, Seoul National University, South Korea

Elizabeth KOH, National Institute of Education, Nanyang Technological University, Singapore

Haipeng WAN, College of Education, China

Yuyao TONG, Guangzhou University, China

Yan ZHOU, Zhejiang Normail University, China

Pei-Yi LIN, National Tsing-Hua University, Taiwan

Sanjay CHANDRASEKHARAN, Homi Bhabha Centre for Science Education (HBCSE), India

Ben CHANG, National Central University, Taiwan

Heisawn JEONG, Hallym University, South Korea

Alwyn Vwen Yen LEE, Nanyang Technological University, Singapore

Aloysius ONG, National Institute of Education, Singapore

Seng Chee TAN, Nanyang Technological University, Singapore

Shikha GUPTA, Shaheed Sukhdev College of Business Studies, India

C3: ALT/LA/P/I PC MEMBERS

Mohamed Elsayed AHMED, South Valley University

Radhika AMASHI, BVBCET

Banni Satria ANDOKO, State Polytechnic of Malang

Vishwas BADHE, Indian Institute of Technology, Bombay, India

Abhilash C B, JSSATE-Bangalore

Weiqin CHEN, Oslo Metropolitan University

Clayton COHN, Vanderbilt University

Syaamantak DAS, Indian Institute of Technology Bombay

Yoshiko GODA, Kumamoto University

C3: ALT/LA/P/I PC MEMBERS

Prakash HEGADE, KLE Technological University

Shruti JAIN, Indian Institute of Technology Bombay

Marc JANSEN, University of Applied Sciences Ruhr West

Kapil KADAM, Indian Institute of Technology Bombay

Kabyashree KHANIKAR, Indian Institute of Technology Bombay

Nattapol KRITSUTHIKUL, National Electronics and Computer Technology Center (NECTEC), Thailand

Ramesh KURBET, KLE technological University

Huiyong LI, Kyushu University, Japan

Ruixue LIU, Northwest Normal University

Sonsoles LÓPEZ-PERNAS, University of Eastern Finland

Vijayalakshmi M, KLE technological University

Kaushik MALLIBHAT, KLE technological University

Ritayan MITRA, Indian Institute of Technology Bombay

Indrayani NISHANE, Indian Institute of Technology Bombay

Amit PAIKRAO, Indian institute of Technology Bombay

Sonika PAL, Indian Institute of Technology Bombay

Rumana PATHAN, Indian Institute of Technology Bombay

Preeti PATIL, KLE Technological University

Herold PC, Indian Institute of Technology Bombay

Antony PRAKASH, Indian Institute of Technology Bombay

Ramkumar RAJENDRAN, Indian Institute of Technology Bomaby

Shridhar SANSHI, NITK Surathkal

Jyoti SHAHA, Indian Institute of Technology Bombay

Pritam SIL, Indian Institute of Technology Bombay

Anuradha SINGH, Banaras Hindu University

Bam Bahadur SINHA, National Institute of Technology Sikkim, India

Caitlin SNYDER, University of Detroit Mercy

Namrata SRIVASTAVA, Vanderbilt University

C3: ALT/LA/P/I PC MEMBERS

Christine Lourrine TABLATIN, Ateneo de Manila University

Tuan TRAN, Royal Melbourne Institute of Technology

Riina VUORIKARI, Institute for Prospective Technological Studies (IPTS)

Didin WAHYUDIN, Departemen Pendidikan Teknik Elektro

Saumya YADAV, IIIT-Delhi, Delhi

Xiaotian ZOU, Indiana University

Pankaj CHAVAN, Indian Institute of Technology Jodhpur

Ashutosh RAINA, Indian Institute of Technology Bombay

C4: TEML PC MEMBERS

Patcharin PANJABUREE, Khon Kaen University, Thailand

Peter SEOW, Nanyang Technological University, Singapore

Rwitajit MAJUMDAR, Kumamoto University, Japan

Emmanuel AYEDOUN, Kansai University, Japan

Danijel MLINARIC, University of Zagreb, Croatia

Nicole, Yin YANG, The Education University of Hong Kong, Hong Kong SAR, China

Manpreet SINGH, The Education University of Hong Kong, Hong Kong SAR, China

Leisi PEI, The Education University of Hong Kong, Hong Kong SAR, China

Xiaojing WENG, The Education University of Hong Kong, Hong Kong SAR, China

Dan SUN, Hangzhou Normal University, China

Maja GLIGORA MARKOVIC, University of Rijeka, Croatia

Kristian STANCIN, University of Rijeka, Croatia

Gordan DUROVIC, University of Rijeka, Croatia

Nehal HASNINE, Hosei University, Japan

Chia-yu HSU, Kyoto University, Japan

Yuko TOYOKAWA, Kyoto University, Japan

Kousuke MOURI, Hiroshima University, Japan

Huiyong LI, Kyushu University, Japan

Pawat CHAIPIDECH, Khon Kaen University, Thailand

C5: EGG PC MEMBERS

Dan SUN, Hangzhou Normal University, China

Dongping ZHENG, University of Hawai i at Mānoa, USA

Haesol BAE, University at Albany – State University of New York, USA

Hui SHI, Florida State University, USA

Jeremy NG, University of Hong Kong, Hong Kong SAR, China

Jessica DUBIOS, Ball State University, USA

Jewoong MOON, University of Alabama, USA

Jiyae BONG, University of North Carolina at Charlotte, USA

Joseph PETERS, University of Hawai i at Mānoa, USA

Larry NGUYEN, University of Hawai i at Mānoa, USA

Liuyufeng LI, University of Hong Kong, Hong Kong SAR, China

Matthew DUVALL, University of Pennsylvania, USA

Meina ZHU, Wayne State University, USA

Mete AKCAOGLU, Georgia Southern University, USA

Nuodi ZHANG, Florida State University, USA

Shiyao WEI, Florida State University, USA

Xiaoxue DU, Massachusetts Institute of Technology, USA

Yanjun PAN, Southern Methodist University, USA

C6: TELL PC MEMBERS

Kae NAKAYA, Osaka University, Japan

Yang GAO, Xi'an Jiaotong University, China

Yunfei DU, Chongqing University of Technology, China

Letizia CINGANOTTO, University for Foreigners of Perugia, Italy

Nina DASKALOVSKA, Goce Delcev University, Republic of North Macedonia

Jo MYNARD, Kanda University of International Studies, Japan

Gilbert DIZON, Himeji Dokkyo University, Japan

C6: TELL PC MEMBERS

Zheng YUAN, King's College London, UK

Laia CANALS, Universitat Oberta de Catalunya, Spain

Antigoni PARMAXI, Cyprus University of Technology, Cyprus

Elena MARTÍN-MONJE, National University of Distance Education, Spain

Chan Chio LAO, University of Macau, Macau SAR, China

Ju Seong LEE, The Education University of Hong Kong, Hong Kong SAR, China

Luyao (Louie) LIANG, Western Sydney University, Australia

Di ZOU, Hong Kong Polytechnic University, Hong Kong SAR, China

Fang HUANG, Shanghai International Studies University, China

Rining WEI, Xi'an Jiaotong-Liverpool University, China

Chen DING, Hefei University of Technology, China

Csaba SZABO, The University of Edinburgh, UK

Carolina BUSTAMANTE, SUNY Old Westbury, USA

Rui LI, Hunan University, China

Yongliang WANG, North China University of Water Resources and Electric Power, China

Artem ZADOROZHNYY, The Education University of Hong Kong, Hong Kong SAR, China

Chian-Wen KAO, Yuan Ze University, Taiwan

Cheng-Ting CHEN, Chung Yuan Christian University, Taiwan

C7: PTP PC MEMBER

Anil MAMMEN, CETE, TISS, India

Arlene VALDERAMA, Jose Rizal University, Mandaluyong City, Philippines

Briju THANKACHAN, EdTech Society, India

DurgaPrasad KARNAM, Illumine Knowledge Resources, India

Ekaterina MURAVEVSKAIA, Indiana University, USA

Herold PC, Indian Institute of Technology Bombay, India

Juhi BANSAL, Ahmedabad University, India

Kapil KADAM, DYPCET, India

C7: PTP PC MEMBER

Mahesha DE SILVA, Tallinn University, Estonia

Manisha KHULBE, Tallinn University, Estonia

Mas-Nida KHAMBARI, Universiti Putra Malaysia, Malaysia

Merike SAAR, Tallinn University, Estonia

Milena SARMIENTO, Tallinn University, Estonia

Nimita PANDEY, Amrita University, India

Prajish PRASAD, FLAME University, India

Pramod GOYAL, Delhi Skill and Entrepreneurship University, India

Priscilla MOSES, Universiti Tunku Abdul Rahman, Malaysia

Rajesh KHAMBAYAT, NITTTR Bhopal, India

Reet KASEPALU, Tallinn University, Estonia

Rex BRINGULA, University of the East, Manila, Philippines

Ruchi KUMAR, CETE, TISS, India

Sadaqat MULLA, CETE, TISS, India

Sambit PRAHARAJ, KIIT, Bhubaneswar, India

Samvet KURIL, Ahmedabad University, India

Sunita RASTE, Indian Institute of Technology Bombay, India

Vikram VINCENT, Freethought Labs, India

Vivek SABANWAR, Texas A&M University, USA

Yannis DIMITRIADIS, Universidad de Valladolid, Spain

Yogendra PAL, ETmantra eLearning Solutions, India

Yogesh VELANKAR, Symbiosis Open Education Society, Pune, India

DISTINGUISHED RESEARCHER AWARD WINNER (2025)



Dr. Ting Chia HSU
National Taiwan Normal University,
Taiwan

Dr. Ting-Chia Hsu is currently a Distinguished Professor at the National Taiwan Normal University. She has long been committed to research in technology-enhanced learning, AI literacy, and computational thinking education, demonstrating both academic rigor and practical impact.

She has received numerous prestigious honors from Taiwan's National Science and Technology Council (NSTC), including the Wu Ta-You Memorial Award, the Outstanding Young Scholar Program, the Distinguished Talent Award, and the International Young Scholar Fellowship. In 2020, her innovative AI2 Robot City board game—which integrates AI-based image recognition—won the NSTC Future Tech Award for its pedagogical innovation and interdisciplinary design. Her original board game, Robot City (focused on computational thinking), was successfully commercialized in 2017 and expanded to the European market in 2024. For these achievements, she received NTNU's Industry-Academia Collaboration Excellence Award in both 2020 and 2023.

In 2024, Professor Hsu served as lead coordinator for the "AI Pedagogy and Teacher Professional Development" section of the draft of the AI Literacy Policy Analysis Book in MOE of Taiwan. In 2025, she is the leader of the "AI literacy framework for K-12 students" division in the revised version of the AI Literacy Policy Analysis Book in MOE of Taiwan. Moreover, she is the editor of the AI education policy white book in Taipei City, Taiwan.

She was honored with the Early Career Researcher Award by the Asia-Pacific Society for Computers in Education (APSCE) in 2018, and later received its highest distinction—the Distinguished Researcher Award in 2025. From 2020 to 2025, she was listed among the world's top 2% most influential scientists by Stanford University, and in 2023 and 2024, ranked fifth in Educational domain in Taiwan. Beyond her teaching experience at senior high schools and universities, Professor Hsu has made systemic contributions across all educational levels in Taiwan. She serves as the consulting professor for the New Taipei City edition of the Kang Hsuan elementary IT textbook, and as the editor for its junior high school counterpart. She also leads the development of digital curricula for the senior high and vocational tracks on Taiwan's Adaptive Learning Platform, including courses on "Information Technology" and "Digital Technology Fundamentals." In higher education, she leads the division of the "Educational Big Data Microcredential Program," which across three periods has involved 13, 14, and 14 universities respectively.
The program supports nationwide data literacy development and facilitates university —government collaboration for learning analytics in K—12 education, promoting data-driven personalized learning in schools.

At the international level, Professor Hsu currently serves as Editor-in-Chief of Research and Practice in Technology Enhanced Learning, a journal indexed in Scopus and ESCI. She has also chaired three major
APSCE Special Interest Groups (TELL, CUMTEL, and CTE-STEM), contributing to the advancement of AI and educational technology research across the Asia-Pacific region. She has held visiting scholar positions at the National Institute of Education (NIE), Nanyang Technological University (Singapore, 2011), MIT Computer Science and Artificial Intelligence Laboratory (USA, 2019), and Vilnius University (Lithuania, 2024). In recent years, she has led the design of innovative instructional modules that combine generative AI, image recognition, and gamification, enhancing students' motivation and AI competencies with strong potential for global diffusion.

LAST TEN YEARS' DISTINGUISHED RESEARCHER AWARD WINNERS

2022 — APSCE Distinguished Researcher Award

Maiga CHANG, Athabasca University, Canada

2021 — APSCE Distinguished Researcher Award

Maria Mercedes T. RODRIGO, Ateneo de Manila University, Philippines

2020 — APSCE Distinguished Researcher Award

Wenli CHEN, Nanyang Technological University, Singapore

2015 — APSCE Distinguished Researcher Award

Lung-Hsiang WONG, Nanyang Technological University, Singapore

2014 — APSCE Distinguished Researcher Award

Hiroaki OGATA, Kyushu University, Japan

EARLY CAREER RESEARCHER AWARD WINNER (2025)



Yun WEN
Nanyang Technological University (NTU), Singapore

Dr Wen Yun is an Assistant Professor in the Learning Sciences and Assessment Department at the National Institute of Education (NIE), Nanyang Technological University, Singapore (NTU). Prior to this, she was a Lecturer at the Singapore Centre for Chinese Language (SCCL), providing in-service training for teachers. She obtained her PhD from NIE/NTU and completed her Postdoc at the Computer-Human Interaction in Learning and Instruction (CHILI) Lab in EPFL, Switzerland.
Positioned at the intersection of empirical inquiry and theoretical advancement, her research focuses on technology-enhanced learning innovations in schools. In particular, she investigates how emerging technologies, such as Artificial Intelligence (AI) and Augmented Reality (AR), can be leveraged to promote collaborative learning in classrooms. She examines how these technologies reshape teaching and learning processes, enhance student engagement, and suggest design principles and teacher support mechanisms for meaningful learning experiences. She currently serves as an Associate Editor for Educational Technology & Society and the Asia Pacific Journal of Education.

A key focus of her research lies in computer-supported collaborative learning for Chinese as a second language. Building on this, she pioneered the use of AR with paper interfaces in classrooms. Her proven track record in school-based innovations also extends to collaborative science inquiry. A key contribution lies in addressing sustainability issues of integrating AR into science inquiry learning for young learners. She leads interdisciplinary teams to translate research into practical educational applications, developing impactful and sustainable technology-enhanced learning solutions.

To date, she has secured 6 competitive research grants totaling approximately USD 1.8 million as Principal Investigator. She currently leads a prestigious project funded by the Social Science & Humanities Research (SSHR) Fellowship (USD 650,676) from the Singapore Social Science Research Council. In recognition of her research achievements, she received the NIE Excellence in Research Award in 2021. She was also honored to receive the Nanyang Research Award (Young Investigator) in 2021, the university's highest accolade for individuals who have made outstanding contributions to advancing the frontiers of knowledge.
Dr Wen has been actively participating in and contributing to APSCE since the early stages of her research career. She served as chair or co-chair of the Technology-Enhanced Language Learning (TELL) sub-conference track and the Work-in-Progress Posters (WIPP) at ICCE. She also has been involved in the organization of the International Conference on AI in Language Learning and Literacy (AI3L), formerly the International Conference on Future Language Learning (ICFULL), since its inception in 2022.

LAST TEN YEARS' EARLY CAREER RESEARCHER AWARD WINNERS

2024 — APSCE Early Career Researcher Award

Dr. Mas Nida Md. KHAMBAI, Universiti Putra Malaysia, Malaysia

2023 — APSCE Early Career Researcher Award

Rwitajit MAJUMDAR, Kyoto University, Japan

2022 — APSCE Early Career Researcher Award

Daner SUN, The Education University of Hong Kong, Hong Kong

2021 — APSCE Early Career Researcher Award

Bo JIANG, East China Normal University, China

2020 — APSCE Early Career Researcher Award

Kaushal Kumar BHAGAT, Indian Institute of Technology, Kharagpur, India

2019 — APSCE Early Career Researcher Award

Chengjiu YIN, Kobe University, Japan

2018 — APSCE Early Career Researcher Award

Ting-Chia HSU, National Taiwan Normal University, Taiwan

2017 — APSCE Early Career Researcher Award

Jon MASON, Charles Darwin University, Australia

2015 — APSCE Early Career Researcher Award

Morris Siu-yung JONG, The Chinese University of Hong Kong, Hong Kong

SPEAKERS OF APSCE WEBINAR SERIES (December 2024 to November 2025)

APSCE Webinar #53: Learning with and about AI in K–12 education

Date : 5 May 2025 (Monday)
Speaker : **Assoc. Prof. Linda Mannila** (University of Helsinki, Finland)
Moderator : **Prof. Ting-Chia HSU** (National Taiwan Normal University, Taiwan), **Prof. Daner Sun** (Educational University of Hong Kong)
Curated by : APSCE Computational Thinking in Education / STEM (CTE/STEM) SIG

APSCE Webinar #54: Learning Analytics and Learning Design as Cornerstones of Evidence-Based Decision-Making in Higher Education

Date : 5 June 2025 (Thursday)
Speaker : **Prof. Blaženka DIVJAK (University of Zagreb, Croatia)**
Moderator : **Prof. Ivica BOTICKI** (University of Zagreb, Croatia)
Curated by : Technology Enhanced Learning for Mobility of Learners and Learning Experiences (TEML) SIG

APSCE Webinar #55: From Player to Creator: Game-Based Learning through Fan Fiction and AI-powered Narratives

Date : 16 July 2025 (Wednesday)
Speaker : **Prof. Sangmin-Michelle LEE** (Kyung Hee University, Korea)
Moderator : **Prof. Jewoong MOON** (The University of Alabama, United States)
Curated by : APSCE Educational Gamification and Game-based Learning (EGG) SIG

APSCE Webinar #56: Repercussions of Generative AI on Teaching and Learning: Insights from ICT Education

Date : 31 July 2025 (Thursday)
Speaker : **Prof. Ritesh CHUGH (Central Queensland University, Australia)**
Moderator : **Dr. Nur Aira Abd Rahim (Universiti Putra Malaysia, Malaysia), Dr. Ryan Ebardo (De La Salle University, Philippines)**
Curated by : APSCE Development of Information and Communication Technology in the Asia-Pacific Neighborhood (DICTAP) SIG

SPEAKERS OF APSCE WEBINAR SERIES (December 2024 to November 2025)

APSCE Webinar #57: Immersive Language Learning with AI: Shared Experiences and Emerging Practices

Date : 15 August 2025 (Friday)

Speaker : **Yun WEN** (Nanyang Technological University, Singapore),
Rustam SHADIEV (Zhejiang University, China),
Yanjie SONG (The Education University of Hong Kong, Hong Kong)

Curated by : APSCE Technology Enhanced Language Learning (TELL) SIG

APSCE Webinar #58: The evolution of computer–supported collaborative learning: Human–to–human vs. human–to–AI

Date : 23 September 2025 (Tuesday)

Speaker : **Elizabeth Koh** (Nanyang Technological University, Singapore),
Young Hoan Cho (Seoul National University, South Korea)

Moderator : **Lenka Schnaubert** (University of Nottingham, United Kingdom),
Alwyn Vwen Yen Lee (Nanyang Technological University, Singapore),
Aditi Kothiyal (Indian Institute of Technology, India)

Curated by : APSCE Computer Supported Collaborative Learning (CSCL) and the Learning Sciences (LS) Special Interest Group (SIG) 2

APSCE Webinar #59: Effects of Instructor Manual Feedback and ChatGPT Intelligent Feedback on Collaborative Programming

Date : 14 October 2025 (Tuesday)

Speaker : **Dr. Fan Ouyang** (Zhejiang University, China)

Moderator : **Dr. Bo Jiang** (East China Normal University, China)

Curated by : APSCE Artificial Intelligence in Education (AIED) / Intelligent Tutoring Systems (ITS) Adaptive Learning (AL) SIG

SPEAKERS OF APSCE WEBINAR SERIES (December 2024 to November 2025)

APSCE Webinar #60: Strategic scaffolding: Data-driven Design for Computational Problem Solving

Date : 28 October 2025 (Tuesday)

Speaker : **Tiffany Barnes** (North Carolina State University, United States)

Moderator : **Ashwin Tudur Sadashiva** (Vanderbilt University, United States)

Curated by : APSCE Learning Analytics and Educational Data Mining (LAEDM) SIG

APSCE Webinar #61: When the Machines Answer Too Quickly: In Praise of the Pause

Date : 7 November 2025 (Friday)

Speaker : **Chris Bigum** (Griffith Institute for Educational Research, Australia)

Moderator : **Shitanshu Mishra** (UNESCO MGIEP, India)

Curated by : APSCE Educational Use of Problems/Questions in Technology-Enhanced Learning (EUPQ) SIG

KEYNOTE SPEAKER

C2: ICCE Sub-Conference on Computer-Supported Collaborative Learning / Learning Sciences (CSCL/LS)



Manu Kapur
Singapore ETH Center

Productive Failure and the Future of EdTech

If learning from failure is intuitively compelling, why do we wait for it to happen. Why don't we intentionally design for it, and harness it for deep learning? In my talk, I will describe my research program on Productive Failure, how, when and why intentionally designing for failure in a safe way can lead to deep learning, and its implications for how we can leverage the latest generative AI technologies to design powerful learning environments for developing conceptual understanding, creativity, and transfer.

Manu is currently the Director of the Singapore-ETH Centre, and a Professor of Learning Sciences and Higher Education at ETH Zurich, Switzerland. With a strong technical background in engineering and statistics as well as doctoral training in the learning sciences, Manu brings a unique interdisciplinary skill set to the study of human learning, both in terms of the fundamental mechanisms of human learning as well as developing applications for translating these mechanisms for teaching and learning. Manu is widely known mainly for his work on learning from Productive Failure, is a sought-after keynote speaker, and has delivered two TEDx talks. His contributions extend across high-profile journals and conferences, influencing educational policies and practices internationally.

KEYNOTE SPEAKER



John Stamper
Carnegie Mellon University

C1: ICCE Sub-Conference on Artificial Intelligence in Education/Intelligent Tutoring System (AIED/ITS) and Adaptive Learning

Are we in a Post-Data Edtech World?
There is no doubt that LLMs have altered the landscape of Artificial intelligence in many fields including education. While over the past 20 years we have seen a huge shift in AI in education towards data informed decision making and model building, which arguably launched or at least greatly enabled the Educational Data Mining and Learning Analytics communities, many of the tasks and models where data was needed can now be “data free" with the use of LLMs.

Few-shot and zero-shot AI has been reinforced with prompting strategies and has some strong results. In this talk, I will discuss the growth and use of data over the past two decades with special focus on my role leading the DataShop repository, and explore the use of data in the world of LLMs.

John Stamper is an Associate Professor at the Human-Computer Interaction institute at Carnegie Mellon University. He is also the Technical Director of the Pittsburgh Science of Learning Center DataShop. His primary areas of research include Educational Data Mining and Intelligent Tutoring Systems. As Technical Director, John oversees the DataShop, which is the largest open data repository of transactional educational data and a set of associated visualization and analysis tools for researchers in the learning sciences. John received his PhD in Information Technology from the University of North Carolina at Charlotte, holds an MBA from the University of Cincinnati, and a BS in Systems Analysis from Miami University. Prior to returning to academia, John spent over ten years in the software industry including working with several start-ups.

KEYNOTE SPEAKER

C7: Practice-driven Research, Teacher Professional Development and Policy of ICT in Education (PTP)

Large Scale Interventions in Higher Education: Lessons learnt from our attempts at reducing Practice–Research gap

The higher education system in India has seen significant expansion in the last two decades with an increased number of institutions and corresponding increases in enrolment. However, there are systemic challenges, including: low employability of graduates, especially for jobs aligned to information technology, and the quality of and access to both content and faculty at postgraduate levels, which are being worked upon by Ministry of Education as part of its National Education Policy 2020 (NEP2020).
This talk will focus on the interventions done at the higher education level, primarily from the point of view of practitioner in the higher education space. I will take the case of the two initiatives taken up by IIT Madras — National Program for Technology Enhanced Learning (NPTEL) and the BS Program in Data Science.

The talk will look at the evolution of the programs from across the years and how the research in education technology is slowly being integrated into the day-to-day practice within them. The talk will focus on the interplay between complex systems (technology, academic institutions, students, government officials, etc) seen in such large initiatives and the lessons that we can take up from them. One of the major learning that I will be detailing is on the need to manage the tension between practice and research which is very critical in such large-scale initiatives and the time required for knowledge from research to seep into day-to-day practice. The second key learning that I would be highlighting will be on the impact of community networks that are grown and nurtured from grounds up as part of these initiatives to sustain the intervention.



Andrew Thangaraj
IIT Madras

Andrew Thangaraj received his B.Tech in Electrical Engineering from the Indian Institute of Technology (IIT), Madras, India in 1998 and a PhD in Electrical Engineering from the Georgia Institute of Technology, Atlanta, USA in 2003. He was a post-doctoral researcher at the GTL-CNRS Telecom lab at Georgia Tech Lorraine, Metz, France from August 2003 to May 2004. From June 2004, he has been with the Department of Electrical Engineering, IIT Madras, where he is currently a professor. Andrew Thangaraj's research interests are in the broad areas of information theory, error-control coding and information-theoretic aspects of learning and cryptography. From Jan 2012 to Jan 2018, he served as Editor for the IEEE Transactions on Communications. From July 2018 to July 2022, he served as an Associate Editor for Coding Techniques for the IEEE Transactions on Information Theory.

Since Oct 2011, he has been serving as NPTEL coordinator at IIT Madras. He has played a key role in initiating and running NPTEL online courses and certification. He is currently the Principal Investigator of the SWAYAM project of the Ministry of Education, Government of India. Since May 2020, he has been serving as Coordinator for the BS (Data Science) Program at IIT Madras. Since May 2024, he has been serving as Chair for the Centre for Outreach and Digital Education (CODE) at IIT Madras. Both the initiatives, NPTEL and IITM BS, have been significant initiatives in the area of large-scale online education and have been game changers in making quality higher education accessible in India.

KEYNOTE SPEAKER

C3: ICCE Sub-Conference on Advanced Learning Technologies, Learning Analytics, Platforms and Infrastructure (ALT/LA/P/I)



Minhong Wang
The University of Hong Kong

Rethinking How People Learn in the Age of AI

How people learn has long been discussed, revealed by many learning theories, explored in extensive practices, and analyzed by numerous studies. In the age of AI, technology is playing an increasingly important role in changing and improving human learning. This talk will present a high-level view of human learning from four fundamental perspectives, that is, learning through interaction with content (C), learning through interaction with other people (O), learning through interaction with self (S), and learning through interaction with tasks or practices (T), the so-called COST framework. Moreover, the human learning process involves multiple dimensions such as affect, behaviour, and cognition, which influence each other. Based on the COST framework and the affect-behaviour-cognition model, this talk will summarize how AI and related technologies can support human learning, how to design effective learning to address learners' needs and challenges, and how to make meaningful analysis of technology-supported human learning.

Dr. Minhong (Maggie) Wang is Professor and Director of the Laboratory for Knowledge Management & E-Learning, Faculty of Education, The University of Hong Kong. She is also Kuang-Piu Chair Professor at Zhejiang University, Eastern Scholar Chair Professor at East China Normal University, and Visiting Research Professor at the Advanced Innovation Center for Future Education of Beijing Normal University. She is the Editor-in-Chief of Knowledge Management & E-Learning (impact factor 2.5). Her expertise includes learning technologies for cognitive development, creative thinking and complex problem solving, knowledge management and visualization, and artificial intelligence applications. Her studies focus on making tacit knowledge and complex cognitive processes visible and accessible to learners and helping them to develop expert-like knowledge and performance in solving complex problems. She has published more than 200 items, including 136 journal articles (83 in SSCI/SCI indexed journals; 33 articles in top 10 journals and 19 articles in top 3 journals in multiple disciplines). Based on Clarivate's Essential Science Indicators (ESI), she is recognized as a top 1% Scholar in (a) Social Sciences, General, and (b) Economics & Business.

THEME-BASED INVITED SPEAKER



Martina Holenko Dlab

University of Rijeka,

Faculty of Informatics and Digital Technologies

C4: ICCE Sub-Conference on Technology Enhanced Learning for Mobility of Learners and Learning Experiences (TEML)

Aligning educational technology innovation with learning objectives through learner-centered design

Technologies offer new opportunities to address persistent challenges in education, particularly the need to motivate learners and support their active engagement in cognitively demanding learning processes. Approaches that integrate educational technologies such as artificial intelligence, gamification platforms, immersive environments and educational robotics are increasingly seen as promising ways to create more engaging and personalized learning experiences.

While the adoption of these technologies continues to grow, their impact on learning remains uncertain. Many learners actively engage with them, yet their effectiveness and efficiency are not consistently supported by research. In some cases, implementation occurs without clear alignment to specific learning objectives or pedagogical frameworks. This highlights the need to move beyond a technology-first approach. To fully realize the potential of these tools, their integration should be intentionally designed to support teaching and learning. This requires alignment with learning objectives, guidance from sound pedagogical principles, and a foundation in effective learning design.

Martina Holenko Dlab is an Associate Professor and Head of the Chair of Multimedia Systems and E-Learning at the Faculty of Informatics and Digital Technologies, University of Rijeka, Croatia. Her research focuses on technology-enhanced learning, with particular interests in blended learning, computer-supported collaborative learning (CSCL), game-based learning, recommender systems, and the use of artificial intelligence in education. She has co-authored numerous scientific and professional papers in these fields. Her research is regularly presented at international conferences and contributes to the advancement of digital education practices. She has extensive experience in designing, implementing, and evaluating technology-enhanced learning models, as well as in delivering teacher training through national and international projects.

THEME-BASED INVITED SPEAKER



Jie Chi Yang
National Central University, Taiwan

C5: ICCE Sub-Conference on Educational Gamification and Game-based Learning (EGG)

Unpacking the Impacts of Learner Differences through Digital Game-Based Learning

Digital Game-Based Learning (DGBL) has emerged as a powerful educational approach to enhance learners' motivation and engagement. However, increasing evidence indicates that diversities exist among learners' reactions to DGBL, due to a range of individual differences. This talk synthesizes key findings from over a decade of empirical research examining how human factors, such as prior knowledge, gaming experience, gender, and anxiety levels, influence both learning outcomes and in-game behavior within DGBL. Drawing on empirical evidence from multiple studies, I will demonstrate how specific learner characteristics can affect the impacts of learning strategies within DGBL. Finally, this talk will highlight design solutions for DGBL with the ultimate goal of informing the development of inclusive learning environments that can accommodate a wide range of individual needs and preferences.

Jie Chi Yang is a Distinguished Professor in the Graduate Institute of Network Learning Technology at National Central University (NCU), Taiwan. He also currently serves as the Director of the Research Center for Science and Technology for Learning at the NCU. Dr. Yang's research interests include digital game-based learning, computer-assisted language learning, human factors, and interactive learning environments. He has published over 100 peer-reviewed journal articles and conference papers, with the majority of his journal articles appearing in top-tier international journals. He has obtained substantial research funding from Taiwan's National Science and Technology Council (NSTC) to support the development of adaptive instructional strategies and game-based learning platforms. Dr. Yang has played key roles in organizing major international conferences. Notably, he served as the International Program Committee Coordination Chair for ICCE 2018. In recognition of his contributions to the field, he acquired the Outstanding Research Award from the NSTC in 2019.

THEME-BASED INVITED SPEAKER

C6: ICCE Sub-Conference on Technology Enhanced Language Learning (TELL)



Yun WEN
Nanyang Technological University (NTU), Singapore

Evolving Learner Agency: From AI Support to Learner-AI Shared in Second Language Learning

Recent advancements in generative artificial intelligence (GenAI) have led to significant transformations across educational contexts, with particularly profound implications for language learning. For young second language learners, who are still developing cognitive, linguistic, and self-regulatory capacities, these changes call for the thoughtful and holistic design of GenAI-supported learning environments. This talk will draw on concrete cases to examine why learner agency is essential and how such agency is shaped, supported, or constrained in the AI-powered learning environment. This talk invites a rethinking of AI's potential in second language learning contexts, provides insights into the design of AI-powered language learning environments, and highlights methodological considerations for investigating their effectiveness.

Wen Yun is an Assistant Professor in the Learning Sciences and Assessment Department at the National Institute of Education (NIE), Nanyang Technological University, Singapore (NTU). Her research investigates how people learn through interaction and dialogue in multimodal environments, and how emerging technologies, such as Augmented Reality (AR) and Artificial Intelligence (AI), can foster high-quality interactions to promote teaching and learning. She leads interdisciplinary teams to translate research into practical educational applications, developing impactful and sustainable technology-enhanced learning solutions. She was awarded the Social Science and Humanities Research Fellowship by the Singapore Social Science Research Council in 2022, and the first recipient from the NIE to receive the prestigious NTU Nanyang Research Award. She currently serves as an Associate Editor for Educational Technology & Society and the Asia Pacific Journal of Education.

Table of Contents

C1: Artificial Intelligence in Education/Intelligent Tutoring and Adaptive Learning (AIED/ITS)

A Computer Vision–Based Approach for Assessing Student Behavioral Engagement in the Classroom ____ 2
Yaping Xu, Haogang Bao, Yaqian Zheng, Siyu Wang, Sisi Wu and Yanyan Li

Evaluating the Effectiveness of Large Language Models for Course Recommendation Tasks ____ 11
Boxuan Ma, Md Akib Zabed Khan, Tianyuan Yang, Agoritsa Polyzou and Shin'Ichi Konomi

Investigating Algorithmic Bias in Affect Detectors with Constructed Categories of Student Identity ____ 21
Nidhi Nasiar, Clara Belitz, Haejin Lee, Frank Stinar, Ryan S. Baker, Jaclyn Ocumpaugh, Stephen Fancsali, Steve Ritter and Nigel Bosch

Coached by My Future Self: A Syllabus–Driven AI Mentor to Build Grit, Agency, and Learning Performance through a Mantle–of–the–Expert Journey ____ 31
Yao–En Chen, Vando Gusti Al Hakim, Su–Hang Yang, Jen–Hang Wang, Rayvin Lai and Gwo–Dong Chen

Investigating the Effectiveness of Explanations in Active Video Watching ____ 41
Raul Vincent Lumapas, Antonija Mitrovic, Matthias Galster, Sanna Katariina Malinen and Jay Holland

Applicability of Facial Video–Based Alertness Estimation Model in Real Online and In–Person Classrooms ____ 51
Atsushi Ashida, Ryosuke Kawamura, Hideaki Hayashi and Hajime Nagahara

Towards Scalable Annotation of Math Word Problems: Knowledge Component Tagging via LLMs and Sentence Embeddings ____ 61
Chor Seng Tan, Chengwei Wei and Jung–jae Kim

Identifying Cognitive and Metacognitive Processes Using Retrospective and Concurrent Think–Aloud Protocols ____ 71
Rumana Pathan, Sahana Murthy and Ramkumar Rajendran

Did You Mispronounce or Did I Mishear? — Detecting Kanji Mispronunciations in Children's Oral Reading ____ 81
Takuya Matsuzaki and Arata Saito

GAI Pedagogical Agents and Teacher–Guided Prompt Use in EFL Debate Training ____ 90
Farhad Cassim and Jie Chi Yang

Relevant Infinite Content Genesis with Vector Store and AI agents: A PERT-Driven Hybrid Framework for NLP – Enhanced Semantic Retrieval and Evaluation 100
Ayushman Pranav, Rajesh Kumar Modi, Umesh Gupta, Ankit Dubey and Pankaj Mishra

Can GAI improve the writing speed and writing quality? A survey study of college students in China 106
Bijun He

Can GPT Detect Gaming the System in Text Replays? 112
Cristina Maier and Ryan Baker

Coded Narratives: Leveraging Women AI Scientists' Memoirs for Inspiration and Knowledge in K-12 AI Education 118
Shuchi Grover

A Federated Learning Neural Network For Student Dropout Prediction 124
Anuradha Kumari Singh and Karthikeyan S.

Automatic Distractor Generation in Multiple-Choice Questions Using Large Language Models with Expert-Informed Distractor Strategies 130
Yusei Nagai and Masaki Uto

The Construction of Evaluation Indicators for Intelligent Educational Software and the Measurement of Its Developmental Level 136
Bo Zhang, Yuanyuan Li, Qianchao Xing, Yuhan Chen and Xinying Wang

A Study of Precise Instructional Intervention Strategies Based on Learner Portraits 142
Jingyi Zhang, Xinyu Sun, Nan Dang, Meng Li, Huilong Li, Minsheng Fan and Weitong Guo

Generative AI pedagogical agent as learners' inquiry assistant: Designing and implementing Situational Teaching 148
Lin Li and Yan Liu

Ambiguity Resolution with Human Feedback for Code Writing Tasks 154
Aditey Nandan and Viraj Kumar

Write or not to Write. Attitudes of Students and Academics towards Written Dissertations in the Light of Artificial Intelligence Development 160
Ewa Duda, Ewelina Młynarczyk-Karabin and Julia Więckiewicz-Modrzewska

Small but Significant: On the Promise of Small Language Models for Accessible AIED 166
Yumou Wei, Paulo Carvalho and John Stamper

SRL profiles in math problem-solving: The Essential Role of Monitoring and Translating for Outcomes and Beliefs 172
Nidhi Nasiar, Jaclyn Ocumpaugh, Ryan S. Baker, Caitlin Mills, Samuel Rhodes and Sheela Sethuraman

Effects of Gen AI in Enhancing Pre-service Teachers' Instructional Design Competency: A Quasi-Experimental Study 178
Le Zhang, Hongchao Peng and Mingrui Zhang

An Exploration of the Impact of Using ChatGPT on Students' On-Task Performance and Task Success 184
Arnel Ocay and Maria Mercedes Rodrigo

RoboTuB: Retrieval-Augmented Tutoring for Adaptive Learning in STEM MOOCs 190
Avijit Pandey, Gaurav Misra, Sunny Prajapati and Syaamantak Das

Idea Space Change Process Visualization System for Supporting Interpretation of Thinking Strategy in Research Instruction 196
Kota Kunori and Tomoko Kojiri

Audience Robot Control System Adapted to Presenter's Skills for Presentation Practice 202
Yuya Kishimoto and Tomoko Kojiri

A Human-AI Collaborative Assessment of AI-Generated vs. Human-Created MCQ Distractors 208
Zifeng Liu, Priyadharshini Ganapathy Prasad, Bach Ngo, Xinyue Jiao and Wanli Xing

AI Support for Translanguaging in International School Classrooms 214
Alina Guha, Gyeongri Kim, Shamya Karumbaiah, Esther Bettney Heidt and Mariana Castro

Bridging Computational Thinking skill gaps with ASSET and Mindspark CT 223
Ashwini Chandrashekhar and Aditya Vikram Singhania

Educational Cone Model in Embedding Vector Spaces 228
Yo Ehara

Generative AI and Multimodal Profiling in Programming Education: A Systematic Review of Tools, Pedagogical Impacts, and Future Competencies 234
Aboul Hassane Cisse and Shoya Ishimaru

Designing an AI Reading Assistant to Support Home—School Engagement 237
Jhih-Kai Tang, Yi-Cheng Tsai, Jia-Yan Lee, Chang-Yen Liao and Tak-Wai Chan

Alternative Story Construction Support System for Narrative Therapy through Fact Recall Assistance 240
Rei Muguruma, Yano Shogo and Tomoko Kojiri

Fostering Creative Thinking in Animation Scriptwriting through AI-Driven Analysis and Socratic Questioning 243
Panadda Jaiboonlue, Wasan Na Chai and Nattapol Kritsuthikul

C2: Computer-Supported Collaborative Learning and Learning Sciences (CSCL/LS)

Enhancing Learning Achievement through GenAI Assisted Collaborative Drama-Based Role-Play Learning Metaverse in Higher Education 247
Su-Hang Yang, Mahesh Liyanawatta, Tzu-Hsien Chuang, Hao Tang, Jen-Hang Wang and Gwo-Dong Chen

How Student-Led Collaborative Metaverse World Creation Fosters Information Literacy and Intercultural Competence: A Quantitative Analysis of a Synchronous International Virtual Exchange 257
Masako Hayashi, Takehiro Suzuki, Hirokazu Hayama and Keisuke Goto

Understanding the Triggers and Dynamics of Socially Shared Metacognitive Regulation in Teams with Divergent Outcomes 267
Vishwas Badhe, Ramkumar Rajendran and Chandan Dasgupta

Persistence in Electronic-Making Projects: A Socio-Cognitive Perspective 277
Nisumba Soodhani, Ramkumar Rajendran and Sridhar Iyer

Comparing students' social knowledge construction in the Metaverse-based and chat-based collaborative learning 283
Suwandi Suwandi, Darmawansah Darmawansah, Benni Ichsanda Rahman Hz and Hsueh-Hu Chuang

The Use of Sensor Data for the Multimodal Analysis of Mathematical Proof Activity in a CSCL Environment 289
Masataka Kaneko, Hironori Egi, Takuya Kitamoto and Takeo Noda

The Desirability of Peer-Assisted Learning Spaces for Online Student-Centered Error Correction: Preliminary Evaluation Study 295
Fu-Yun Yu

Development of a Maker Education Instructional Design Model: Delphi-based Analysis 298
Guang Chen, Xinru Liu and Lei Wang

C3: Advanced Learning Technologies, Learning Analytics, Platforms and Infrastructure (ALT/LA/P/I)

Impact of LLM Feedback on Learner Persistence in Programming ____ 302
Yiqiu Zhou, Maciej Pankiewicz, Luc Paquette and Ryan Baker

Too Detailed to Share? Towards Risk-Based Privacy Protection of Fine-Grained Educational Data ____ 312
Hibiki Ito, Chia-Yu Hsu and Hiroaki Ogata

Logged In and Mindset On: Log Data Based Insights on Engagement and Learnings in Programming CBLEs ____ 322
Indrayani Nishane, Ramkumar Rajendran and Sridhar Iyer

How do students do design? A data-driven approach for understanding students' engineering design behaviors ____ 332
Shruti Jain, Namrata Srivastava, Clayton Cohn, Naveeduddin Mohammed and Gautam Biswas

Quantifying the Benefits of the Reread- Before-Answer Strategy in Japanese High- School EFL ____ 342
Isanka Wijerathne, Brendan Flanagan and Hiroaki Ogata

Investigating the Impact of Time Constraints on Sudoku Solving: A Multimodal Analysis of Cognitive and Affective States ____ 352
Deepika Yadav, Tushar Patle, Sonika Pal, Amit Paikrao and Ritayan Mitra

Unpacking student engagement with everyday ethics in GenAI mediated dialogue ____ 362
Lakshmi T G, Gurudeo Thakare, Alekh Velayudhan and Karishma Shanghvi

Dialogic Expertise in Practice: How Expert Teachers Orchestrate Classroom Talk in Distinct Mathematics Lesson Types ____ 372
Siyu Wang, Yaqian Zheng, Mingze Sun, Yaping Xu and Yanyan Li

A Deep-Learning Technique for Converting Bengali Handwritten Answer-Book Images into Text ____ 382
Moumita Moitra, Malay Kumar Majhi and Sujan K Saha

VisionTutor: An Adaptive Tutoring Platform for Real-Time Progressive Learning ____ 392
Sunil Gopal C V, Shreyas Gutti, Sudhan S, Sushanth R, Abhilash C B, Rohitaksha K, Manjunath K V and Ashwin T S

Do Categorical Emotions Have Intensity? Modeling Emotional Intensity Transitions by Task, Peer, and Environment Causes ____ 402
Ryan Li, Ashwin T S, Divya Mereddy and Gautam Biswas

Evaluating Students' Scientific Inquiry Strategies in Large-Scale Digital Assessments ____ 412
Namrata Srivastava, Emma Linsenmayer, Mario Piacentini and Gautam Biswas

A Secured Knowledge Management System to enhance the User Experience in Engineering Education 422
Rajeswari K., Vivekanandan N., Sakshi Kulkarni, Janhavi Mali and Vinayak Shete

Challenge-Based Learning and ICT Integration in Mechanical Engineering Education: A Case Study on Mechanical Engineering Design Courses 428
Vivekaandan N, Rajeswari K and Yuvraj Kanna N V

User Evaluation of a Virtual Patient for Philippine Medical Education 434
Maria Mercedes T. Rodrigo, Samantha Castaneda, James Alvir Maclin V. Alaan, Paolo Santino P. Caoile, Alec Isaiah Dayupay, Roswold Jemuel Sanchez and Eric Cesar Jr. E. Vidal

Critical Thinking as a Dynamic System: Tracking Nonlinear Development through Reading Study Worksheets in Elementary Science 440
Manyu Huang, Xueqing Zhou and Guang Chen

Assessment of Comment Quality in Active Video Watching using Deep Learning 446
Chantelle D. M. Baldwin and Antonija Mitrovic

Exploring the Relationship Between System Operation Behaviors and Learning Achievement in Agricultural Education 452
Haiqiao Liu, Tsubasa Minematsu, Chengjiu Yin, Sijie Xiong and Atsushi Shimada

CodeRunner Agent: Integrating AI Feedback and Self-Regulated Learning to Support Programming Education 458
Huiyong Li and Boxuan Ma

Trial Analysis on Behavioral Logs to Identify Risk Tactics and Infer Non-Cognitive Skills in Elementary Students 464
Shiang-Yi Chou, Taku Kawada, Yasuhiro Sato, Tadashi Inagaki and Toru Nagahama

Learning from Others' Codes in Game-based Robot Programming: Behavioral Patterns and Outcomes 470
Shintaro Maeda, Kento Koike and Tomoto Takahito

Teacher-Involved Automatic Characteristics Classification in Handwritten Math Answer Process 476
Shunsuke Tonosaki, Taito Kano, Chia-Yu Hsu and Hiroaki Ogata

Recomposition Based Learning for Promoting Structural Understanding – From Reconstruction of External Representations to Recomposition of Internal Representation– 482
Tsukasa Hirashima and Kodai Watanabe

Behavior-Aware Argumentation: Integrating Toulmin's Model for Adaptive Learning in Digital Environments 488
Banni Satria Andoko, Vivin Ayu Lestari, Rosa Andrie Asmara, Usman Nurhasan, Galih Putra Riatma, Ananda Galih Pratiwi and Tsukasa Hirashima

Dual Analysis of Literature and Existing Tools for Establishment of Acceptable xAPI Profile for Collaborative Learning 494
Mitsushiro Ezoe, Yasuhisa Tamura and Ren Komatsuzaki

Enhancing Retrospective Think-Aloud Protocols with Eye-Gaze data 500
Amit Paikrao, Debarshi Nath, Ramkumar Rajendran and Ritayan Mitra

Improvement of a Japanese Language Learning Support System that Enables Word Accent Learning 506
Kazuki Tomita, Satoru Kogure, Yasuhiro Noguchi, Koichi Yamashita, Tatsuhiro Konishi and Makoto Kondo

INSTd: A Computer-Adaptive Assessment of Definition-Based Comprehension Skills 509
Noriko Arai, Naoya Todo, Teiko Arai, Shingo Sugawara, Takuya Matsuzaki and Koken Ozaki

A Dual-Framework Approach to Evaluate Marker less AR Application: Insights from SUS and HARUS Questionnaires 512
Ajay Shankar Tiwari, Toluchuri Shalini Shanker Rao, Utanko Mitra and Kaushal Kumar Bhagat

Multimodal Trait Scoring for Video Interviews Using Neural Models with Handcrafted Features and Trait-Attention 515
Taichi Kitajima and Masaki Uto

Learning Support System Visualizing Behavior of Student Program in Object Oriented Language Based on Teacher's Intent of Instruction 518
Koichi Yamashita, Kenzo Kobayashi, Takuma Kokubo, Satoru Kogure, Yasuhiro Noguchi, Raiya Yamamoto, Tatsuhiro Konishi and Yukihiro Itoh

Development of a Coding Framework for Cognitive Engagement Based on Online Asynchronous Discussion 521
Junhao Wan and Guang Chen

C4: Technology Enhanced Learning for Mobility of Learners and Learning Experiences (TEML)

A Cross-Platform Engagement Analysis for Supporting Knowledge Transformation Through Complex Learning Settings 525
Manabu Ishihara, Izumi Horikoshi and Hiroaki Ogata

A Study of Strategies to Enhance Student Engagement in Middle School Classrooms with Instructional Feedback in a Smart Paper – Pen Environment 535
Nan Dang, Jingyi Zhang, Xinyu Sun, Xiaoyan Wang, Huilong Li, Minsheng Fan and Xiaodong Wei

Predicting Students' Online Learning— Character Strengths, Online Self-Regulated Learning, and Learning Satisfaction 541
Chin Lee Ker and Gaoxia Zhu

Meta AI for Leadership Training: A Pilot Study on Mobile Social Media—Based Blended and Self-Regulated Learning 547
Riska Saputra, Zainul Arifin, Tsaqufal Jaliliy and Husni Mubarok

C5: Educational Gamification and Game-based Learning (EGG)

Impact of Gamification on Engagement and Learning in Video-Based Platforms 554
Pasan Peiris, Matthias Galster, Antonija Mitrovic, Sanna Malinen, Ehsan Bojnordi and Jay Holland

Impact of Digital Learning Game Narratives on Affective Experiences and Learning 564
J. Elizabeth Richey, Daniel Pollock, Mahboobeh Mehrvarz, Aditi Haiman, Hayden Stec, Leah Teffera and Bruce M. McLaren

From Roof Cracks to Wound Healing: How SBF Analogy Shapes Students' Reasoning in Complex Biology Processes 574
Meera Pawar, Rajashri Priyadarshini and Sahana Murthy

Integrating Generative AI into Escape Room-Based Learning: Effects on Nursing Students' Clinical Skills, Flow Experience, and Self-Efficacy 584
Intan Setiani, Ching Yi Chang, Riska Saputra and Jie Chi Yang

Design and Testing of a VR Escape Room Game for Philippine Martial Law History 590
Eric Cesar Jr. E. Vidal, Nicko R. Caluya, Johanna Marion R. Torres, Jesus Alvaro C. Pato and Kenneth King L. Ko

C6: Technology Enhanced Language Learning (TELL)

Exploring EFL Learners' Perceptions of Generative AI through the Technology Acceptance Model: Insights from an AI Literacy Program 594
Michiyo Oda, Ryo Saito and Tatsuya Horita

Chatbot Personalisation for EFL Learning: Integrating BKT and Task Context 604
Steve Woollaston, Brendan Flanagan and Hiroaki Ogata

Designing Metaverse-based language learning: How pre-service teachers applied the Triple-E framework 614
Thien-Bao Nguyen, Darmawansah Darmawansah and Van-Phu Nguyen

Web Application to Help Learn English Using Medical Bilingual Corpus – Updates to Improve Learners' Writing Skills 620
Riku Yasuda, Yoshinori Miyazaki, Motoko Asano and Miho Fujieda

Investigating the Impact of AI Generated Mnemonic Illustrations on Learning Japanese Kanji 626
Mayur Bhurle, Vishwas Badhe and Ramkumar Rajendran

EmotiCon: Enhancing Public Speaking Training with Context-Aware Feedback on Emotional Delivery 632
Yash Desai, Praneeth Kasiraju, Syaamantak Das and Ramkumar Rajendran

C7: Practice-driven Research, Teacher Professional Development and Policy of ICT in Education (PTP)

Developing Teachers' Digital Competence through Expansive Learning: A Study of Teacher-Researcher Collaborative Teaching Research 639
Junyang Liu, Bian Wu and Dong Li

Computational Thinking (CT) Proficiency Among Malaysian Teacher Trainees: A Cross-Institutional Analysis in the Central Zone 649
Ahmad Sarji Abdul Hamed, Su Luan Wong, Mohd Zariat Abd Rani, Mas Nida Md Khambari, Nur Aira Abd Rahim, Priscilla Moses and Nur Izzati Mat Zin

Predicting Learners' State of Mind from Cross-Contextual Educational Log Data 658
Junya Atake, Chia-Yu Hsu and Hiroaki Ogata

Empowering, Not Replacing: Using Generative AI to Coach Educators in Providing Effective Feedback 668
Antonette Shibani, Jayakrishnan Warriem, Lisa-Angelique Lim, Gnana Bharathy and Arup Chattopadhyay

Pre-Digital to AI Eras: Changing Role of a Teacher with Technology 678
Vinay Yadav, Ishika Ishika and Durgaprasad Karnam

An Exploratory Study on the Influence of Anonymity, Personality, and Social Factors on Participation in Online College Communities 688
Ryan Ebardo, Eleana Louise Hernandez, Vanessa Claire Ma, Marco Lorenzo Mitra and Marco Gabriel Tarrosa

Confronting the Impact and Challenges of the AI Era: Teachers' Creative Use of Digital Tools to Enhance Digital Teaching Competence 694
Lin Li and Yan Liu

Teaching Machine Learning to Senior Primary Students: Evaluating a Course for AI Literacy in a Hong Kong Primary School ____ 700
Siu Cheung Kong, Yin Yang and So Sum Chow

Co-Designing AI and Multimodal Analytics Integrated Reflective Learning in Japanese High School Context ____ 706
Rwitajit Majumdar, Shin-Ichiro Kubota and Huiyong Li

Visualizing Knowledge-Based Learning Logs to Support Multi-Perspective K-12 Math Instruction ____ 712
Yudai Okayama, Junya Atake, Kensuke Takii, Changhao Liang and Hiroaki Ogata

Conflow: An Online Concept-Mapping Tool to Measure Conceptual Fluency ____ 718
Shivin Agarwal, Anveshna Srivastava and Nisheeth Srivastava

An Exploratory Study of a Generative AI Teacher Coach ____ 724
Karishma Shanghvi and Sridhar Iyer

Learning Sciences in Action: A Reflective Case on the Pedagogy of a Learning Sciences Course for Postgraduates ____ 730
Durgaprasad Karnam and Amrit Pal Singh

Analysis and Comparative of the Academic Draft on Coding and Artificial Intelligence Education in Indonesia Primary and Secondary Schools ____ 737
Riska Saputra and Adventhius Immanuel Karo Karo

Configurational Influence of Self-Efficacy, Learning Strategies, and Emotions on MOOC Learning Outcomes: An fsQCA Analysis ____ 743
Felix Arokiya Raj A.P. and Arul Aram I.

An Online Drill-and-Practice System Anchoring the Cultivation of a Growth Mindset: Design Concepts, Core Components, and Preliminary Evaluation ____ 749
Fu-Yun Yu

Attaining the Program Educational Objectives of the IT Program Three to Five Years After Graduation: An Evidence-Based Approach ____ 752
Arlene Mae Celestial-Valderama, Roel Traballo, Don Erick Bonus and Sharmaine Justyne Maglapuz

Effect of ISIED Education Activities on The Development of Students' Key Competency ____ 760
Yuhan Zhou, Jiaru Wu and Guang Chen

Teachers as Learners: Exploring the Role of Course Structure on Learners' Agency ____ 763
Sumitra Sadhukhan, Leena Bhattacharya and Sridhar Iyer

Panels

Developing K–12 AI Literacy in Different Countries: Approaches and Challenges 767
Shuchi GROVER, Balaraman RAVINDRAN, Siu-Cheung KONG, Maria M.T.RODRIGO, Lung-Hsiang WONG Weena NAOWAPRATEEP

Professional Development of STEM Teachers Using the DECODE Model in India and Taiwan 774
P. S. SREEDEVI, P RAJASEKARAN, Chun-Yen CHANG, Marison Sudianto MANALU

C1
Artificial Intelligence in Education/Intelligent Tutoring and Adaptive Learning (AIED/ITS)

A Computer Vision-Based Approach for Assessing Student Behavioral Engagement in the Classroom

Yaping XU[a,b], Haogang BAO[c], Yaqian ZHENG[d], Siyu WANG[a], Sisi WU[a], & Yanyan LI[a,*]
[a]*Faculty of Education, Beijing Normal University, Beijing, China*
[b]*School of Educational Technology, Northwest Normal University, Lanzhou, China*
[c]*The China National Academy of Educational Sciences, Beijing, China*
[d] *College of Education, Hebei University, Baoding, China*
*liyy@bnu.edu.cn

Abstract: With increasing emphasis on enhancing classroom engagement and learning outcomes, efficient and accurate assessment of student behavioral engagement has become essential. Traditional methods such as classroom observations and self-reports are often subjective, time-consuming, and lack scalability. To address these limitations, this study proposes a computer vision-based approach for automatically assessing student behavioral engagement in classroom environment. Specifically, an improved VGG16-Attn model is proposed by integrating attention mechanisms to enhance feature extraction and boost behavior recognition performance. Our results indicate that the proposed method not only accurately detects student behavior types but also, by further aggregating and calculating these behaviors, objectively and effectively captures the dynamic evolution of behavioral engagement. This method offers a promising solution for educators to gain actionable insights into student engagement, ultimately contributing to more personalized and effective teaching strategies.

Keywords: Behavioral engagement, computer vision, classroom environment

1. Introduction

Student engagement can be defined as the degree to which students participate in learning activities, encompassing their attention, involvement, and effort (Fredricks et al., 2004). Research has shown a significant positive correlation between student engagement and academic achievement (Gunuc, 2014). Student engagement is conceptualized as a multidimensional construct consisting of cognitive, emotional, and behavioral components (Fredricks et al., 2004; Cooper, 2014), where cognitive and emotional engagement are relatively implicit and difficult to observe directly. In contrast, behavioral engagement is more easily assessed as it is directly reflected in students' observable actions, such as classroom interactions, answering questions, and completing assignments. Therefore, behavioral engagement is often the most accessible dimension to evaluate and is the most commonly used indicator in many studies. Traditionally, engagement has been assessed through self-reports, teacher observations, and surveys, but these methods often suffer from subjectivity, lack of real-time feedback, and difficulties in scaling for large classrooms.

Recent advancements in computer vision and deep learning have provided new avenues for automating the assessment of student behavioral engagement in classroom environments. These technologies offer the potential to capture real-time behavioral data with greater accuracy and consistency compared to traditional methods. A growing body of research has demonstrated the potential of using video-based analysis for tracking student behaviors (Ngoc Anh et al., 2019; Guo, 2022; Trabelsi et al., 2023). Despite these advancements, video-based behavioral analysis still faces several challenges in practical applications. First, the performance of behavior recognition models is limited by the scale and diversity of available

datasets. Many datasets fail to adequately cover various classroom environments and behavior types. Second, existing research often lacks long-term tracking of student behaviors, which hinders a comprehensive analysis of the evolution of students' behavioral engagement. Additionally, most studies primarily focus on the technical aspects of behavior recognition. While these studies have achieved relatively accurate recognition of student behaviors, they have not effectively integrated this behavioral data into a comprehensive assessment of behavioral engagement.

To address the aforementioned challenges, this study constructed a large-scale, high-quality dataset specifically designed to support the task of student behavior recognition in classroom contexts. On the basis of this dataset, a computer vision–driven framework was proposed for the quantitative assessment of students' behavioral engagement during live classroom sessions. This framework integrates advanced visual analysis techniques to capture and interpret students' actions, enabling a more objective and scalable measurement of engagement levels. The findings of this study contribute to providing teachers with accurate classroom feedback and offer data support for optimizing teaching strategies and implementing personalized education.

2. Literature Review

2.1 Behavioral Engagement and Its Typical Measurement

The definition of classroom behavioral engagement can be categorized into two main perspectives: one focuses on student compliance with rules, specifically reflecting behaviors such as attendance rates. For instance, Finn (1989) suggests that learning engagement helps identify the process of students gradually becoming disengaged and alienated from school, thus enabling timely interventions to support students in completing their academic work. The other perspective examines the extent of students' deep involvement in learning activities, including effort, persistence, and concentration (Fredricks et al., 2004). This is typically reflected in behaviors such as completing assignments, attending classes, answering questions, and participating in discussions. This study, building upon the foundation of prior research, conceptualizes behavioral engagement as the proactive and purposeful interaction of students with various learning resources, instructors, and peers within the classroom context. Such engagement encompasses actions that reflect students' participation, persistence, and effort in academic tasks. By systematically observing students' behavioral states throughout the learning process, the degree of their engagement in classroom activities can be assessed.

Traditional methods for measuring classroom behavioral engagement primarily include self-reports and classroom observations. Early research primarily involved learners self-reporting their levels of engagement. This method is highly practical and provides large-scale data at a relatively low cost. However, self-reports are prone to subjective bias, and students may not always provide truthful responses. Additionally, this approach fails to offer process-oriented analysis, making it difficult to track the dynamic changes in student engagement over time. Classroom observation, on the other hand, uses pre-established observation scales to capture students' behavioral characteristics during the learning process, thereby assessing their behavioral engagement (Volpe et al., 2005). The primary limitation of this method lies in the substantial investment of time and effort required from researchers, rendering it impractical for large-scale investigations and unsuitable as a routine instrument for process-oriented assessment. However, the advent of computer vision technology has introduced new opportunities to enhance and transform traditional classroom observation approaches. Leveraging cameras, sensors, and sophisticated algorithms, computer vision systems are capable of capturing and processing visual data in real time, thereby offering researchers granular and dynamic insights into student behaviors and interactions. In particular, recent advancements in deep learning have significantly expanded the analytical capabilities of these systems, enabling more accurate, scalable, and automated behavioral analysis. As a result, research integrating computer vision with educational observation has attracted growing

scholarly interest in recent years, paving the way for more efficient and data-rich methods of studying classroom engagement.

2.2 Classroom Behavioral Engagement Recognition via Computer Vision

Computer vision technology aims to detect, recognize, track, and understand objects through the processing and analysis of image or video data. Researchers can utilize this technology to perform fine-grained recognition of students' body movements in classroom videos, enabling the automated assessment of behavioral engagement. Early studies in this field primarily focused on using traditional machine learning methods to recognize classroom engagement (Karimah & Hasegawa, 2022). For instance, Zaletelj et al. (2017) trained a classifier to automatically assess students' attention levels in the classroom by extracting features from facial expressions and body posture, and combining them with various traditional machine learning algorithms. The results indicated that the three-level attention classifier achieved a moderate accuracy of 75.3%. The advantage of traditional machine learning methods lies in their relatively simple model structures, lower computational requirements, and suitability for smaller datasets. However, these methods are highly dependent on manual feature extraction and have limited effectiveness in handling complex and dynamic student behaviors.

With the advancement of artificial intelligence, deep learning methods have demonstrated their superiority in image recognition tasks. The end-to-end learning process not only simplifies the workflow of traditional machine learning tasks but also enhances the capability of deep neural networks through a hierarchical feature extraction process. Consequently, in recent years, an increasing number of studies on student behavior recognition have adopted deep learning methods. For example, Li et al. (2023) used classroom videos from a smart classroom to generate an image dataset containing seven typical learning behaviors. Based on this dataset, they developed a behavior recognition network model, which includes a backbone network (SlowFast R-101) for feature extraction, a target detector (Faster R-CNN), and an action classifier composed of fully connected layers. Ikram et al. (2023) employed a VGGNet16 model based on transfer learning to calculate students' learning engagement levels in real classroom settings by extracting external behavioral features. Additionally, Xiong et al. (2024) proposed a CNN-Transformer model that simultaneously captures both coarse- and fine-grained information, significantly improving the accuracy of learning engagement recognition in classroom contexts.

In summary, research on classroom behavioral engagement recognition based on computer vision has evolved from traditional methods based on handcrafted features to deep learning approaches. Although deep learning methods are widely favored for their ability to provide more accurate recognition of student behavior, they often suffer from lower interpretability and typically require large amounts of labeled data for training. Existing datasets in current research are relatively limited in size, which can lead to overfitting issues. Furthermore, due to privacy concerns and other factors, acquiring datasets from external institutions presents significant challenges. In response to these issues, this study has constructed a large-scale student behavior dataset by collecting real classroom videos and has conducted an in-depth analysis of engagement levels based on behavior recognition results.

3. Methodology

3.1 Student Behavior Dataset

Due to the absence of publicly available datasets tailored to real-world classroom settings, we constructed a large-scale student behavior dataset to support behavior recognition model development and evaluation. The dataset includes 152,823 annotated images extracted from actual classroom videos, each labeled with one of nine representative behaviors: listening, reading, writing, raising hand, standing, leaning the body, looking around, lying on desk, and other, Figure 1 shows example images. These categories were selected for their pedagogical

relevance and importance in evaluating student engagement. The dataset captures diverse classroom environments, camera angles, postures, and occlusion conditions, making it highly representative. This resource provides a solid foundation for advancing computer vision-based behavior analysis in education.



Figure 1. Example images from the behavior dataset.

3.2 Data Augmentation

The annotated behavior dataset exhibits class imbalance, which may lead to biased predictions if directly used to train deep learning models. In such cases, the model is likely to favor the majority class, resulting in poor recognition performance for the minority classes. To expand and balance the image dataset, this study employed various data augmentation (DA) techniques, including rotation, flipping, noise addition, and blurring, as shown in Figure 2. After applying these augmentation methods, the final dataset used to train the behavior recognition model contained 312,459 images.



Figure 2. Some examples of DA.

3.3 Improved VGGNet16 for Recognizing Student Behavior

Although VGGNet16 (Simonyan & Zisserman, 2019) is a classic convolutional neural network with strong performance in image classification, it lacks the ability to adaptively focus on important spatial and channel-wise features. This limitation reduces its effectiveness in distinguishing key behavioral cues from complex classroom backgrounds, potentially lowering recognition accuracy.

To overcome this, we propose an improved model, referred to as VGG16-Attn, by embedding a Convolutional Block Attention Module (CBAM) at the end of each convolutional block in the original VGGNet16 (see Figure 3). The CBAM module introduces a two-stage attention mechanism: the channel attention sub-module emphasizes meaningful feature channels by learning inter-channel relationships, while the spatial attention sub-module highlights relevant regions in the feature map by learning spatial dependencies. Structurally, the modified model retains the five convolutional blocks of VGGNet16, each followed by a CBAM module, and ends with three fully connected layers for classification. This enhancement

enables the network to better capture discriminative features relevant to student behavior, improving both accuracy and robustness.

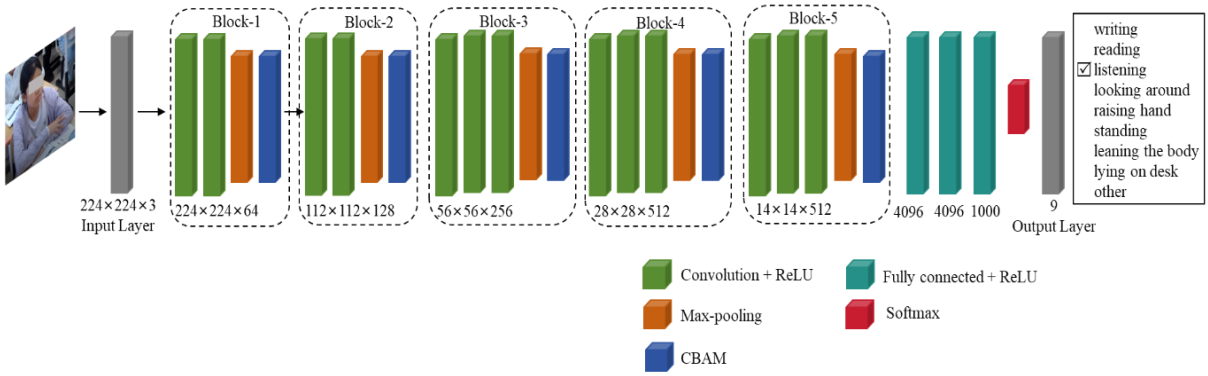


Figure 3. Network architecture of the VGG16-Attn model.

3.4 Behavioral Engagement Analysis

Chi and Wylie (2014) proposed the ICAP (Interactive, Constructive, Active, Passive) framework, which distinguishes students' engagement levels in the classroom based on their observable behaviors. This study builds on the ICAP framework to systematically categorize learning behaviors and classifies them into four types: interactive behavior, active behavior, passive behavior, and disengaged behavior. Specifically, interactive behavior emphasizes students' interaction with others (teacher or peers) in the classroom, and is demonstrated by actions such as standing up to answer questions or leaning the body to communicate. Active behavior reflects active participation and autonomous learning tendencies, such as writing or raising hands. Passive behavior refers to actions where students primarily receive information in the classroom, such as listening or reading. Disengaged behavior refers to a lack of meaningful learning participation, such as looking around or lying on the desk.

The levels of engagement for these four types of behavior are ranked from highest to lowest as follows: active behavior > interactive behavior > passive behavior > disengaged behavior. To quantitatively reflect this hierarchy, weights representing each behavior's contribution to overall engagement are assigned using the Analytic Hierarchy Process (AHP) (Tavana et al., 2023), as shown in Table 1.

Table 1. *Weight Distribution Information for Different Behavior Types*

Behavior Type	Weight	Description	Explicit Indicators
Active behavior	0.4824	Actively participate in learning, but do not involve interaction.	writing, raising hand
Interactive behavior	0.3213	It Involves collaboration and knowledge-building activities.	standing, leaning the body
Passive behavior	0.1457	Limited to information reception, indicating a lower level of engagement.	listening, reading
Disengaged behavior	0.0506	Not engaged in classroom learning.	looking around, lying on desk, other

After obtaining the behavior recognition results for all students in the entire class using the trained model, the occurrence proportion of each behavior type within each time segment (5 seconds in this study) can be calculated. The behavior engagement level for each time segment can then be computed using the following formula:

$$Behavioral\ Engagement = \sum_{i=1}^{n} (W_i \times P_i)$$

In the formula, W_i represents the weight of the i-th behavior type, P_i denotes the proportion of occurrences of the i-th behavior type, and n refers to the total number of behavior types (n=4).

4. Result and Discussion

This section presents the baseline results obtained using the dataset. The preliminary evaluation offers insights into the dataset's usability. Additionally, the trained model is applied to recognize student behaviors in real classroom videos, facilitating an analysis of the evolution of behavioral engagement.

4.1 Experimental Setup and Evaluation Metrics

The proposed model was evaluated on the self-constructed student behavior dataset, split into training, validation, and test sets in a 7:1:2 ratio. Experiments were conducted using Python with Keras 2.2.0 and TensorFlow-GPU 2.6.2. The model was trained for 50 epochs with a batch size of 32 and a learning rate of 0.01. To address class imbalance, a weighted cross-entropy loss function was employed, and parameter optimization was carried out using the Adam optimizer.

The performance of the proposed model was rigorously evaluated using four widely recognized metrics—Accuracy, Precision, Recall, and F1-score—thereby ensuring a comprehensive assessment of its predictive capabilities. All metrics were computed on the test set using macro-averaging, which ensures equal weight across all behavior categories regardless of class size. This approach enables a fair evaluation of the model's performance on both majority and minority classes.

4.2 Student Behavior Recognition Results

To comprehensively evaluate the effectiveness of the proposed VGG16-Attn model in classroom student behavior recognition, a series of comparative experiments were conducted against several representative state-of-the-art deep learning architectures. Specifically, the comparison set included InceptionV3 (Szegedy et al., 2016), Xception (Chollet, 2017), DenseNet121 (Huang et al., 2017), MobileNetV3 (Howard et al., 2017), and the original VGGNet16 (Simonyan & Zisserman, 2014). These models are widely used in computer vision tasks and exhibit distinct architectural advantages: InceptionV3 employs parallel multi-scale convolution modules to enhance feature extraction; Xception uses depthwise separable convolutions to reduce computational cost; DenseNet121 promotes feature reuse across layers to improve gradient flow and parameter efficiency; MobileNetV3 combines neural architecture search with optimized activation functions for lightweight deployment; and VGGNet16, though structurally simple, remains effective in many visual tasks due to its stable convolutional stacking.

The detailed results are summarized in Table 2. The experimental results indicate that although the proposed VGG16-Attn model has a higher computational complexity than the baseline models, it delivers the best classification performance across all core metrics. Specifically, it achieves an Accuracy of 0.91, a Macro-Precision (Macro-P) of 0.91, a Macro-Recall (Macro-R) of 0.90, and a Macro-F1 score of 0.91. Compared to the original VGGNet16, the VGG16-Attn model demonstrates a 5% improvement in accuracy, confirming the effectiveness of the CBAM attention mechanism in enhancing feature representation. In contrast, InceptionV3, Xception, and DenseNet121 exhibit lower computational costs, with Xception performing slightly better in select metrics. However, none of these models achieve an accuracy above 0.85, suggesting limited capacity to capture fine-grained behavior patterns in complex classroom scenarios. Notably, MobileNetV3 achieves the lowest computational complexity (only 0.44G FLOPs), underscoring its efficiency. Nonetheless, its classification accuracy is over 4% lower than that of VGG16-Attn, indicating that lightweight models may face challenges in extracting sufficiently rich features for recognizing complex student behaviors.

Table 2. *Comparison Results Between the VGG16-Attn and State-of-the-art Models*

Models	Accuracy	Macro-P	Macro-R	Macro-F1	FLOPs(G)
InceptionV3	0.77	0.77	0.76	0.76	5.69
Xception	0.84	0.87	0.83	0.84	9.11
DenseNet121	0.70	0.71	0.67	0.68	5.67
MobileNetV3	0.87	0.86	0.87	0.87	0.44
VGGNet16	0.86	0.87	0.86	0.86	15.47
VGG16-Attn (This study)	0.91	0.91	0.90	0.91	30.95

Figure 4 displays the results of student behavior detection in different classroom scenarios using the trained model. As shown in the figure, the majority of students' behavior states are accurately recognized, indicating that the model performs well under ideal lighting conditions and with relatively clear visibility. However, for students seated at the back or in corner positions of the classroom, the model faces challenges in behavior recognition due to limited visibility and reduced resolution.

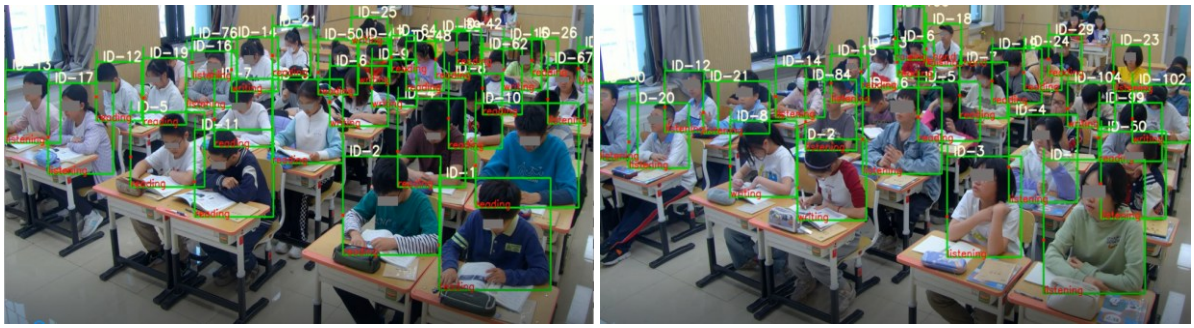


Figure 4. Examples of student behavior recognition.

4.3 Analysis of Classroom Behavioral Engagement

This study analyzed classroom behavioral engagement using video recordings from two class sessions. Prior to data collection, informed consent was obtained, and students were clearly informed about the purpose of the study. All data were anonymized to ensure student privacy and comply with ethical standards.

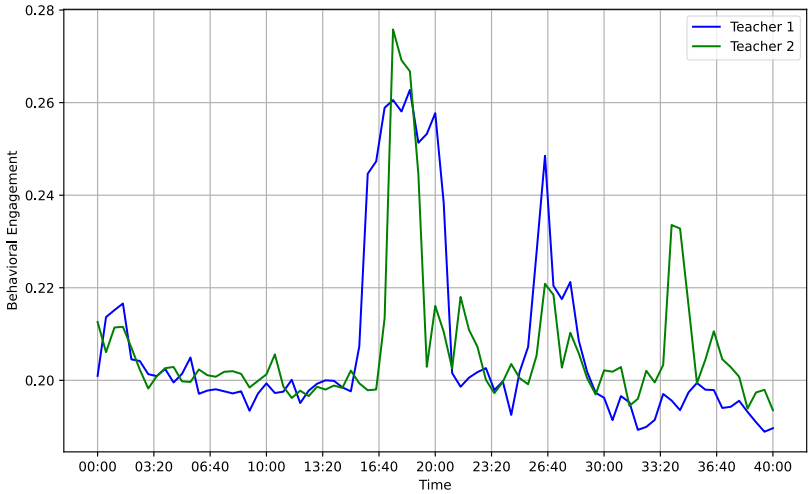


Figure 5. Temporal evolution of student behavioral engagement in two classrooms.

Aggregated behavior recognition results revealed distinct engagement patterns (Figure 5): both classes showed low engagement during the first 15 minutes, with noticeable increases

around the 16th and 27th minutes. However, in the final 10 minutes, engagement diverged—remaining low in Teacher 1's class, while fluctuating in Teacher 2's class. Video analysis indicates that low engagement corresponded to passive lecture segments, whereas peaks aligned with interactive activities such as practice and questioning. Despite identical instructional content, the engagement trends reflect differences in teaching pace and style. Teacher 2's class included more frequent but shorter practice sessions, consistent with lesson evaluation notes stating that "the pace of the lesson was faster, and when students made mistakes, the teacher quickly moved on without targeted feedback". While this approach covered more content, it limited opportunities for in-depth learning. These findings are consistent with classroom observations and further validate the effectiveness of the proposed automatic behavioral engagement recognition method.

5. Conclusions and Future Work

This study integrates computer vision technology with an improved deep learning model to enable automated assessment of classroom behavioral engagement, aiming to address the limitations of traditional methods—namely, subjectivity, lack of real-time feedback, and poor scalability. Furthermore, the proposed approach was applied to real classroom video analysis, and the experimental results confirmed its effectiveness and reliability in practical settings. This provides a scientific foundation for teachers to dynamically monitor students' learning processes.

Although this study successfully achieves automatic recognition of classroom behavioral engagement using classroom video data, it has not yet focused on more implicit aspects of engagement, including emotion and cognition. Future research should integrate other sources of data, such as classroom dialogues, and further develop a more comprehensive, multimodal approach to automated learning engagement assessment. Additionally, future work should explore how to better integrate the proposed method into teaching practices to enhance educational efficiency and quality, thereby contributing to the development of educational informatization and intelligence.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No: 62277006, 62267008).

References

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4), 219-243. https://doi.org/10.1080/00461520.2014.965823

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In Proceedings of *the IEEE Conference on Computer Vision and Pattern Recognition*, 1251-1258. https://doi.org/10.48550/arXiv.1610.02357

Cooper, K. S. (2014). Eliciting engagement in the high school classroom: A mixed-methods examination of teaching practices. *American Educational Research Journal*, 51(2), 363-402. https://www.jstor.org/stable/24546691

Finn, J. D. (1989). Withdrawing from school. Review of Educational Research, 59(2), 117-142. https://doi.org/10.2307/1170412

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, 74(1), 59-109. https://doi.org/10.3102/00346543074001059

Gunuc, S. (2014). The relationships between student engagement and their academic achievement. *International Journal on New Trends in Education and their Implications*, 5(4), 216-231. https://doi.org/10.1007/s40299-013-0095-8

Guo, Q. (2022). System analysis of the learning behavior recognition system for students in a law classroom: Based on the improved SSD behavior recognition algorithm. *Scientific Programming*, 2022(1), 3525266. https://doi.org/10.1155/2022/3525266

Howard, A. G. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of *the IEEE Conference on Computer Vision and Pattern Recognition*, 4700-4708. https://doi.org/10.48550/arXiv.1608.06993

Ikram, S., Ahmad, H., Mahmood, N., Faisal, C. N., Abbas, Q., Qureshi, I., & Hussain, A. (2023). Recognition of student engagement state in a classroom environment using deep and efficient transfer learning algorithm. *Applied Sciences*, 13(15), 8637. https://doi.org/10.3390/app13158637

Karimah, S. N., & Hasegawa, S. (2022). Automatic engagement estimation in smart education/learning settings: A systematic review of engagement definitions, datasets, and methods. *Smart Learning Environments*, 9(1), 31. https://doi.org/10.1186/s40561-022-00212-y

Li, Y., Qi, X., Saudagar, A. K. J., Badshah, A. M., Muhammad, K., & Liu, S. (2023). Student behavior recognition for interaction detection in the classroom environment. *Image and Vision Computing*, 136, 104726. https://doi.org/10.1016/j.imavis.2023.104726

Ngoc Anh, B., Tung Son, N., Truong Lam, P., Phuong Chi, L., Huu Tuan, N., Cong Dat, N., ... & Van Dinh, T. (2019). A computer-vision based application for student behavior monitoring in classroom. *Applied Sciences*, 9(22), 4729. https://doi.org/10.3390/app9224729

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. https://doi.org/10.48550/arXiv.1409.1556

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of *the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818-2826. https://doi.org/10.48550/arXiv.1512.00567

Tavana, M., Soltanifar, M., & Santos-Arteaga, F. J. (2023). Analytical hierarchy process: Revolution and evolution. *Annals of Operations Research*, 326(2), 879-907. https://doi.org/10.1007/s10479-021-04432-2

Trabelsi, Z., Alnajjar, F., Parambil, M. M. A., Gochoo, M., & Ali, L. (2023). Real-time attention monitoring system for classroom: A deep learning approach for student's behavior recognition. *Big Data and Cognitive Computing*, 7(1), 48. https://doi.org/10.3390/bdcc7010048

Volpe, R. J., DiPerna, J. C., Hintze, J. M., & Shapiro, E. S. (2005). Observing students in classroom settings: A review of seven coding schemes. *School Psychology Review*, 34(4), 454-474. https://doi.org/10.1080/02796015.2005.12088009

Xiong, Y., Xinya, G., & Xu, J. (2024). CNN-Transformer: A deep learning method for automatically identifying learning engagement. *Education and Information Technologies*, 29(8), 9989-10008. https://doi.org/10.1007/s10639-023-12058-z

Zaletelj, J., & Košir, A. (2017). Predicting students' attention in the classroom from Kinect facial and body features. *EURASIP Journal on Image and Video Processing*, 2017, 1-12. https://doi.org/10.1186/s13640-017-0228-8

Evaluating the Effectiveness of Large Language Models for Course Recommendation Tasks

Boxuan MA[a*], Md Akib Zabed KHAN[b], Tianyuan YANG[a], Agoritsa POLYZOU[b] & Shin'ichi KONOMI[a]
[a]*Kyushu University, Japan*
[b]*Florida International University, United States*
*boxuan@artsci.kyushu-u.ac.jp

Abstract: Large Language Models (LLMs) have made significant strides in natural language processing and are increasingly being integrated into recommendation systems. However, their potential in educational recommendation systems has yet to be fully explored. This paper investigates the use of LLMs as a general-purpose recommendation model, leveraging their vast knowledge derived from large-scale corpora for course recommendation tasks. We explore prompt and fine-tuning methods for LLM-based course recommendation and compare their performance against traditional recommendation models. Extensive experiments were conducted on a real-world MOOC dataset, evaluating using LLMs as course recommendation systems across a variety of key dimensions. Our results demonstrate that fine-tuned LLMs can achieve good performance comparable to traditional models, highlighting their potential to enhance educational recommendation systems. These findings pave the way for further exploration and development of LLM-based approaches in the context of educational recommendations.

Keywords: Course Recommendation, Large Language Models, MOOC Courses

1. Introduction

Course recommendation systems are increasingly used in the field of education and have become an essential tool in addressing information overload and enhancing user experience for learning (Ma et al., 2020). They can offer personalized course suggestions that align with a student's interests, career goals, or skill development needs and can enhance the educational experience by helping students navigate the vast array of available courses and make more informed decisions about their learning journey (Jiang et al., 2019).

Over the past decade, significant advancements have been made in course recommendation technologies. Traditional recommendation models, including collaborative filtering and content-based methods, have long been employed in practical settings. These approaches typically rely on user-item interaction data or explicit features to provide personalized recommendations (Ma et al., 2021). While successful, traditional recommendation models face notable limitations, such as a lack of generalization and the need for task-specific data for training (Ma et al., 2024). On the other hand, recent deep learning models have demonstrated considerable potential in enhancing prediction accuracy, but they require extensive training and often suffer from a lack of explainability, making them less transparent to users (Dai et al., 2023).

Large Language Models, such as ChatGPT, have gained significant attention due to their great performance for natural language processing tasks, such as text generation, question answering, and language comprehension. These models, with their adaptability and vast knowledge derived from large-scale corpora, present an appealing opportunity for recommendation systems. Previous research has indicated LLMs can be directly used as recommendation systems with prompts, and a growing body of research has begun to explore the potential of LLMs in recommendation tasks, evaluating their performance across various

recommendation scenarios and datasets from different domains (Di et al., 2023; Dai et al., 2023; Liu et al., 2023). On the other hand, many researchers have started using LLMs as part of recommendation systems to enhance their performance, such as through feature extraction, feature augmentation, or knowledge representation.

Despite the rapid development of LLMs, most research on LLM-based recommendation systems has focused on domains like music, movies, and books. There has been limited research on applying LLMs specifically to course recommendations within the context of Massive Open Online Courses (MOOCs), and whether LLMs can perform well on course recommendation tasks remains an open question. Therefore, this paper aims to bridge this gap by evaluating the effectiveness of LLMs in recommending courses based on user learning history. Our study offers a comparative analysis between LLMs and traditional recommendation models and investigates the promise of LLMs in addressing key challenges in educational recommendation systems.

2. Related Work

2.1 Course Recommendation

The rise of Massive Open Online Courses (MOOCs) and the increasing number of students have led to the widespread application of course recommendation systems (Ma et al., 2021). Since the introduction of the first course recommendation system based on constraint satisfaction (Parameswaran et al., 2011), various methods have been developed. Content-based approaches recommend courses by matching students' interests with course descriptions and content (Morsomme et al., 2019; Morsy et al., 2019). Matrix Factorization (MF) techniques have also been applied to course recommendation, particularly for predicting future course selections based on students' past courses and grades (Elbadrawy et al., 2015; Sweeney et al., 2016). Other methods have explored the mining of historical course enrollment data to uncover relationships and patterns. For example, Aher et al. (2013) employed association rule mining combined with clustering to identify course relationships, recommending courses based on historical enrollment patterns. Similarly, Bendakir et al. (2006) used association rules with user ratings to enhance the recommendation results, while Polyzou et al. (2019) introduced Scholars Walk, which captures sequential course relationships through a random-walk approach.

As deep learning techniques have gained popularity, they have also been applied to course recommendation systems (Yang et al., 2025; Zhang et al., 2019; Jiang et al., 2019). For instance, Pardos et al. (2019a) modified the skip-gram model to generate course vectors from historical course enrollment data, which are then used to recommend courses similar to a student's previously favored courses. In a similar vein, they proposed the course2vec model, which employs a neural network to generate course recommendations by taking multiple courses as input and predicting a probability distribution over potential course selection (Pardos et al., 2019b).

2.2 LLMs for Recommendation

LLMs have demonstrated their adaptability and significant improvements in a wide range of NLP tasks by leveraging the extensive knowledge from large-scale corpora. Inspired by its successes, there has been a growing interest in applying LLMs to recommendation systems. Recent works have leveraged prompt-based techniques to transform recommendation tasks into natural language tasks, utilizing LLMs without task-specific fine-tuning. For instance, LMRecSys (Zhang et al., 2021) and P5 (Geng et al., 2022) focus on converting recommendation tasks into multi-token cloze tasks using prompts to tackle zero-shot and data efficiency issues. GPT4Rec (Li et al., 2023) and M6-Rec (Cui et al., 2022) utilize LLMs to learn both item and user embeddings. Liu et al. (2023) evaluated ChatGPT's performance on five recommendation scenarios, including rating prediction, sequential recommendation, direct recommendation, explanation generation, and review summarization. Dai et al. (2023) investigated ChatGPT's ranking capabilities, including point-wise, pair-wise, and list-wise

ranking. Moreover, the ability of LLMs has also been explored in cold-start scenarios where few user interaction data are available (Wu et al., 2023).

Besides the direct use of LLMs as recommendation systems, LLMs are increasingly being used as components to enhance traditional recommendation models. These approaches integrate LLMs into existing systems through feature extraction, feature augmentation, knowledge representation, and ranking functions (Wu et al., 2024). For instance, Gao et al. (2023) are among the first to use ChatGPT to augment traditional recommender systems by injecting user preferences into the recommendation process through conversational interaction. Another example, Zhang et al. (2023) enhance recommendation system with LLMs by designing prompts for different recommendation settings, where LLM takes candidates from a Recall model for re-ranking.

Despite the growing body of research on LLM-based recommendation systems in different domains, to our knowledge, there has been limited research on applying LLMs specifically to course recommendations aside from work by Khan et al. (2022), while they did not focus on the evaluation of LLMs' potential and only used local models. Yang et al. (2024a, 2024b) use LLMs to generate knowledge concepts from course descriptions and provide course recommendations based on generated concepts. However, whether LLMs can perform well on course recommendation tasks remains an open question.

3. Methods

The workflow for using LLMs in course recommendation tasks is shown in Figure 1. We explore two approaches for applying LLMs to course recommendations. The first approach involves using pre-trained LLMs as recommendation systems, where they generate recommendations directly based on prompts. The second approach involves fine-tuning the model, enriching its knowledge base with student interaction data, and generating recommendations based on the fine-tuned model. The examples of course information, training set, prompt and training instance for fine-tuning can also be seen in Figure 1.

3.1 Direct Use of LLMs for Recommendation

As shown in Figure 1-B, we directly use LLMs to generate recommendations without re-training or fine-tuning the model. Instead, we craft prompts and feed them into the LLMs. The model then generates recommendation results based on the instructions provided in prompts.

For *zero-shot* recommendation (Figure 1-B-1), we provide the LLM with knowledge about all available courses (including course IDs, names, and descriptions), along with the student's prior course registration history as prompt input. The LLM is tasked with recommending a set of courses based on this input. In the case of *few-shot* (Figure 1-B-2), we incorporate additional training data as knowledge context and prompt the LLM to recommend courses based on the complete set of provided prompts. To mitigate hallucination issues and address token limitations, we represent courses using only their IDs in our prompt design.

3.2 Using Fine-tuned LLMs for Recommendation

We also fine-tune LLMs to enhance their knowledge with historical data relevant to our task. We fine-tune two open-source models, Llama-3 (Touvron et al., 2023) and GPT-2 (Radford et al., 2019), as they are freely available and easy to use. Following prior work in item recommendation (Liu et al., 2023), we use students' course enrollment histories to fine-tune the LLMs, enabling them to capture historical enrollment patterns. After fine-tuning, we provide prompts that include a student's prior course registration history and ask the fine-tuned models to recommend a set of courses (see Figure 1-C).

We use <|*user*|> and <|*assistant*|> tokens to indicate the input and output in each training instance, where the input is the student's prior course list, and the output is the subsequent courses they are likely to take. Additionally, we include course descriptions as input to help the model capture the semantic similarity between courses.

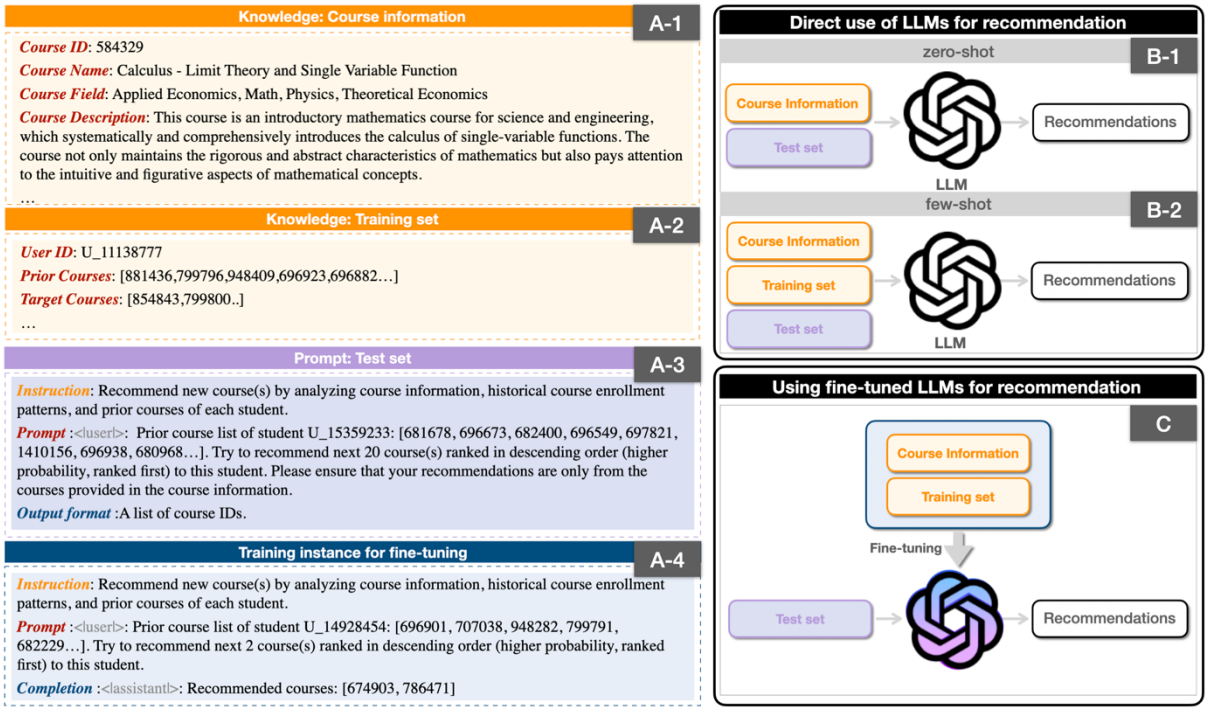


Figure 1. Workflow of Using LLMs to Perform Course Recommendation Tasks. A-1: Example Knowledge for Course Information. A-2: Example Knowledge for Training Set. A-3: Example Prompt of Course Recommendation Task (Test Set). A-4: Example of A Training Instance for Fine-Tuning. B-1: Direct Use of LLMs for Recommendation (Zero-Shot). B-2: Direct Use of LLMs for Recommendation (Few-Shot). C: Using Fine-Tuned LLMs for Recommendation.

4. Evaluation

4.1 Dataset

In the context of this work, we focus on the scenario of course recommendation within a MOOC environment. The dataset MOOCCubeX (Yu et al., 2021) in our analysis is collected from the XuetangX[1], one of the largest MOOC websites in China. This dataset consists of 4,216 courses and 3,330,294 students.

4.2 Evaluation Baselines and Metrics

We explore two approaches for using LLMs in course recommendation: direct prompting and fine-tuning with student interaction data. For direct prompting, we use GPT4-turbo and GPT4o due to their popularity and affordability. Following Khan et al. (2022), we fine-tune two open-source models: Llama-3 (Touvron et al., 2023) and GPT2 (Radford et al., 2019).

We also compared LLMs to following traditional baselines: **Random**, recommend random items. **Pop** (Elbadrawy et al., 2016), recommend most popular items. **PMF** (Salakhutdinov et al., 2007), probabilistic matrix factorization relies solely on the user-item interaction matrix. NMF (Lee et al., 1999), non-negative matrix factorization factorizes a non-negative matrix into the product of two or more non-negative matrices based on the user-item interaction matrix. **Item-based KNN** (Sarwar et al., 2001), models user and item based on item similarity obtained by interaction information. **User-based KNN** (Sarwar et al., 2001), models user and item based on user similarity obtained by interaction information. **KEAM** (Yang et al., 2024a), a knowledge graph-based autoencoder model that use both user-item interactions and course concepts.

[1] http://www.xuetangx.com

Table 1. *Accuracy Performance Comparison (%)*

Model	K = 5					K = 10				
	HR	**Recall**	**Precision**	**F1**	**nDCG**	**HR**	**Recall**	**Precision**	**F1**	**nDCG**
Random	0.100	0.005	0.020	0.010	0.020	0.610	0.009	0.080	0.090	0.080
GPT4-turbo zero-shot	0.210	0.100	0.040	0.060	0.050	0.410	0.310	0.040	0.070	0.130
GPT4o zero-shot	0.405	0.080	0.080	0.075	0.075	0.815	0.185	0.080	0.110	0.110
Item-based KNN	0.510	0.070	0.100	0.080	0.125	1.225	0.400	0.130	0.195	0.215
GPT4o few-shot	0.800	0.400	0.160	0.230	0.230	0.800	0.400	0.080	0.130	0.230
GPT4-turbo few-shot	1.002	0.113	0.201	0.147	0.015	1.000	0.110	0.100	0.110	0.100
PMF	1.630	0.285	0.325	0.304	0.435	3.370	0.680	0.340	0.425	0.515
NMF	1.630	0.285	0.325	0.304	0.435	3.370	0.680	0.340	0.425	0.515
User-based KNN	2.960	1.080	0.595	0.755	0.955	4.595	1.645	0.460	0.715	1.100
Pop	8.195	3.300	1.680	2.215	2.680	15.165	5.950	1.660	2.580	3.640
KEAM	13.200	6.053	2.840	3.866	5.977	20.800	9.643	2.300	3.714	7.514
GPT2 Fine-tuning	16.903	7.438	3.524	4.782	11.492	22.643	9.560	2.452	3.903	13.498
Llama3 Fine-tuning	**21.677**	**12.434**	**4.852**	**6.980**	**15.266**	**28.857**	**15.166**	**3.424**	**5.587**	**17.496**
Model	**K = 15**					**K = 20**				
	HR	**Recall**	**Precision**	**F1**	**nDCG**	**HR**	**Recall**	**Precision**	**F1**	**nDCG**
Random	1.230	0.150	0.090	0.110	0.110	1.435	0.190	0.070	0.100	0.115
GPT4-turbo zero-shot	0.410	0.310	0.030	0.050	0.130	1.230	0.940	0.060	0.120	0.280
GPT4o zero-shot	1.225	0.450	0.085	0.135	0.185	1.225	0.450	0.060	0.105	0.185
Item-based KNN	2.450	0.590	0.175	0.270	0.320	3.165	0.845	0.175	0.295	0.385
GPT4o few-shot	0.800	0.400	0.500	0.444	0.230	0.800	0.400	0.400	0.400	0.230
GPT4-turbo few-shot	1.000	0.110	0.070	0.080	0.100	3.000	0.118	0.150	0.270	0.370
PMF	5.205	1.465	0.345	0.560	0.885	5.100	0.980	0.260	0.405	0.620
NMF	5.205	1.465	0.345	0.560	0.885	5.100	0.980	0.260	0.405	0.620
User-based KNN	6.530	2.335	0.455	0.760	1.335	8.165	2.830	0.440	0.760	1.530
Pop	17.515	6.655	1.305	2.175	3.865	21.920	8.270	1.295	2.240	4.425
KEAM	26.700	12.229	2.053	3.516	8.667	32.000	14.830	1.915	3.392	9.752
GPT2 Fine-tuning	26.622	10.896	1.990	3.365	14.296	30.793	12.799	1.896	3.303	15.204
Llama3 Fine-tuning	**34.008**	**17.193**	**2.793**	**4.805**	**18.693**	**38.294**	**19.399**	**2.393**	**4.261**	**19.709**

To get a comprehensive evaluation that sheds light on LLMs' performance in the course recommendation task, we select the different metrics following previous works (Dai et al., 2023; Di et al., 2023; Liu et al., 2023). For accuracy metrics, we utilize metrics including ***HR*** (Hit Ratio), ***Recall***, ***Precision***, ***F1***, and ***nDCG*** (normalized Discounted Cumulative Gain). For coverage and novelty metrics, we select ***Coverage*** (quantifies the coverage of all available items that can potentially be recommended), ***Gini Index*** (assesses the distribution of items) and ***EPC*** (Expected Popularity Complement, measures the expected number of relevant recommended items that were not previously seen by the user, showing the model's ability to introduce novelty in the recommendations). Higher scores in these metrics indicate better recommendations.

4.3 Implementation Details

We first split the dataset based on user history, using 80% of the user interaction data for training and 20% for testing. After processing, we randomly sampled 1000 records from the test set for evaluation due to token limitations and expensive costs. For each user, we input their previously interacted items in order and use the LLM to recommend a list of course IDs they might interact with next.

For the pre-trained models, we access GPT4-turbo and GPT4o using OpenAI's API. We also fine-tuned Llama-3-8B and GPT-2-1.5B models. Following previous work (Khan et al., 2022), the Llama-3 model is tokenized using Autotokenizer, and the GPT-2 model is tokenized using the GPT-2tokenizer. We use BitsandBytes[2] and PEFT[3] libraries for fine-tuning efficiently by reducing computing requirements. The PEFT library offers a LoraConfig class to enable QLoRA (Dettmers et al., 2023), which combines quantization with low-rank adapters. By quantizing the base LLM to 4 bits per parameter, we drastically reduce its memory footprint. QLoRA then freezes all original model weights and injects a small set of trainable, low-rank adapter matrices. This approach lets us fine-tune large models using far less GPU memory and compute resources.

[2] https://github.com/bitsandbytes-foundation/bitsandbytes
[3] https://github.com/huggingface/peft

Table 2. *Diversity and Novelty Performance Comparison (%)*

Model	K = 5			K = 10		
	Coverage	**Gini Index**	**EPC**	**Coverage**	**Gini Index**	**EPC**
Pop	0.160	80.000	3.410	0.320	90.000	4.505
PMF	0.220	80.455	0.945	0.395	90.175	1.205
NMF	0.220	80.455	0.945	0.395	90.175	1.205
GPT4-turbo zero-shot	1.140	96.050	0.050	1.910	97.860	0.090
Item-based KNN	6.035	96.375	0.305	9.675	97.775	0.400
User-based KNN	6.270	96.850	3.435	9.295	98.175	1.650
GPT4-turbo few-shot	14.872	99.771	0.250	27.540	99.871	0.250
GPT4o zero-shot	17.375	97.720	0.115	32.910	98.640	0.170
KEAM	12.048	95.560	8.115	18.554	97.466	9.257
Llama3 Fine-tuning	22.505	99.696	**14.399**	31.006	99.692	**15.991**
GPT2 Fine-tuning	23.004	95.896	10.393	25.307	95.699	11.301
GPT4o few-shot	32.400	99.890	0.230	53.750	99.890	0.230
Random	**54.160**	**99.930**	0.090	**78.795**	**99.950**	0.150
Model	**K = 15**			**K = 20**		
	Coverage	**Gini Index**	**EPC**	**Coverage**	**Gini Index**	**EPC**
Pop	0.480	93.330	4.800	0.640	95.000	5.180
PMF	0.585	93.525	1.350	0.760	95.095	1.335
NMF	0.585	93.525	1.350	0.760	95.095	1.335
GPT4-turbo zero-shot	1.910	97.860	0.090	2.890	98.740	0.130
Item-based KNN	12.485	98.375	0.690	12.485	98.375	0.690
User-based KNN	12.005	98.735	1.825	14.280	99.020	1.955
GPT4-turbo few-shot	37.740	99.900	0.250	47.010	99.900	0.360
GPT4o zero-shot	48.810	98.640	0.170	65.090	99.230	0.200
KEAM	22.988	98.217	9.867	26.747	98.614	10.284
Llama3 Fine-tuning	36.735	99.832	**16.710**	39.802	99.807	**17.198**
GPT2 Fine-tuning	25.894	96.004	11.798	26.092	96.505	12.291
GPT4o few-shot	68.960	99.940	0.230	79.030	99.950	0.230
Random	**90.580**	**99.955**	0.265	**95.900**	**99.960**	0.195

5. Results

5.1 Recommendation Performance

To assess the recommendation capability of large language models (LLMs), we conducted experiments comparing pre-trained and fine-tuned LLMs with traditional models. Table 1 presents the results in percentages.

In summary, we found that the performance of LLMs in the zero-shot prompting setup was relatively low compared to baseline models, outperforming only random recommendation approaches. In contrast, the few-shot prompting setup generally yielded better results, suggesting that providing historical enrollment data helps LLMs identify enrollment patterns and improve recommendation accuracy. However, overall, pre-trained LLMs still fall short of traditional recommendation methods. As a result, relying solely on LLMs for sequential recommendation tasks may not be optimal. Further research is needed to integrate additional guidance and constraints to help LLMs accurately capture historical user interests and produce meaningful recommendations. However, we found that fine-tuned Llama3 consistently outperformed all other methods across different K values, while fine-tuned GPT-2 surpassed traditional baselines and achieved performance similar to KEAM, which is the state-of-the-art model. These results suggest that fine-tuning enables LLMs to specialize in the recommendation task, allowing them to more effectively model user behavior and course relationships and thereby produce more accurate predictions.

5.2 Diversity and Novelty Performance

We also aim to assess the extent of diversity (**Coverage, Gini Index**) and novelty (**EPC**) in the recommendations generated by LLMs. Table 2 presents the results in percentages.

Overall, the Random model achieves the highest Coverage and Gini Index across all K values, outperforming all other models. This result is not surprising, given its random nature, which ensures a broad range of items are recommended. In contrast, LLMs perform relatively better in diversity and novelty dimensions. Notably, the advanced model, GPT4o, outperforms

Table 3. *Performance Comparison (%) on Cold Start Scenario*

	K = 5							
Model	**HR**	**Recall**	**Precision**	**F1**	**nDCG**	**Coverage**	**Gini Index**	**EPC**
Random	0.000	0.000	0.000	0.000	0.000	**53.080**	**99.930**	0.000
PMF	0.000	0.000	0.000	0.000	0.000	0.190	80.190	0.000
NMF	0.000	0.000	0.000	0.000	0.000	0.190	80.190	0.000
GPT4-turbo	0.200	0.200	0.040	0.070	0.090	33.290	98.710	0.050
Item-based KNN	0.430	0.430	0.090	0.140	0.220	11.660	97.280	0.150
User-based KNN	0.430	0.430	0.090	0.140	0.430	11.280	98.650	0.430
GPT4o	1.080	1.080	0.220	0.360	0.820	23.630	98.710	0.740
Pop	4.730	4.730	0.950	1.580	3.030	0.160	80.000	2.470
GPT2 Fine-tuning	4.409	4.409	0.882	1.470	3.030	9.037	83.844	1.920
KEAM	7.800	7.800	1.560	2.600	6.406	34.294	97.108	4.933
Llama3 Fine-tuning	**13.613**	**13.613**	**2.723**	**4.538**	**11.810**	7.795	83.029	**11.473**
	K = 10							
Model	**HR**	**Recall**	**Precision**	**F1**	**nDCG**	**Coverage**	**Gini Index**	**EPC**
Random	0.220	0.220	0.020	0.040	0.060	**77.570**	**99.950**	0.020
PMF	0.000	0.000	0.000	0.000	0.000	0.350	90.100	0.000
NMF	0.000	0.000	0.000	0.000	0.000	0.350	90.100	0.000
GPT4-turbo	0.200	0.200	0.200	0.400	0.090	61.750	99.240	0.050
Item-based KNN	1.080	1.080	0.110	0.200	0.650	15.410	99.130	0.520
User-based KNN	1.720	1.720	0.170	0.310	0.650	18.460	98.310	0.310
GPT4o	1.510	1.510	0.150	0.270	0.960	41.520	99.240	0.800
Pop	6.670	6.670	0.670	1.210	3.640	0.320	90.000	2.710
GPT2 Fine-tuning	7.214	7.214	0.721	1.312	3.420	9.156	90.953	2.266
KEAM	11.000	11.000	1.100	2.000	7.945	56.292	98.484	5.346
Llama3 Fine-tuning	**16.515**	**16.515**	**1.651**	**3.003**	**12.632**	7.795	90.631	**11.904**

GPT4-turbo model across all dimensions. Upon a closer examination of the generated recommendations, we observed that GPT4-turbo often exhibits “lazy behaviors”, generating similar or repetitive recommendation lists, which leads to low diversity. Furthermore, few-shot models consistently outperform zero-shot models. By learning from user-specific data, few-shot models can generate more personalized and relevant recommendations, enhancing both diversity and novelty. Finally, fine-tuned models like GPT-2 and Llama3 not only excel in diversity but also significantly surpass other models in novelty. This indicates that directly applying LLMs to recommendation tasks is challenging because the data used for pre-training LLMs differs from the specific requirements of recommendation tasks. However, fine-tuned LLMs, when trained on specific data, can deliver better results.

5.3 Cold Start Scenario

Cold start is a well-known challenge in course recommendation systems, especially in MOOC environments. It refers to the difficulty of recommending relevant courses to new users who lack sufficient interaction data. To investigate the performance of LLMs in cold start scenarios for course recommendations, we adopt two different experiments inspired by previous studies (Di et al., 2023; Dai et al., 2023).

First, we identify cold-start users by dividing the users of the dataset into quartiles based on their historical interaction data. The lower quartile, representing users with the least interaction, is selected as the subset of cold-start users. This allows us to evaluate the models under consistent cold-start conditions, ensuring that all models are tested with a similar subset of users (note that we fine-tuned our LLMs using the same cold-start training set). The results of this evaluation are presented in Table 3. We observe that off-the-shelf LLMs outperform traditional models for cold start scenarios when only limited training data is available. LLMs do not require extensive training data to function as recommendation systems, as their pre-trained knowledge allows them to make informed predictions. The Pop method, by contrast, performs well because it simply recommends the most popular courses in the dataset. Moreover, with minimal training data, fine-tuned Llama3 achieved the highest scores, while fine-tuned GPT-2 outperformed Pop and ranked second only to KEAM. This demonstrates that the reasoning abilities and extensive knowledge encoded in LLMs enable them to generate superior recommendations.

Secondly, we investigate the amount of training data required for traditional recommendation models to achieve performance comparable to or better than LLMs.

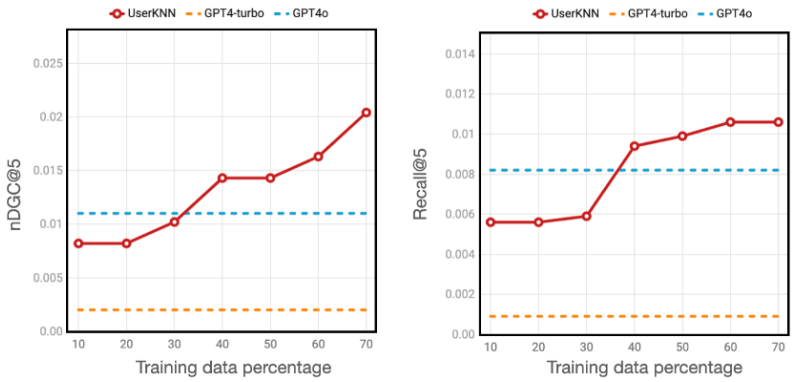


Figure 2. Comparison with UserKNN Using Different Percentages of Training Data.

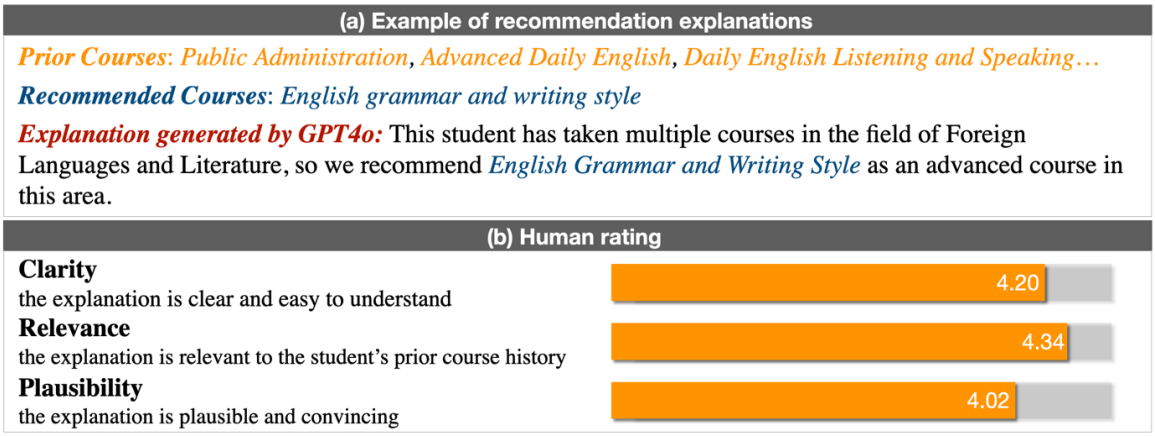


Figure 3. (a) Example of Explanation Generated by LLMs (b)Human Rating Results.

Specifically, we chose the User-based KNN model as it performed well in the first experiment. We then evaluated their performance after training on varying proportions of training data and compared their performance to that of LLMs. *Recall@5* and *nDCG@5* are reported in Figure 2. As expected, the performance of User-based KNN improves with increasing amounts of training data. Also, we can observe that although GPT4-turbo's performance is not good, direct use of GPT4o as a recommendation system without training data still outperforms User-based KNN that trained on few data, i.e., less than 30%. Based on these findings, we conclude that using LLMs as course recommendation systems is a promising approach for mitigating the cold-start problem, offering effective solutions when traditional methods may struggle.

5.4 Explanation

Providing students with clear and understandable explanations for course recommendations can enhance the transparency and perceived trustworthiness of the system, thereby improving overall user satisfaction (Ma et al., 2021; Ma et al., 2024). We also prompt LLMs to produce a textual explanation for each recommended course.

To assess the quality of these explanations, we conducted a small-scale human evaluation on 50 sampled outputs. Each explanation was rated along three dimensions, including ***Clarity*** (the explanation is clear and easy to understand), ***Relevance*** (the explanation is relevant to the student's prior course history), and ***Plausibility*** (the explanation is plausible and convincing), using a 5-point Likert scale by two human researchers, the example explanation and human rate results are shown in Figure 3. Most explanations received scores of 4 or higher across all dimensions, suggesting that the explanations were generally interpretable and credible. These findings indicate that LLMs can effectively generate intuitive and trustworthy justifications by contextualizing recommendations within students past learning experiences.

6. Conclusions

In this paper, we evaluate the performance of LLMs in course recommendation tasks and compare them with traditional recommendation methods across diverse settings. Our results indicate that LLMs without fine-tuning perform worse than baseline models. However, fine-tuned LLMs significantly outperform traditional approaches, particularly in cold-start scenarios. Moreover, LLMs demonstrate strong capabilities in generating high-quality recommendation explanations. These findings provide valuable insights into the strengths and limitations of LLMs in educational recommendation systems.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP24K20903 and Dissertation Fellowship awarded by the University Graduate School of Florida International University.

References

Aher, S.B., Lobo, L.: Combination of machine learning algorithms for recommendation of courses in e-learning system based on historical data. *Knowledge-Based Systems* 51, 1–14 (2013)

Bendakir, N., Aïmeur, E.: Using association rules for course recommendation. In: *Proceedings of the AAAI workshop on educational data mining*. vol. 3, pp. 1–10 (2006)

Cui, Z., Ma, J., Zhou, C., Zhou, J., Yang, H.: M6-rec: Generative pretrained language models are open-ended recommender systems. *arXiv preprint* arXiv:2205.08084 (2022)

Dai, S., Shao, N., Zhao, H., Yu, W., Si, Z., Xu, C., Sun, Z., Zhang, X., Xu, J.: Uncovering chatgpt's capabilities in recommender systems. In: *Proceedings of the 17th ACM Conference on Recommender Systems*. pp. 1126–1132 (2023)

Di Palma, D., Biancofiore, G.M., Anelli, V.W., Narducci, F., Di Noia, T., Di Sciascio, E.: Evaluating chatgpt as a recommender system: A rigorous approach. *arXiv preprint* arXiv:2309.03613 (2023)

Elbadrawy, A., Karypis, G.: Domain-aware grade prediction and top-n course recommendation. In: *Proceedings of the 10th ACM conference on recommender systems*. pp. 183–190 (2016)

Elbadrawy, A., Studham, R.S., Karypis, G.: Collaborative multi-regression models for predicting students' performance in course activities. In: *proceedings of the fifth international conference on learning analytics and knowledge*. pp. 103–107 (2015)

Gao, Y., Sheng, T., Xiang, Y., Xiong, Y., Wang, H., Zhang, J.: Chat-rec: Towards interactive and explainable llms-augmented recommender system. *arXiv preprint* arXiv:2303.14524 (2023)

Geng, S., Liu, S., Fu, Z., Ge, Y., Zhang, Y.: Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In: *Proceedings of the 16th ACM Conference on Recommender Systems*. pp. 299–315 (2022)

Jiang, W., Pardos, Z.A., Wei, Q.: Goal-based course recommendation. In: *Proceedings of the 9th international conference on learning analytics & knowledge*. pp. 36–45 (2019)

Khan, M.A.Z., Polyzou, A., Bennamane, N.: How can we use llms for edm tasks? the case of course recommendation (2022)

Lee, D., Seung, H.: Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 788–91 (11 1999). https://doi.org/10.1038/44565

Li, J., Zhang, W., Wang, T., Xiong, G., Lu, A., Medioni, G.: Gpt4rec: A generative framework for personalized recommendation and user interests interpretation. *arXiv preprint* arXiv:2304.03879 (2023)

Liu, J., Liu, C., Zhou, P., Lv, R., Zhou, K., Zhang, Y.: Is chatgpt a good recommender? a preliminary study. *arXiv preprint* arXiv:2304.10149 (2023)

Liu, J., Liu, C., Zhou, P., Ye, Q., Chong, D., Zhou, K., Xie, Y., Cao, Y., Wang, S., You, C., et al.: Llmrec: Benchmarking large language models on recommendation task. *arXiv preprint* arXiv:2308.12241 (2023)

Ma, B., Lu, M., Taniguchi, Y., Konomi, S.: Exploration and explanation: An interactive course recommendation system for university environments. In: *IUI Workshops* (2021)

Ma, B., Lu, M., Taniguchi, Y., Konomi, S.: CourseQ: the impact of visual and interactive course recommendation in university environments. *Research and practice in technology enhanced learning* 16, 1–24 (2021)

Ma, B., Lu, M., Taniguchi, Y., Konomi, S.: Investigating course choice motivations in university environments. *Smart Learning Environments* 8(1), 1–18 (2021)

Ma, B., Taniguchi, Y., Konomi, S.: Course recommendation for university environments. *International educational data mining society* (2020)
Ma, B., Yang, T., Ren, B.: A survey on explainable course recommendation systems. In: Streitz, N.A., Konomi, S. (eds.) *Distributed, Ambient and Pervasive Interactions*. pp. 273–287. Springer Nature Switzerland, Cham (2024)
Morsomme, R., Alferez, S.V.: Content-based course recommender system for liberal arts education. *International educational data mining society* (2019)
Morsy, S., Karypis, G.: Will this course increase or decrease your gpa? towards grade-aware course recommendation. *Journal of Educational Data Mining* 11(2), 20–46 (2019)
Parameswaran, A., Venetis, P., Garcia-Molina, H.: Recommendation systems with complex constraints: A course recommendation perspective. *ACM Transactions on Information Systems* (TOIS) 29(4), 1–33 (2011)
Pardos, Z.A., Fan, Z., Jiang, W.: Connectionist recommendation in the wild: on the utility and scrutability of neural networks for personalized course guidance. *User modeling and user-adapted interaction* 29, 487–525 (2019a)
Pardos, Z.A., Jiang, W.: Combating the filter bubble: Designing for serendipity in a university course recommendation system. *arXiv preprint* arXiv:1907.01591 (2019b)
Polyzou, A., Nikolakopoulos, A.N., Karypis, G.: Scholars walk: A markov chain framework for course recommendation. *International Educational Data Mining Society* (2019)
Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. *OpenAI blog* 1(8), 9 (2019)
Salakhutdinov, R., Mnih, A.: Probabilistic matrix factorization. In: *Proceedings of the 20th International Conference on Neural Information Processing Systems*. pp. 1257–1264 (2007)
Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Item-based collaborative filtering recommendation algorithms. In: *Proceedings of the 10th international conference on World Wide Web*. pp. 285–295 (2001)
Sweeney, M., Rangwala, H., Lester, J., Johri, A.: Next-term student performance prediction: A recommender systems approach. In: *arXiv preprint* arXiv:1604.01840 (2016)
Yang, T., Ren, B., Ma B., He, T., Gu, C., Konomi, S.: Boosting course recommendation explainability: A knowledge entity aware model using deep learning. In: *International Conference on Computers in Education* (2024a)
Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. *arXiv preprint* arXiv:2302.13971 (2023)
Wu, L., Zheng, Z., Qiu, Z., Wang, H., Gu, H., Shen, T., Qin, C., Zhu, C., Zhu, H., Liu, Q., et al.: A survey on large language models for recommendation. *World Wide Web* 27(5), 60 (2024)
Wu, X., Zhou, H., Yao, W., Huang, X., Liu, N.: Towards personalized cold-start recommendation with prompts. *arXiv preprint* arXiv:2306.17256 (2023)
Yang, T., Ren, B., Gu, C., Ma, B., Konomi, S.: Leveraging ChatGPT for automated knowledge concept generation. In: *21st International Conference on Cognition and Exploratory Learning in the Digital Age, CELDA* 2024. pp. 75–82. IADIS Press (2024b)
Yu, J., Wang, Y., Zhong, Q., Luo, G., Mao, Y., Sun, K., Feng, W., Xu, W., Cao, S., Zeng, K., et al.: MOOCCubeX: A large knowledge-centered repository for adaptive learning in MOOCs. In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. pp. 4643–4652 (2021)
Zhang, J., Hao, B., Chen, B., Li, C., Chen, H., Sun, J.: Hierarchical reinforcement learning for course recommendation in moocs. In: *Proc. AAAI conf artificial intelligence*. vol. 33(1), pp. 435–442 (2019)
Zhang, J., Xie, R., Hou, Y., Zhao, X., Lin, L., Wen, J.R.: Recommendation as instruction following: A large language model empowered recommendation approach. *ACM Transactions on Information Systems* (2023)
Zhang, Y., DING, H., Shui, Z., Ma, Y., Zou, J., Deoras, A., Wang, H.: Language models as recommender systems: Evaluations and limitations. In: I (Still) Can't Believe It's Not Better! *NeurIPS* 2021 Workshop (2021)
Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, *36*, 10088-10115. (2023)
Frej, J., Shah, N., Knezevic, M., Nazaretsky, T., & Käser, T. Finding paths for explainable mooc recommendation: A learner perspective. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. pp. 426-437 (2024).
Yang, T., Ren, B., Gu, C., Ma, B., He, T., & Konomi, S. I. Towards better course recommendations: integrating multi-perspective meta-paths and knowledge graphs. In *Proceedings of the 15th international learning analytics and knowledge conference*. pp. 137-147 (2025).

Investigating Algorithmic Bias in Affect Detectors with Constructed Categories of Student Identity

Nidhi NASIAR[a*], Clara BELITZ[b], HaeJin LEE[b], Frank STINAR[b], Ryan S. BAKER[c], Jaclyn OCUMPAUGH[d], Stephen E. FANCSALI[e], Steve RITTER[e], Nigel BOSCH[e]
[a]*University of Pennsylvania, USA*
[b]*University of Illinois Urbana–Champaign, USA*
[c]*Adelaide University, Australia*
[d]*University of Houston, USA*
[e]*Carnegie Learning, Inc., USA*
*nasiar@upenn.edu

Abstract: Algorithmic bias research often evaluates models in terms of traditional demographic categories (e.g., U.S. Census), but these categories may not capture nuanced, context-dependent identities relevant to learning. This study evaluates four affect detectors (boredom, confusion, engaged concentration, and frustration) developed for an adaptive math learning system. Metrics for algorithmic fairness (AUC, weighted F1, MADD) show subgroup differences across several categories that emerged from a free-response social identity survey (Twenty Statements Test; TST), including both those that mirror demographic categories (i.e., race and gender) as well as novel categories (i.e., Learner Identity, Interpersonal Style, and Sense of Competence). For demographic categories, the confusion detector performs better for boys than for girls and underperforms for West African students. Among novel categories, biases are found related to learner identity (boredom, engaged concentration, and confusion) and interpersonal style (confusion), but not for sense of competence. Results highlight the importance of using contextually grounded social identities to evaluate bias.

Keywords: algorithmic bias, affect, fairness, social identity, twenty statements test

1. Introduction

The scaled adoption of adaptive learning technologies, which in many cases have led to improved student outcomes (Rai & Murthy, 2022), has raised concerns about algorithmic fairness (Kizilcec & Lee, 2022). A critical step in addressing these concerns is detecting these biases, which have been investigated in terms of race (Hu & Rangwala, 2020), gender (Christie et al., 2019), neurodivergence (e.g., ADHD; Lee et al., 2025), socioeconomic status (Yu et al., 2021), urbanicity (Ocumpaugh et al., 2015), and regional groups (Svabensky et al., 2024). The range of subgroups potentially affected by these biases accentuates the complexity of this issue. To date, many studies rely solely on institutionally defined demographic labeling systems, such as race or gender, based on legal protections in some countries. These labels often deliberately merge smaller categories in order to facilitate large-scale comparisons, but may fail to capture the nuanced, context-dependent dimensions of student identity in the classroom. In practice, heterogeneous groups (e.g., English language learners) are treated as monoliths, and significant within-group diversity is left unanalyzed or even misrepresented (Wang et al., 2022).

In addition to collapsing heterogeneity, institutionally defined labels may also miss the unique, context-specific dimensions of students' self-perceived identities (e.g., learner identity), which may be especially relevant in classroom settings (Cribbs et al., 2015; Crossley et al., 2018). Research shows that a student's sense of identity—like seeing oneself as "good at math"—can drive learning outcomes and academic success (Cribbs et al., 2015). Likewise,

while there are well-documented links between large-scale demographic categories (e.g., race, gender) and learners' performance, as well as help-seeking and motivational behaviors(e.g., Karumbaiah et al., 2022), incorporating non-traditional, locally meaningful categories can provide a richer, more nuanced picture of students and their learning experiences. These context-specific identities particularly matter in the case of modeling emotions, which are inherently social and situated (Barrett et al., 2019).

To investigate this, we examine whether affect detectors built for MATHia (an adaptive learning system) show biases against middle school students in terms of their self-reported identities, as elicited from a free-response survey. Doing so helps us understand whether any of these less-studied categories of identities are important for determining if automated detectors are less effective for some groups of students.

2. Literature Review

2.1 Algorithmic Bias in Education

Research on algorithmic bias in education has largely focused on broad demographic categories like race, ethnicity, nationality, and gender (Baker & Hawn, 2022). Such studies have revealed racial disparities across models predicting college retention (Kai et al., 2017), high school dropout (Christie et al., 2019), and course failure risk (Hu & Rangwala, 2020), with inconsistent results highlighting the complexity of intra-group heterogeneity. Geographic and linguistic differences have also emerged in automated essay scoring (Bridgeman et al., 2009) and in help-seeking models favoring domestic learners (Ogan et al., 2015). Gender disparities also appear but are often intertwined with other sociocultural factors (Yu et al., 2021).

However, reliance on these traditional categories risks over-simplification, as aggregating heterogeneous subgroups can obscure contextually salient identities (Belitz et al., 2023). Emerging work has explored less-studied categories, such as urbanicity (Ocumpaugh et al., 2014), neurodivergent groups (Lee et al., 2025), and regional groups in the Philippines (Svabensky et al., 2024), revealing biases tied to dimensions of identity that are finer-grained and more contextually situated than what is often studied. Yet these efforts remain fragmented, and few frameworks address how self-reported identities affect algorithmic biases. This gap highlights the need for more nuanced approaches to subgroup definition, ensuring that fairness evaluations reflect the lived experiences of learners rather than overly broad, externally chosen demographic variables.

2.2 Twenty Statements Test (TST)

Determining which factors are relevant potential sources of bias for a given student remains an underexplored area (Belitz et al., 2023). Gender, racial, and economic categories are often analyzed because of longstanding patterns of disparity in the treatment and educational outcomes of these groups (Kao & Thompson, 2003). While these issues remain important, the categories used in these contexts—which are designed to apply to many learners for ease of comparison at scale—are not always the most relevant in a specific situation. For example, populations in South Asia and East Asia have ethnic groups with vast cultural and linguistic diversity, which are often grouped under the same label, even though these groups (and their teachers) likely consider themselves distinct from one another. Efforts to capture identities that could supplement the existing commonly studied categories exist (Crossley et al., 2018), but the ability to do so at scale is quite challenging. One promising approach is the Twenty Statements Test (TST), which elicits a free-response survey to ask students how they would describe themselves (Kuhn, 1954). This technique is designed to surface salient categories that emerge from students' responses, which may not always be captured by traditional approaches.

2.3 Algorithmic Bias in Affect Detection

Existing research on algorithmic bias in education has examined predictive models of dropout (Christie et al., 2019), automated essay scoring (Bridgeman et al., 2009), and help-seeking behaviors (Ogan et al., 2015), but research involving affect detectors remains limited. For observation-based detectors, Ocumpaugh et al. (2014) found detectors failed to work across differences in urbanicity, and Chiu (2020) found gendered affect patterns—female STEM students persisted when less bored and off-task, while male students thrived with higher concentration and lower frustration. More recently, with facial-recognition-based detectors, Ashwin & Biswas (2024) reported higher misclassification rates for students with darker skin tones, underscoring the need for diverse training data. These limited efforts highlight the importance of examining bias in affect detectors across varied subpopulations. Furthermore, since affect is rooted in cultural and linguistic differences (Barrett et al., 2019), it may be important to incorporate variables of this nature in investigating the algorithmic bias that impacts detectors of affect.

3. Methods

3.1 Participants and Study Context

The study was conducted with middle school students across several classes in a small city in the northeastern U.S. with a large immigrant population. These students used Carnegie Learning's MATHia software in 2023-2024 as their regular math instruction. Participation in the study involved two different types of data collection. First, 219 students completed the Twenty Statements Test to determine relevant social categories for the students. Second, field observation of 163 students was conducted to collect training labels for automated detectors. Both parental consent and student assent were acquired.

The labeled affect data from the observers was mapped to action logs in MATHia for each student, as well as the corresponding TST identity labels when present. TST responses from students without affect labels were dropped since the affect labels served as ground truth and were necessary to build the detectors. A final total of 95 students' data were used for analysis. Students who didn't have TST responses for a specific category of interest were noted as NR (not reported) for that category, and evaluated as a separate subgroup for that category (explained in detail below).

3.2 BROMP-based Affect Detectors

3.2.1 BROMP Field Observations

Affect training labels were obtained using the Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP; Ocumpaugh et al., 2015a), a classroom observation method that has been used to produce detectors for more than two dozen online learning systems (Baker et al., 2020). BROMP observers use a momentary time sampling method, implemented by an open-source Android app (Ocumpaugh et al., 2015b), to label students' affective engagement. As part of the BROMP certification process, the researchers achieved a kappa of greater than 0.6 on all affective states for middle school students' observations, indicating good agreement among the researchers. Following this, coders independently coded over five class periods, observing for boredom, confusion, delight, engaged concentration, and frustration. As advised in the BROMP manual, these coders did not observe the same student simultaneously, maximizing data collection (Ocumpaugh et al., 2015a). In total, 591 observations of affect were obtained from 163 students. Engaged concentration was the most prevalent affective state (N=591; 71.21%), followed by confusion (N=109; 13.13%), frustration (N=42; 5.06%), boredom (N=86; 10.36%), and delight (N=2; 0.24%). This distribution is typical for BROMP research (Karumbaiah et al., 2021). Due to low N, a delight detector was not built.

3.2.2 Feature Engineering

To construct affect detectors, we extracted 157 features from the action log data of students. These included features like error counts, number of attempts, and changes in hint levels., based on actions of students. As typical in BROMP-based detectors, features within the 20 seconds (defined as a "clip") before the BROMP label's timestamp were prioritized, but earlier features were used to add contextual information (Baker et al., 2012; Zambrano et al., 2024).

3.2.3 Detector Construction and Validation

We performed stratified 4-fold student-level cross-validation. This ensured that affect labels were balanced across each fold while making sure that each student's labels appeared only in either the training or test set. To address class imbalances, we used a synthetic oversampling technique (SMOTE; Chawla et al., 2002) on training data. Features were selected through a forward feature selection process, evaluating performance using the Area Under the Receiver Operating Characteristic Curve (AUC ROC; AUC for short). Following common practice, each detector was trained as a one vs. rest classification (Baker et al., 2012; Zambrano et al., 2024). We tested 5 machine-learning algorithms from the Scikit-learn library for Python: Logistic Regression (LR), Random Forests (RF), Extreme Gradient Boosting (XGB), Support Vector Machines (SVM), and Decision Trees. We used the default hyperparameters for all models, as the study's goal was not to optimize overall model performance but to assess bias for the novel TST categories.

3.3 Twenty Statements Test (TST)

3.3.1 TST Implementation

This study measures student identity using the Twenty Statements Test (Kuhn, 1954), which asks students to fill in the blank—twenty times—for the question "Who am I?". This free-form survey has been used to produce nuanced identity classifications across urbanity samples (Somech, 2000), age groups (McRae & Costa, 1988), and nationalities (Santamaria et al., 2010). Notably, the TST facilitates identifying locally relevant categories that more fully reflect the identities of those being sampled.

3.3.2 TST Data Categorization

Qualitative coding of TST data was adapted from Gordon and Gergen (1968); these codes were then combined into umbrella categories specific to our sample. Table 1 shows the N for each subgroup across five umbrella categories, with examples given for those that do not align with typical census-style data. Each umbrella category also includes a Not Reported (NR) subgroup, as each umbrella category had students whose answers did not reflect that category. This large number of students being identified as NR is expected for open-ended methods and was included because not reporting a category serves as an indicator of lack of salience of that category in this context.

Table 1. *TST Categories. N≥10 are shown in grayscale and included in the fairness analyses*

Categories	**Subgroups**	**Description/example**	**N**
Gender	B	Boy	13
	G	Girl	21
	NR	not reported	149
Race / Ethnicity / Nationality	Black	-	13
	West African	Deidentified West African Nationality	18
	Caribbean	Deidentified Caribbean Nationality	9
	White	-	4

Categories	Subgroups	Description/example	N
	South Asian	Deidentified South Asian Nationality	2
	East Asian	Deidentified East Asian Nationality	3
	South American	Deidentified South American Nationality	2
	Mixed	-	3
	NR	not reported	142
Learner Identity (LI)	Pos	e.g., “good student” or “good at math”	36
	Neg	e.g. “not great at math” “don’t like school”	10
	Neut	e.g., “student”, or “high schooler”	10
	NR	not reported	127
Interpersonal Style (IPS)	Pos	e.g., “ nice” or “kind” or ”respectful”	33
	Neg	e.g., “rude”	3
	Neut	e.g., “shy”	2
	NR	not reported	145
Sense of Competence (SoC)	Pos	e.g., “smart” “leader”, “talented”	37
	Neg	e.g., “idiot”	1
	NR	not reported	145

Traditional Demographic Categories. Two categories aligned with traditional census measures emerged in this data: race and gender. Gender produced two named subgroups from the data: boy and girl, with NR included as a third subgroup. Race (a merger of race, ethnicity, and nationality) has many named subgroups but only two large enough (N≥10) for analysis: Black and West African (a pseudonym for a specific nationality that students provided, which we use to prevent re-identification of this school district). Because many students identified as both groups, we focus only on the West African subgroup. This decision reflects the higher specificity of the national label, and the concern that the Black subgroup would be more heterogenous. Students who reported being Black but were not in the West African subgroup had N<10 and were thus not analyzed. For analyses, we compare to “Other” labels and “not reported” (NR).

Novel Social and Contextual Categories. Three named categories emerged that did not reflect typical census data. Learner identity (LI) emerged with students reporting to be “a good student” (positive LI), “not good at math” (negative LI), or simply “a student” (neutral LI). Likewise, Interpersonal Style (IPS) also emerged with positive, negative and neutral subgroups, while Sense of Competence (SoC) had positive and negative but no neutral subgroups. An NR subgroup is included for all categories.

To facilitate fairness analysis, final subgroups within categories are mutually exclusive. In cases where TST identity categories are not mutually exclusive (e.g., valence subgroups, where students sometimes provided multiple, contradictory responses), a resolution process totaled the number of positive and negative responses and applied the majority label (e.g., if a student provided four neutral responses and one positive, the final code was neutral). Since no student had equal numbers of contradictory codes, no further resolutions were necessary.

3.4 Testing for Algorithmic Bias

We used three complementary fairness metrics—difference in AUC and weighted F1 (Wtd F1), and Mean Absolute Distribution Difference (MADD)—to test bias against TST subgroups with N≥10. AUC and Wtd F1 are standard predictive-performance metrics, while MADD is a novel fairness metric (Verger et al., 2023). AUC and Wtd F1 were calculated separately for each subgroup. MADD was calculated by comparing one subgroup to all others.

For AUC (range: 0-1; chance = 0.5), a threshold of ΔAUC≥0.1 was used to label bias. AUC reflects the trade-off between true positive and false positive rates, and it can be interpreted as a measure of how well an algorithm ranks individual cases. For the weighted F1 score (range: 0-1), the same threshold was chosen to label bias (ΔwtdF1≥0.1). This score combines precision and recall, while also taking class imbalance into account. Finally, MADD measures the divergence between the predicted probability distributions of a model across

subgroups, with scores ranging from 0 (most fair/identical distributions) to 2 (most biased/completely divergent distributions). For this study, we apply a conservative threshold of MADD>1 for labeling bias (Verger et al., 2023).

As small differences could arise stochastically (Kohavi et al., 2022), we chose conservative thresholds for our metrics. Our decision aligns with calls for precautionary fairness checks that are needed in all contexts (Passi & Barocas, 2019) but perhaps especially in educational AI to proactively mitigate harm (Holstein & Doroudi, 2021). These differences should not be treated as definitive evidence of bias but as signals to prioritize subgroups for validation in larger, representative samples (Mitchell et al., 2021).

4. Results

4.1 Affect Detector Models

The performance of different algorithms was compared for each binary affect classifier, and we selected the model for investigation of algorithmic bias based on the highest AUC. Table 2 shows performances for the final detectors that are comparable to previously published interaction-based affect detectors (Baker et al., 2014; Tabanao & Rodrigo, 2018): Random Forest for boredom (AUC=0.68), Logistic Regression for confusion and engaged concentration (AUC=0.65 and 0.66, respectively), and XGBoost for frustration (AUC=0.74).

Table 2. *Performance of detectors across algorithms with AUC values*

Algo.	**BOR AUC**	**CONF AUC**	**Eng CONC AUC**	**FRU AUC**
Decision Tree	0.62	0.56	0.60	0.63
SVM	0.62	0.61	0.59	0.65
LR	0.67	**0.65**	**0.66**	0.71
RF	**0.68**	0.64	0.64	0.69
XGBoost	0.67	0.64	0.64	**0.74**

4.2 Algorithmic Bias

We next evaluate bias in these models across the four affect detectors. Although we report bias evaluation both for categories similar to traditional, census-style demographics (i.e., gender and race) and for novel categories (i.e., Learner Identity, Interpersonal Style, and Sense of Competence), all data have emerged from students' free form responses in the TST.

4.2.1 Traditional Demographic Categories

Table 3 reports detector performance across TST subgroups and Table 4 shows AUC and F1 differences between subgroups for traditional demographic categories (gender and race).

Table 3. *Model evaluation results for all categories.*

		Gender			Race			Learner Identity				IPS		SoC	
Affect	Metrics	boy	girl	NR	W. Afr.	Other	NR	NR	pos	neg	neut	NR	pos	NR	pos
BOR (RF)	AUC	0.58	0.64	0.70	0.63	0.61	0.69	0.72	0.56	0.67	0.74	0.70	0.70	0.69	0.65
	Wtd F1	0.73	0.73	0.75	0.76	0.69	0.74	0.75	0.73	0.66	0.82	0.76	0.74	0.76	0.71
	MADD	0.45	0.49	0.40	0.49	0.40	0.49	0.31	0.36	0.64	0.66	0.34	0.37	0.35	0.35
CONF (LR)	AUC	0.72	0.62	0.62	0.55	0.66	0.65	0.64	0.64	0.70	0.70	0.58	0.78	0.62	0.67
	Wtd F1	0.73	0.68	0.65	0.64	0.66	0.67	0.67	0.64	0.64	0.79	0.64	0.77	0.64	0.72
	MADD	0.53	0.40	0.41	0.50	0.39	0.50	0.29	0.29	0.66	0.65	0.36	0.41	0.42	0.43
ENG (LR)	AUC	0.63	0.60	0.67	0.65	0.60	0.65	0.66	0.62	0.72	0.60	0.67	0.63	0.68	0.60
	Wtd F1	0.64	0.56	0.61	0.65	0.54	0.60	0.61	0.59	0.63	0.68	0.61	0.63	0.62	0.60

		Gender			Race			Learner Identity				IPS		SoC	
Affect	Metrics	boy	girl	NR	W. Afr.	Other	NR	NR	pos	neg	neut	NR	pos	NR	pos
	MADD	0.44	0.44	0.30	0.38	0.36	0.38	0.28	0.30	0.61	0.73	0.28	0.32	0.29	0.29
FRU (XGB)	AUC	0.71	0.64	0.73	0.79	0.76	0.71	0.70	0.74	0.94	0.80	0.75	0.56	0.72	0.75
	Wtd F1	0.82	0.93	0.90	0.87	0.85	0.89	0.90	0.87	0.92	0.83	0.88	0.92	0.89	0.88
	MADD	0.41	0.27	0.24	0.35	0.26	0.35	0.23	0.28	0.39	0.63	0.17	0.22	0.20	0.20

Table 4. *Differences in fairness metrics for all categories. Diff ≥ |0.1| are highlighted in gray*.

		Gender			Race			Learner Identity						IPS	SoC
Affect	Metrics	B-G	B-NR	G-NR	WA-Ot	WA-NR	Ot-NR	NR-Pos	NR-Neg	NR-Neut	Pos-Neg	Pos-Neut	Neg-Neut	NR-pos	NR-pos
BOR	AUC	-0.06	-0.12	-0.06	0.02	-0.06	-0.08	0.16	0.05	-0.02	-0.11	-0.18	-0.07	0.00	0.04
(RF)	WtdF1	0.00	-0.02	-0.02	0.07	0.02	-0.05	0.02	0.09	-0.07	0.07	-0.09	-0.16	0.02	0.05
CONF	AUC	0.10	0.10	0.00	-0.11	-0.10	0.01	0.00	-0.06	-0.06	-0.06	-0.06	0.00	-0.20	-0.05
(LR)	WtdF1	0.05	0.08	0.03	-0.02	-0.03	-0.01	0.03	0.03	-0.12	0.00	-0.15	-0.15	-0.13	-0.08
ENG	AUC	0.03	-0.04	-0.07	0.05	0.00	-0.05	0.04	-0.06	0.06	-0.10	0.02	0.12	0.04	0.08
(LR)	WtdF1	0.08	0.03	-0.05	0.11	0.05	-0.06	0.02	-0.02	-0.07	-0.04	-0.09	-0.05	-0.02	0.02
FRU	AUC	0.07	-0.02	-0.09	0.03	0.08	0.05	-0.04	-0.24	-0.10	-0.20	-0.06	0.14	0.19	-0.03
(XGB)	WtdF1	-0.11	-0.08	0.03	0.02	-0.02	-0.04	0.03	-0.02	0.07	-0.05	0.04	0.09	-0.04	0.01

Gender. The performance of each affect detector was tested across three gender subgroups: boy, girl, and not reported (NR). Boredom detection produced ΔAUC>0.1 when comparing the NR group (0.70) to boys (0.58). In contrast, confusion's AUCs are higher for boys (0.72) than for both girls and NR (each 0.62). For engaged concentration and frustration, no major disparities are found. Wtd F1 scores are largely comparable across subgroups—with the only notable exception for frustration (better for girls than for boys). MADD values ranged from 0.40 to 0.53, suggesting fair affective models.

Race. Next, we examine the performance of these detectors across three racial subgroups: West African, Others, and NR. For boredom, AUC values are similar across subgroups, but confusion detection does not perform as well for West African students (0.55) as for Others (0.66) and NR (0.65). For engaged concentration and frustration, AUC values showed minimal differences across subgroups. Wtd F1 scores also show few differences, though engaged concentration detection performs better for West African students than for Others (ΔWtd F1=0.11, see Table 3). All MADD values are relatively small (≤0.5).

4.2.2 Novel Identity Categories

Detector performance was tested across three novel TST categories: Learner Identity (LI), Interpersonal Style (IPS), and Sense of Competence (SoC). Among the LI codes, four subgroups are compared (positive, negative, neutral, and NR), and biases were evenly distributed across the affect detectors. Among the IPS and SoC codes, only two subgroups are compared (NR, positive), as only these subgroups had N>10 (Table 1). More differences in performance between the subgroups are seen in these novel categories than for the traditional ones. The specific results are discussed below.

Learner Identity (LI). Testing across the four LI subgroups shows that the boredom detector underperforms for students with a positive learner identity (AUC=0.56) compared to other students (NR: 0.72, negative: 0.67, neutral: 0.74), but for Wtd F1 scores, the detectors performed worse for students with negative LI than students with neutral LI (0.66 vs 0.82). For confusion, AUC values are comparable across subgroups, but Wtd F1 scores show better performance for the neutral subgroup (0.79) compared to the other three subgroups (ΔAUC≥0.12). For concentration, performance is better for the negative subgroup (0.72) than the positive (0.62) and neutral (0.60) subgroups (ΔAUC≥0.10). For frustration, the negative subgroup again performed best on AUC (negative: 0.94 vs. NR: 0.70, positive: 0.74, neutral:

0.80; ΔAUC=0.14 to 0.24). MADD values are highest when comparing the neutral subgroup across affective states (avg=0.67) but remain below 1 in all cases.

Interpersonal Style (IPS). For boredom, AUC values of both subgroups were equivalent (0.70), but for confusion, performance was better for the positive subgroup than the NR subgroup (0.78 vs 0.58; ΔAUC≥0.20). Similar differences were observed in Wtd F1, with the positive subgroup performing better than NR (0.77 vs 0.64; ΔwtdF1=0.13). For engaged concentration, there were no differences (.63 vs 0.67), but frustration detection performed better for the NR subgroup than the positive subgroup (0.75 vs 0.56; ΔAUC≥0.19). Low MADD values (<0.45) were seen across all subgroup comparisons.

Sense of Competence (SoC). Across all affect**,** AUC and Wtd F1 differences were minimal (ΔAUC < 0.1), indicating no significant disparities in model performance, while MADD values were also low, showing no substantial differences in probability distribution across subgroups.

5. Discussion and Conclusion

5.1 Summary of Results

Overall, results show performance differences across groups, emphasizing the importance of novel categories of Learner Identity and Interpersonal Style, in addition to demographic categories of gender and race for analyzing bias in affect detectors.

5.1.1 Performance Across Subgroups

Within gender, four group comparisons showed performance differences. Boys' boredom was not as accurately modeled (vs the NR group), nor was their frustration (vs the girls). However, the confusion detectors outperformed for boys compared to the other two groups. Within race categories, only two groups and two affective states experienced bias. The confusion detectors underperformed for the West African subgroup (compared to either subgroup), while concentration detection underperformed for the Others (vs. the West African subgroup) based on weighted F1.

More differences emerge within the novel categories. As Table 4 shows, no strong trends appear in terms of which detector is underperforming on which subgroup. Detectors underperformed for the positive LI subgroup in six comparisons to other groups, three of which involve boredom. Both the negative LI subgroup and neutral LI are underpredicted for two affective states (negative: boredom and confusion; neutral: frustration and concentration). Finally, the NR group is underpredicted for confusion (one comparison) and frustration (two comparisons). For the Interpersonal Style and Sense of Competence categories, differences only emerge for the positive and NR Interpersonal subgroups (in frustration and confusion).

Overall, the demographics typically studied in algorithmic bias research (gender, race/ethnicity) do not seem to be the categories showing the most bias in affect detectors. Instead, learner identities show the most variability in model performance, suggesting that non-traditional self-categorizations might help us to better understand algorithmic bias and the development of context-aware detectors more generally.

One limitation of the current method is the possibility of self-presentation effects. Combining the TST with validated surveys (e.g., self-efficacy) could disentangle some self-presentation effects from genuine identity differences.

5.1.2 Interpreting Not Reported (NR) Groups

The most common subgroup across all categories was NR, the group that had not provided a TST response relevant to that category. The NR groups raise important questions about our data, since identity salience—shaped by context, self-presentation effects, and societal norms—may influence reporting patterns. For instance, students in sensitive contexts (e.g., immigration status concerns) might underreport that identity, and other marginalized students

may avoid labels to avoid stigmatization. Likewise, cultural norms could make some groups less comfortable reporting accomplishments, which could influence categories like learner identity (e.g., “I’m a good student!”).

Although these issues complicate the interpretation of NR subgroups, additional surveys and efforts to cross-reference this label with the records maintained by a school could help us better understand these patterns. These efforts could be useful if we want to understand why, for example, a frustration detector is underperforming for NR group.

5.2 Conclusion

This study evaluates algorithmic bias within affect detectors that were developed for the MATHia online learning system. Unlike many previous studies of algorithmic bias, which have typically relied on census-style demographic data, we check model performance among social identity categories extracted from the Twenty Statements Test (TST). Evidence of algorithmic bias demonstrate that the bias may occur for characteristics beyond traditional demographic categories (cf. Baker & Hawn, 2024). The inclusion of TST categories enabled us to detect differences that might have been missed using more traditional categories. In particular, learner identity showed multiple differences across affect detectors and subgroups.

These findings offer evidence for the need to evaluate bias across contextually relevant categories (e.g., Learner Identity and Interpersonal Style), but further research is necessary to generalize these results to a broader population. In addition, future work should investigate how self-reported identity categories interact with presentation effects and other dynamics related to students’ identity reporting behaviors. More research on under-studied dimensions of identity, including those shaped by specific regional and ethnic norms, could allow us to better mitigate potential biases against small, contextually situated subgroups. Such efforts could enhance the fairness of models across a broader range of socially and contextually grounded student identities.

References

Ashwin, T. S., & Biswas, G. (2024). Identifying and mitigating algorithmic bias in student emotional analysis. *Int. Conf. Art. Intelligence in Ed., 89-103.* Cham: Springer Nature Switzerland.

Baker, R.S., Gowda, S., Wixon, M., Kalka, J., Wagner, A., Salvi, A., Aleven, V., Kusbit, G., Ocumpaugh, J. Rossi, L. (2012). Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra. *International Conference on Educational Data Mining (2012), 126-133.*

Baker, R. S., & Hawn, A. (2022). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 1-41*.

Baker, R. S., Ocumpaugh, J. L., & Andres, J. M. A. L. (2020). BROMP quantitative field observations: A review. Learning science: Theory, research, and practice, 127-156.

Baker, R.S., Ocumpaugh, J., Gowda, S.M., Kamarainen, A.M. and Metcalf, S.J. (2014). Extending log-based affect detection to a multi-user virtual environment for science.

Barrett, L. F. (2019). How emotions are made. Providence Book Festival, May, 25, 2019.

Belitz, C., Ocumpaugh, J., Ritter, S., Baker, R. S., Fancsali, S. E., & Bosch, N. (2023). Constructing categories: Moving beyond protected classes in algorithmic fairness. *Journal of the Association for Information Science and Technology, 74(6), 663-668.*

Bridgeman, B., Trapani, C., & Attali, Y. (2009). Considering fairness and validity in evaluating automated scoring*. In annual meeting of the National Council on Measurement in Education, San Diego, CA.*

Chawla et al. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16, 321-357.*

Chiu, M. S. (2020). Gender Differences in Predicting STEM Choice by Affective States and Behaviors in Online Mathematical Problem Solving: Positive-Affect-to-Success Hypothesis. *Journal of Educational Data Mining, 12(2), 48-77.*

Christie, S. T., Jarratt, D. C., Olson, L. A., & Taijala, T. T. (2019). Machine-Learned School Dropout Early Warning at Scale. *International Educational Data Mining Society.*

Cribbs, J.D., Hazari, Z., Sonnert, G., Sadler, P.M. (2015). Establishing an explanatory model for mathematics identity. *Child Development 86(4), 1048-1062.*

Crossley, S., Ocumpaugh, J., Labrum, M., Bradfield, F., Dascalu, M., Baker, R.S. (2018). Modeling Math Identity and Math Success through Sentiment Analysis and Linguistic Features. *International Educational Data Mining Society.*

Gordon, C., and Gergen, K.J. (1968). Self-conceptions: Configurations of content.

Holstein, K., & Doroudi, S. (2021). Equity and artificial intelligence in education: Will aied amplify or alleviate inequities in education?. arXiv preprint arXiv:2104.12920.

Hu, Q., & Rangwala, H. (2020). Towards Fair Educational Data Mining: A Case Study on Detecting At-Risk Students. International Educational Data Mining Society.

Kao, G., & Thompson, J. S. (2003). Racial and ethnic stratification in educational achievement and attainment. Annual Review of Sociology, 29(1), 417-442.

Karumbaiah, S., Baker, R. S., Ocumpaugh, J., & Andres, J. M. A. L. (2021). A re-analysis and synthesis of data on affect dynamics in learning. *IEEE Transactions on Affective Computing*, *14*(2), 1696-1710.

Karumbaiah, S., Ocumpaugh, J., & Baker, R. S. (2022). Context matters: Differing implications of motivation and help-seeking in educational technology. Int. J. Artificial Intel. in Ed., 32(3), 685-724.

Kizilcec, R. F., & Lee, H. (2022). Algorithmic fairness in education. The ethics of artificial intelligence in education, 174-202. Routledge.

Kohavi, R., Deng, A., & Vermeer, L. (2022). A/B testing intuition busters: Common misunderstandings in online controlled experiments. Proc. 28th ACM SIGKDD Conf. on Knowledge Discovery & Data Mining, 3168-3177.

Kuhn, M. & McPartland, T. (1954). An empirical investigation of self-attitudes. American Sociological Review 19, 68-76.

Lee, H., Belitz, C., Nasiar, N., & Bosch, N. (2025). XAI Reveals the Causes of Attention Deficit Hyperactivity Disorder (ADHD) Bias in Student Performance Prediction.

McCrae, R. R., & Costa Jr, P. T. (1988). Age, personality, and the spontaneous self-concept. *Journal of Gerontology, 43(6), S177-S185.*

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application, 8.*

Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C. (2014). Population validity for educational data mining models: A case study in affect detection. British Journal of Educational Technology, 45(3), 487-501.

Ocumpaugh, J., Baker, R.S. and Rodrigo, M.M.T. (2015). Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual.

Ogan, A., Walker, E., Baker, R., Rodrigo, M. M. T., Soriano, J. C., & Castro, M. J. (2015). Towards understanding how to assess help-seeking behavior across cultures. *International Journal of Artificial Intelligence in Education, 25, 229-248.*

Passi, S., & Barocas, S. (2019). Problem formulation and fairness. *In Proceedings of the conference on fairness, accountability, and transparency, 39-48.*

Rai, R. D., & Murthy, S. (2022). Evidence Over Intuition: Identifying Factors That Influence the Effectiveness of Large Scale Edtech Initiatives. In *International Conference on Computers in Education, 261-267*.

Santamaría, A., de la Mata, M. L., Hansen, T. G., & Ruiz, L. (2010). Cultural self-construals of Mexican, Spanish, and Danish college students: Beyond independent and interdependent self. Journal of Cross-Cultural Psychology, 41(3), 471-477.

Somech, A. (2000). The independent and the interdependent selves: Different meanings in different cultures. *International Journal of Intercultural Relations, 24(2), 161-172.*

Švábenský, V., Verger, M., Rodrigo, M. M. T., Monterozo, C. J. G., Baker, R. S., Saavedra, M. Z. N. L., & Shimada, A. (2024). Evaluating algorithmic bias in models for predicting academic performance of filipino students. EDM, 2024.

Tabanao, E., & Rodrigo, M. M. (2018). Investigating the Generalizability of Affect Detectors from Facial Expressions. In *International Conference on Computers in Education*.

Verger, M., Lall, S., Bouchet, F., & Luengo, V. (2023). Is Your Model "MADD"? A Novel Metric to Evaluate Algorithmic Fairness for Predictive Student Models. arXiv preprint arXiv:2305.15342.

Wang, A., Ramaswamy, V. V., & Russakovsky, O. (2022). Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. *In Proc. of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 336-349.*

Yu, R., Lee, H., & Kizilcec, R. F. (2021). Should college dropout prediction models include protected attributes?. In Proc. of the 8th ACM conference on learning@ scale (pp. 91-100).

Zambrano, A. F., Nasiar, N., Ocumpaugh, J., Goslen, A., Zhang, J., Rowe, J., ... & Hutt, S. (2024). Says Who? How different ground truth measures of emotion impact student affective modeling. *In Proceedings of the 17th Int. Conference on Educational Data Mining, 211-223.*

Coached by My Future Self: A Syllabus-Driven AI Mentor to Build Grit, Agency, and Learning Performance through a Mantle-of-the-Expert Journey

Yao-En CHEN[a], Vando Gusti AL HAKIM[b*], Su-Hang YANG[c], Jen-Hang WANG[d], Rayvin LAI[a] & Gwo-Dong CHEN[a]
[a]*Department of Computer Science and Information Engineering, National Central University*
[b]*Department of Electrical Engineering Education, Universitas Negeri Yogyakarta*
[c]*Department of Hospitality Management, Chien Hsin University of Science and Technology*
[d]*Research Center for Science and Technology for Learning, National Central University*
[acd]*Taiwan*
[b]*Indonesia*
*vandogustialhakim@uny.ac.id

Abstract: Learning is a transformative journey from not knowing to mastery, guided by syllabi that outline weekly topics, learning objectives, tasks, and assessment rubrics. While these documents scaffold academic content, they rarely engage students as developing professionals or motivate them through long-term identity growth. Many learners struggle to stay on track—grappling with procrastination, low persistence, and a diminished sense of ownership over one's learning—especially when weekly tasks feel disconnected from their future goals. This paper introduces a new approach to AI-enhanced learning: a "Future Self Coach" that transforms a standard course syllabus into a personalized mentor. Rather than functioning as a virtual assistant, this large language model (LLM)-powered agent embodies the student's envisioned expert self at the end of the course. The system leverages existing course syllabi—specifically their objectives and rubrics, framed as identity-progress dashboards—as inputs to generate personalized guidance aligned with weekly expectations, without requiring teachers to author complex content. Through this interaction, students are not only supported in completing tasks, but also gradually adopt the mindset promoted by the Mantle-of-the-Expert approach, where learning is framed as stepping into expert roles. A quasi-experimental study with 97 vocational students in hospitality training compared three conditions: (1) a learning management system (LMS) with the Future Self Coach, (2) an LMS with a general-purpose AI assistant, and (3) a standard LMS. Results showed that students supported by the Future Self Coach significantly outperformed the other groups across all measures—demonstrating higher learning performance, greater grit, stronger sense of agency, and engaged more frequently. This study presents a scalable, psychologically grounded framework for integrating AI into existing courses without burdening instructors. By leveraging the syllabus as both a roadmap and motivational scaffold, the system bridges the gap between short-term tasks and long-term identity, helping learners stay focused, resilient, and committed to their professional growth.

Keywords: *Future Self, LLM, Grit, Sense of Agency, Personalized Learning*

1. Introduction

Learning is a cumulative journey toward expertise, structured not only by the content students consume but also by how they engage with it over time. Many students know *what* to do in a course, but struggle to stay motivated about *why* it matters (Reeve, 2024). Weekly tasks often feel like isolated checkboxes, disconnected from personal growth or professional identity. This

disconnect leads to familiar struggles: procrastination, declining persistence, and a diminished sense of ownership over one's learning (Kirschner & Hendrick, 2024). These internal struggles—often invisible to instructors—can critically affect learning outcomes. While syllabi offer week-by-week structure—laying out topics, tasks, and assessments—they rarely help students see themselves as future professionals in the making (Harrington & Thomas, 2023). The gap between present effort and future identity is rarely made visible or actionable within the learning process.

This paper reimagines the syllabus not just as a roadmap for content, but as a bridge to the student's future self—the envisioned expert identity they hope to achieve (Hershfield, 2011; Pataranutaporn et al., 2024). We introduce the *Future Self Coach*, an AI-powered mentor built on large language models (LLMs) that uses course syllabi and rubrics to guide students on their journey toward becoming competent or "expert" by the end of a course. Unlike conventional virtual assistants or helpdesk-style chatbots, the pop-up coach speaks as the learner's future professional self—helping them stay focused, track progress, and receive motivational feedback. Our approach is grounded in the *Mantle-of-the-Expert* (*MoE*) pedagogy, which positions learners as professionals-in-training by having them adopt the roles and responsibilities of real-world experts (Heathcote & Bolton, 1994; Taylor, 2025). Rather than inventing new curricula, we turn existing course artifacts—especially the syllabus and its rubrics—into what we call an *identity-progress dashboard* (Novice → Apprentice → Expert) to make identity progress visible. This enables students to track their growth not only through grades, but through gradual mastery of expert habits and mindsets.

Our design rationale activates two essential psychological drivers: **grit**—the ability to sustain effort toward long-term goals (Derakhshan et al., 2025; Duckworth et al., 2007)—and **sense of agency**, or the belief that one's actions can shape meaningful outcomes (Haggard, 2017; McGivney, 2025). Both are essential to learning, but often neglected in digital learning systems, which mostly focus on content delivery and reactive help (Labadze et al., 2023). Such systems lack mechanisms to support long-term motivation or translate course milestones into emotionally resonant, identity-centered guidance. To address this gap, we introduce the Future Self Coach and investigate its effects on student learning performance. A quasi-experimental study was conducted in a university-level foreign language hospitality course, comparing three learning conditions: (1) a standard LMS, (2) an LMS with a general-purpose AI assistant, and (3) an LMS enhanced with the Future Self Coach. The study is guided by the following research questions:

1. **RQ1**: How does the Future Self Coach affect students' learning achievement?
2. **RQ2**: How does it influence students' grit?
3. **RQ3**: How does it impact students' sense of agency?
4. **RQ4**: How does it affect students' engagement frequency over time?

Together, our contribution lies in presenting a zero-authoring, syllabus-driven AI coach that transforms standard course artifacts into a personalized mentoring system—without adding any workload for teachers. The remainder of this paper outlines the theoretical framework, system design and implementation, research methods, results, and implications.

2. Theoretical Framework

2.1 Future Self

The concept of the future self refers to an individual's mental image of who they aspire to become (Hershfield, 2011). This psychological construct has been widely studied in fields such as human-computer interaction (Pataranutaporn et al., 2024), health behavior (Zhao et al., 2022), and behavioral economics (Taylor & Carlson, 2025). A consistent finding across these domains is that when individuals feel a stronger connection to their future selves—a state known as future self-continuity—they are more likely to delay gratification, make responsible decisions, and persist through short-term discomfort in pursuit of long-term goals.

Despite its motivational promise, the future self remains rarely applied in educational settings, and even less so as an embedded element of daily coursework (Feng & Papi, 2020). Most implementations take the form of one-off journaling or career reflections (Beauvais, 2025), disconnected from the flow of instructional content. Our approach fills this gap by treating the future self not as a passive vision, but as an active mentor. The Future Self Coach simulates a student's professional identity at the course's end, guiding their weekly learning while reinforcing alignment with long-term goals. This identity framing lays the foundation for our integration of role-based pedagogies such as Mantle-of-the-Expert.

2.2 Mantle-of-the-Expert (MoE) Pedagogy

Mantle-of-the-Expert (MoE) is a role-based pedagogy that invites students to learn by stepping into professional identities within realistic scenarios (Heathcote & Bolton, 1994). Rather than passively absorbing content, they take on expert roles—such as managers or service staff—to solve problems and build both skills and identity through structured narratives. MoE fosters agency and responsibility by framing learners as competent actors from the outset.

Prior research has shown MoE's value in project-based and interdisciplinary learning, especially in drama education and humanities (Taylor, 2025; Kosma, 2021). However, it remains largely absent from digital or AI-enhanced platforms, which often focus on content rather than professional identity. Our system adapts MoE for the digital space by mapping syllabus goals to role-based progression. Weekly tasks are framed as real-world challenges, supported by an AI coach that reinforces students' expert identity throughout the course.

2.3 LMS-Integrated AI Mentorship

AI has advanced rapidly in education, particularly in adaptive feedback, automated grading, and tutoring systems (Mittal et al., 2024). Recent developments in large language models (LLMs) have enabled richer dialogue-based interactions, with growing interest in conversational agents for learning (McTear, 2022). However, most implementations remain task-centric rather than long-term motivation or identity development.

Several studies have explored AI assistants or virtual agents in LMS environments (Gubareva & Lopes, 2020; Sougleridi et al., 2023), but these typically focus on reactive help—not on representing who the learner could become. To our knowledge, no existing system meaningfully integrates teacher-authored documents—such as syllabi or rubrics—into identity-aware guidance. Our system is best described as an LLM-powered conversational agent, accessible through a chat-like interface embedded in the LMS. It delivers weekly check-ins through messages that link course tasks to future professional roles, encourage self-reflection, and offer motivational support when students encounter challenges. This design shifts the role of the AI assistant from a content responder to an identity-centered mentor.

2.4 Grit and Sense of Agency

Academic success requires more than subject knowledge—it hinges on a combination of non-cognitive skills that enable students to stay motivated, manage time, and cope with frustration. Among these, grit—the capacity to sustain passion and effort over long periods—is a strong predictor of persistence and goal completion (Derakhshan et al., 2025; Duckworth et al., 2007). Equally important is a sense of agency, or the belief that one's actions meaningfully influence outcomes (Haggard, 2017; McGivney, 2025). Students who perceive high agency are more likely to take initiative, persevere, and remain engaged over time.

Some studies have explored grit interventions (e.g., goal setting, mindset prompts), but few tie them to curriculum structure or identity growth (Chen et al., 2024; Ghafouri, 2024). Our system operationalizes these traits by embedding them into the weekly learning process. Students interact with a future-self coach that affirms long-term goals, while the syllabus-aligned rubric is visualized as an identity-progress dashboard. Badge levels (Novice →

Apprentice → Expert) help learners track not just task completion, but the development of mindset, responsibility, and professional habits over time.

3. System Design and Implementation

The Future Self Coach is designed to deliver scalable, identity-based mentorship aligned with the Mantle-of-the-Expert (MoE) principles. In MoE, students are positioned as professionals-in-training from the outset—learning not as passive recipients, but as emerging experts responsible for real-world tasks. The system enables students to interact with an AI agent that guides them through weekly challenges, framed in the voice of their envisioned future self. Rather than viewing assignments as isolated tasks, learners are coached to see each activity as a meaningful step in their journey toward expertise.

Technically, the system operates as a conversational agent powered by a large language model (LLM), embedded into the Learning Management System (LMS) via a pop-up chat interface. OpenAI's GPT-3.5-turbo model was used via API integration. The Future Self Coach uses a role-conditioning technique based on prompt initialization (Al Hakim et al., 2024), which primes the LLM to consistently respond as the student's future professional self. Students engage with the coach weekly to receive personalized feedback, encouragement, and guidance—generated by comparing two structured data streams: (1) the syllabus and its rubric, which serve as a static map of expectations, and (2) real-time student activity data, including task completion and rubric-aligned scores. The first defines *where* students are expected to go; the second reveals *where* they currently are. The system architecture—including all core components and teacher-student roles—is shown in Figure 1.

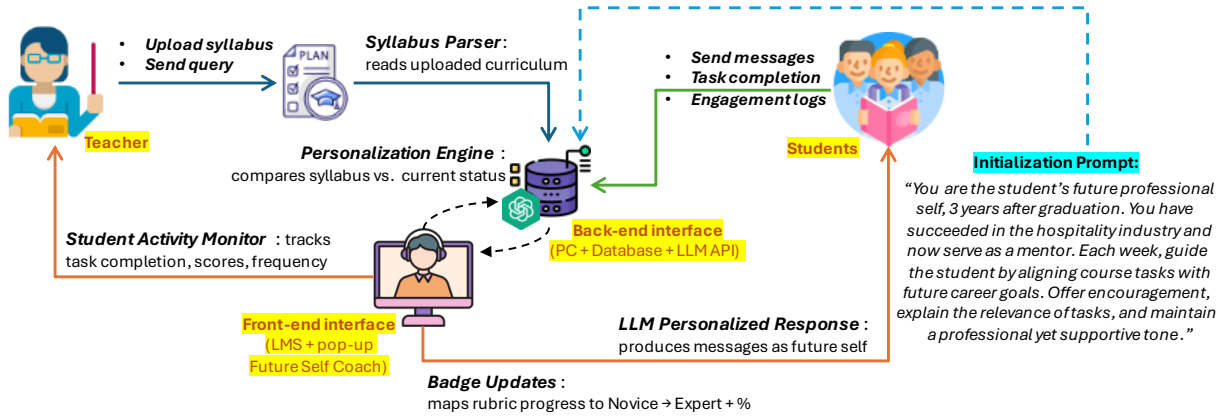


Figure 1. System architecture of the Future Self Coach.

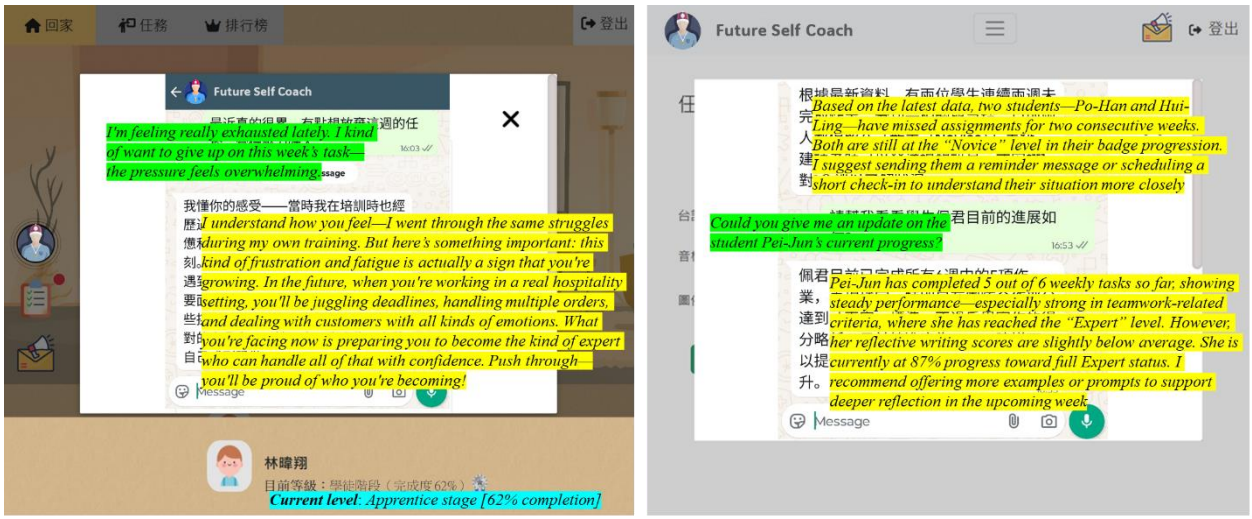


Figure 2. Example interactions with the Future Self Coach.

The system supports the development of grit and sense of agency by reinforcing key behaviors such as time management, tolerance for frustration, growth mindset, and the willingness to take on challenges. Through weekly AI interactions, students receive messages that help them reframe setbacks as growth opportunities and connect momentary difficulty to long-term identity development (see Figure 2, left). This future-facing narrative helps normalize frustration as part of the expert development process. Besides, Student progress is visualized through a badge-based identity dashboard. Learners begin as *Novice*, advance to *Apprentice*, and ultimately reach *Expert* level. Each badge maps directly to rubric dimensions, with progression tracked as a percentage of rubric mastery. This mechanism not only motivates continued effort but also gives students agency to set micro-goals and monitor their development in real time. Teachers play a foundational yet low-effort role in the system. They upload syllabi and its rubrics in structured formats and can monitor student progress via a dashboard. The AI assistant can also respond to teacher queries—for example, summarizing badge progress (see Figure 2, right). These dual-use capabilities allow the system to act as both a mentor for the student and a consultant for the teacher.

System implementation follows a modular and replicable architecture. The backend is built in Python and connects to LLM services via OpenAI API endpoints. The frontend chat interface is implemented in JavaScript and embedded into the LMS via iframe or plugin. Student progress data is securely stored in a cloud-based database, and badge updates are computed by a rubric-aligned scoring engine. The system supports containerized deployment using Docker, with optional Kubernetes orchestration for institutional scalability. Authentication follows OAuth 2.0 standards to ensure secure and role-based access.

4. Experimental Design

4.1 Participants, Ethics, and Learning Content

A quasi-experimental study was conducted in Taiwan with 97 undergraduate students majoring in Hospitality Management, all of whom were preparing for roles in Japanese restaurant service. To ensure group comparability, students were stratified based on their most recent course grades, then randomly assigned to one of three independent groups. Experimental Group A (LMS with the Future Self Coach) included 34 students (20 female, 14 male); Experimental Group B (LMS with a general-purpose AI assistant using API access to ChatGPT) included 32 students (19 female, 13 male); and the Control Group (standard LMS) included 31 students (17 female, 14 male). Unlike Group A, the assistant in Group B did not initiate role-based coaching. Its replies were primarily task-oriented—providing factual answers or clarifications without future-oriented framing or emotional scaffolding. This contrast allowed the study to isolate the effects of identity-based mentorship featured in Group A.

All participants had similar academic backgrounds, were around 21 years old, and were taught by the same instructors to control for instructional variability. Group assignments were completely independent, and no cross-group interaction occurred during the study. The research was approved by the university's institutional review board [IRB No.: NCUEREC-110-011], and all participants provided informed consent. They were informed of their right to withdraw at any point without penalty, and all privacy protocols were strictly followed.

The learning content was designed by the course instructor in the form of structured service scripts, aligning closely with both the course objectives and students' prior training. Scenarios simulated professional roles in Japanese restaurant settings—such as service staff, cashiers, and floor managers—emphasizing not only technical procedures but also customer interaction, emotional adaptability, and real-time problem solving. For example, one script tasked students with handling an emotionally upset customer while role-playing as a cashier who had made a billing error, requiring both empathy and accountability. These scenarios were designed to reflect authentic workplace challenges, supporting both skill development and professional identity formation in line with the MoE approach.

4.2 Instruments and Data Analysis

To evaluate learning achievement (**RQ1**), students completed a pre-test and post-test aligned with the course objectives. The two versions included variations in question phrasing and order to reduce memorization effects. The test consisted of 13 items, scored from 0 to 100, and demonstrated high internal reliability (Cronbach's α = .863). Learning gains across groups were compared using ANCOVA, with pre-test scores as covariates to control for baseline differences.

Grit (**RQ2**) was measured using a 10-item, five-point Likert scale adapted from Duckworth et al. (2007), showing strong internal consistency (Cronbach's α = .88). A sample item includes: "*I finish whatever I begin*." This reflects the sustained passion and perseverance component central to the grit construct. Sense of agency (**RQ3**) was assessed using a 13-item, five-point Likert scale adapted from Tapal et al. (2017), with reliability at Cronbach's α = .78. A representative item is: "*I am in full control of what I do*." This captures the individual's belief in their capacity to influence outcomes through their own actions. One-way ANOVAs were performed separately to compare group differences in both grit and agency scores.

To assess study frequency (**RQ4**), system log data were extracted from the LMS to capture the number of active learning sessions per student per day. Daily averages were calculated for each group, and engagement trends were visualized over time to identify behavioral differences.

In addition, semi-structured interview questions (Al Hakim et al., 2025) were conducted to explore student perceptions and motivational factors influencing their willingness to engage with the learning system.

4.3 Procedure

The experiment was conducted over an eight-week period and fully embedded within students' regular classroom training. The system was designed to complement conventional instruction, allowing seamless integration into scheduled topics without disrupting the existing curriculum. While this study focused on an eight-week implementation, the system itself is scalable and can be extended to support full-semester use, including repeated cycles of learning outcome demonstrations.

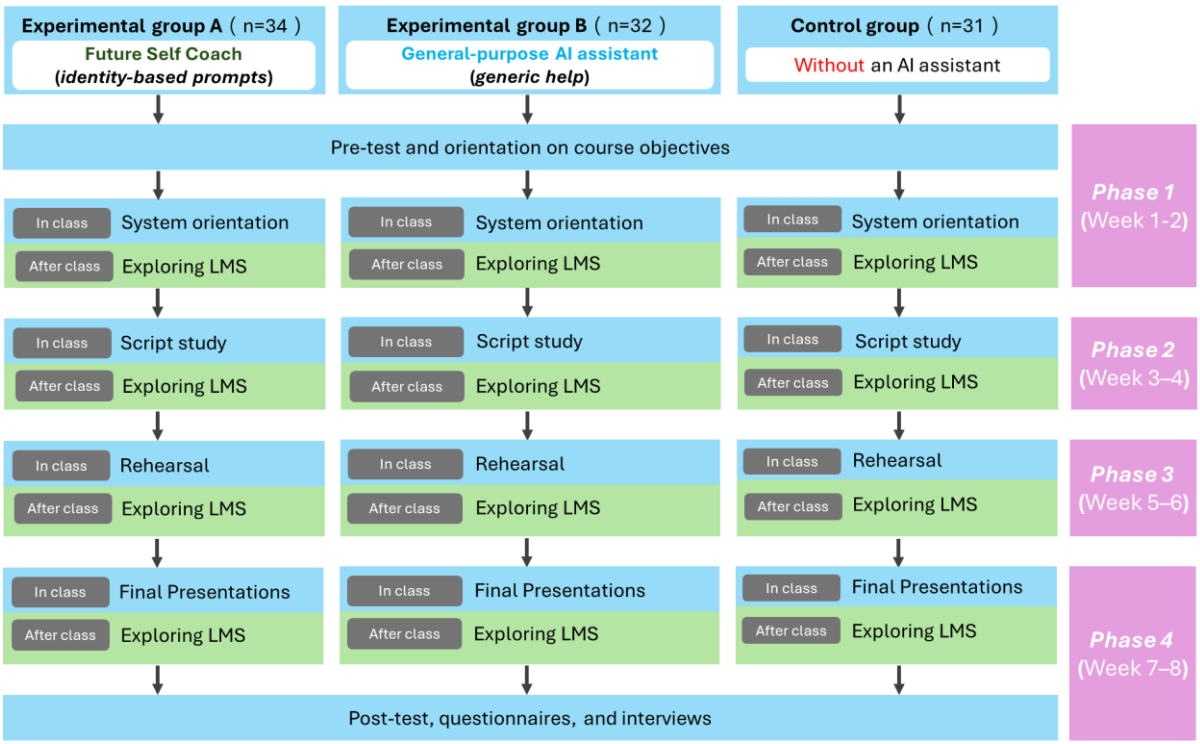


Figure 3. Experimental procedure.

The intervention was structured into four sequential phases, each lasting two weeks. Sessions were held once a week for 100 minutes, in accordance with institutional guidelines. Each phase introduced new tasks aligned with the syllabus, allowing students to gradually build their skills while maintaining a consistent instructional rhythm.

As shown in Figure 3, the experimental procedure included both in-class and out-of-class components. Students progressed through phases involving LMS onboarding, scenario-based script study, rehearsal sessions, and final role-play presentations—mirroring authentic professional development cycles in the hospitality industry.

The nature of system interaction varied by group. Students in the Future Self Coach group received identity-framed guidance, responding in the voice of their future professional self. For example, when a student typed, "*I got confused during the team role-play—I didn't know how to respond in Japanese*," the system replied, "*As your future self managing an international service team, I remember moments like this too. Staying calm and asking teammates for backup helped me handle language gaps professionally. You've just experienced a real challenge experts often face*." In contrast, students in the conventional AI assistant group received generic replies such as: "*That sounds difficult. Try reviewing your script or asking your teacher*." These differences reflect the distinction between identity-based mentorship and standard task-level support.

5. Results and Discussion

5.1 Learning Performance

An ANCOVA was conducted on students' post-test scores, with pre-test scores as the covariate and group intervention as the independent variable. An initial ANOVA showed no significant difference in pre-test scores among the groups ($F = 0.417$, $p > 0.05$), indicating baseline equivalence. The assumption of homogeneity of regression slopes was met ($F = 1.268$, $p > 0.05$), justifying the use of ANCOVA.

Table I shows the ANCOVA results, revealed a significant effect of group assignment on post-test scores ($F = 13.42$, $p < 0.001$), with a large effect size (partial $\eta^2 = 0.21$). Pairwise comparisons showed that students in the Future Self Coach group significantly outperformed those in the AI assistant group and the control group. The AI assistant group also performed significantly better than the control.

Table 1. *ANCOVA results for students' learning performance*

Group	*N*	Mean	SD	Adj. Mean	F-value	Pairwise Comparisons	Effect Sizes (η_p^2)
Exp. A	34	83.1	5.1	84.6	13.42*	A > B	0.21
Exp. B	32	77.8	6.2	78.3		A > C	
Control	31	73.1	4.9	72.8		B > C	

*$p < 0.05$

These results suggest that the personalized, identity-based feedback provided by the Future Self Coach enhanced students' conceptual understanding and application. The causal mechanism may involve a sequence of motivational and cognitive processes: identity-framing increased perceived relevance, which enhanced focus and effort, leading to deeper processing and improved learning outcomes (Burnette et al., 2020). By aligning weekly tasks with future professional goals, the coach helped students sustain attention and make connections across lessons. These patterns can be interpreted through the lens of identity-based motivation theory, which emphasizes the role of long-term goal alignment in driving learning persistence (Oyserman, 2024). One student remarked, "*It felt like I was being guided by my future self. That made the lessons more serious and made me want to do better*." Another noted, "*The coach reminded me what kind of expert I want to become, and that helped me push through even when I didn't feel like studying*."

5.2 Grit

A one-way ANOVA was performed to compare grit scores among the three groups (see Table 2 below). The analysis revealed significant differences ($F = 7.45$, $p < 0.05$), with a moderate effect size (partial $\eta^2 = 0.11$). Pairwise tests showed that the Future Self Coach group reported significantly higher grit than both the AI assistant group and the control group, while the difference between Group B and the control was not statistically significant.

Table 2. *ANOVA results for students' grit and sense of agency*

Measure	**Group**	***N***	**Mean**	**SD**	**F-value**	**Pairwise Comparisons**	**Effect Sizes (η_p^2)**
Grit	Exp. A	34	83.1	5.1	7.45*	A > B	0.11
	Exp. B	32	77.8	6.2		A > C	
	Control	31	73.1	4.9			
Sense of Agency	Exp. A	34	83.1	5.1	6.97*	A > B	0.12
	Exp. B	32	77.8	6.2		A > C	
	Control	31	73.1	4.9			

*$p < 0.05$

The increased grit observed in Group A likely stems from the system's ability to contextualize challenges within a long-term identity narrative. When weekly obstacles were framed as necessary steps toward becoming an expert, students were more likely to persist through setbacks (Derakhshan et al., 2025). This approach resonates with Duckworth et al. (2007), who emphasize that purpose and meaning are central to sustained effort. In contrast, Group B's generic AI lacked future-oriented framing, and students in the control group had no additional support. Interview responses reflected this pattern. One student shared, "*When I felt stuck, the coach reminded me why I was doing this. It helped me not give up*." Another said, "*It felt like someone believed in my future version, so I kept pushing*."

5.3 Sense of Agency

Sense of agency was also assessed via one-way ANOVA (see Table 2 above), revealing significant group differences ($F = 6.97$, $p < 0.05$), with a moderate effect size (partial $\eta^2 = 0.12$). Pairwise analysis indicated that Group A scored significantly higher than both Group B and the control group, while the difference between Group B and the control was not significant.

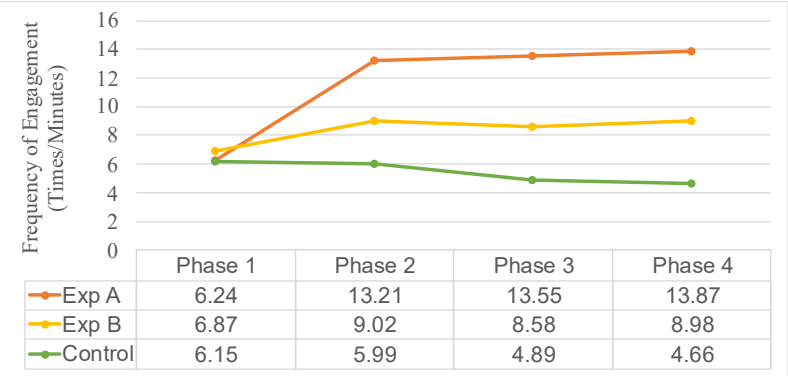
The enhanced sense of agency in the Future Self Coach group can be attributed to the badge-based identity dashboard. By translating rubric criteria into visual badge progress (Novice → Apprentice → Expert), students could track and influence their trajectory in real time. This feedback loop promoted autonomy, as students set micro-goals and monitored their own growth (Al Hakim et al. 2025). The design aligns with Self-Determination Theory, where perceived control over outcomes supports intrinsic motivation (Deci & Ryan, 2020). One participant explained, "*Seeing my badge level change made me feel like I had control. I knew what I needed to work on*." Another noted, "*The AI helped me connect my actions to progress*."

5.4 Engagement Frequency

Analysis of LMS activity logs revealed significant differences in engagement frequency across groups. Students in the Future Self Coach group engaged in more consistent out-of-class activity, averaging 11.72 weekly sessions, compared to 8.36 in the AI assistant group and 5.42 in the control group. Fig. 4 illustrates that Group A's engagement remained stable and slightly increased over time, peaking during rehearsal and presentation phases. In contrast, Groups B and C showed more variable activity, with noticeable drops after the script study phase.

This consistent engagement in Group A stemmed from the Future Self Coach's identity-framed prompts, which framed each week as a step toward real-world expertise. These emotionally resonant messages offered motivational nudges beyond formal instruction,

helping embed learning into daily routines. As one student shared, "*The coach kept checking in and reminding me why the task mattered in the real world. That kept me going even when I was tired*."



	Phase 1	Phase 2	Phase 3	Phase 4
Exp A	6.24	13.21	13.55	13.87
Exp B	6.87	9.02	8.58	8.98
Control	6.15	5.99	4.89	4.66

Figure 4. Observed study frequency among groups.

6. Conclusion, Limitations, and Future Works

This study introduces a new direction for using LLM-powered AI in education through a scalable and psychologically grounded framework that embeds identity-based mentorship into existing course structures. By transforming standard syllabi and rubrics into interactive, future-self-aligned coaching dialogues, the system not only improved learning performance, but also significantly enhanced students' grit, sense of agency, and study frequency. The Future Self Coach repositions AI from a reactive content assistant to a proactive identity guide, showing how syllabus-driven personalization can shift learners' mindsets and behaviors over time. These results mark an important step toward rethinking AI in education—not just as a tutor, but as a partner in professional identity development. The integration of Mantle-of-the-Expert pedagogy into LMS environments demonstrates how expert-role framing can support both hard and soft skill acquisition while maintaining instructional scalability. By embedding motivation into weekly check-ins, the system promotes emotional engagement and professional self-concept—without adding to teacher workload.

While promising, this study has limitations. It was conducted in a single discipline within a specific cultural and institutional context (hospitality training in Taiwan), and over a short-term deployment. Moreover, although the rubric-aligned badge system and conversational coaching provided robust support, the system did not yet adapt task difficulty or pacing in real time. Future work should explore longer-term use, cross-disciplinary replication, and integration with motivational scaffolds like peer coaching, social comparison, or gamified goal-tracking. Enhancements such as behavioral analytics, emotional state recognition, and embodied AI agents may unlock richer, more adaptive interactions. Ultimately, this work lays the foundation for AI mentors that are not only intelligent but personally meaningful—capable of evolving with learners and sustaining expert identity over time.

References

Al Hakim, V. G., Paiman, N. A., & Rahman, M. H. S. (2024). Genie-on-demand: A custom AI chatbot for enhancing learning performance, self-efficacy, and technology acceptance in occupational health and safety for engineering education. *Computer Applications in Engineering Education*, *32*(6), e22800. https://doi.org/10.1002/cae.22800.

Al Hakim, V. G., Yang, S. H., Wang, J. H., Ku, Y. H., Li, Y. J., Chang, C. K., & Chen, G. D. (2025). Robots facilitate self-talk: effects on ownership, self-efficacy, and learning performance. *Interactive Learning Environments*, 1-22. https://doi.org/10.1080/10494820.2025.2465444.

Beauvais, C. (2025). 'Your Future Self Will Thank You!': Creating future childhood memories in time-capsule activity books for children. In *Children, Young People and the Future*. Routledge.

Burnette, J. L., Pollack, J. M., Forsyth, R. B., Hoyt, C. L., Babij, A. D., Thomas, F. N., & Coy, A. E. (2020). A growth mindset intervention: Enhancing students' entrepreneurial self-efficacy and career development. *Entrepreneurship Theory and Practice*, *44*(5), 878-908.

Chen, C. Y., Juan, Y. S., Wang, J. H., Yang, S. H., & Chen, G. D. (2024, July). Integrate an AI Chatbot-Based Learning Butler Digital system to enhance Students' Grit and Growth Mindset for Improving Learning Outcomes. In *2024 IEEE International Conference on Advanced Learning Technologies (ICALT)* (pp. 21-25). IEEE.

Derakhshan, A., Solhi, M., & Azari Noughabi, M. (2025). An investigation into the association between student-perceived affective teacher variables and students' L2-grit. *Journal of Multilingual and Multicultural Development*, *46*(3), 798-814.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92*(6), 1087–1101. https://doi.org/10.1037/0022-3514.92.6.1087.

Feng, L., & Papi, M. (2020). Persistence in language learning: The role of grit and future self-guides. *Learning and Individual Differences*, *81*, 101904.

Ghafouri, M. (2024). ChatGPT: The catalyst for teacher-student rapport and grit development in L2 class. *System*, *120*, 103209.

Gubareva, R., & Lopes, R. P. (2020). Virtual Assistants for Learning: A Systematic Literature Review. *CSEDU (1),* 97-103.

Haggard, P. (2017). Sense of agency in the human brain. *Nature Reviews Neuroscience*, *18*(4), 196-207.

Harrington, C., & Thomas, M. (2023). *Designing a motivational syllabus: Creating a learning path for student engagement*. Taylor & Francis.

Heathcote, D., & Bolton, G. (1994). *Drama for Learning: Dorothy Heathcote's Mantle of the Expert Approach to Education. Dimensions of Drama Series*. Heinemann, 361 Hanover St., Portsmouth, NH 03801-3912.

Hershfield, H. E. (2011). Future self-continuity: How conceptions of the future self transform intertemporal choice. *Annals of the New York Academy of Sciences, 1235*(1), 30-43. https://doi.org/10.1111/j.1749-6632.2011.06201.x.

Kirschner, P. A., & Hendrick, C. (2024). *How learning happens: Seminal works in educational psychology and what they mean in practice*. Routledge.

Kosma, G. (2021). The Effect of Drama Methods on Enhancing Students' Emotional Engagement in the Distance EFL Learning Context; the Mantle of the Expert Paradigm. *Uluslararası Beşeri Bilimler ve Eğitim Dergisi*, *7*(15), 1-16.

Labadze, L., Grigolia, M., & Machaidze, L. (2023). Role of AI chatbots in education: systematic literature review. *International Journal of Educational Technology in Higher Education, 20*(1), 56.

McGivney, E. (2025). Interactivity and identity impact learners' sense of agency in virtual reality field trips. *British Journal of Educational Technology*, *56*(1), 410-434.

McTear, M. (2022). *Conversational ai: Dialogue systems, conversational agents, and chatbots*. Springer Nature.

Mittal, U., Sai, S., & Chamola, V. (2024). A comprehensive review on generative ai for education. *IEEE Access*.

Oyserman, D. (2024). Identity-Based Motivation and the Motivational Consequences of Difficulty. *Social and Personality Psychology Compass*, *18*(12), e70028.

Pataranutaporn, P., Winson, K., Yin, P., Lapapirojn, A., Ouppaphan, P., Lertsutthiwong, M., ... & Hershfield, H. E. (2024, October). Future you: a conversation with an AI-generated future self reduces anxiety, negative emotions, and increases future self-continuity. In *2024 IEEE Frontiers in Education Conference (FIE)* (pp. 1-10). IEEE.

Reeve, J. (2024). *Understanding motivation and emotion*. John Wiley & Sons.

Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary educational psychology*, *61*, 101860.

Sougleridi, E. I., Kopsidas, S., Vavougios, D., Avramopoulos, A., & Kanapitsas, A. (2023, September). Embedding AI into LMS and eLearning Platforms. In *International Conference on Interactive Collaborative Learning* (pp. 363-368). Cham: Springer Nature Switzerland.

Taylor, A., & Carlson, J. (2025). Comparing Who We Are to Who We Could Be: How Future Self-Images Influence Consumer Choices. *Psychology & Marketing*.

Taylor, T. (2025). Mantle of the Expert: Using Story Worlds to Engage Young Minds. *Childhood Education*, *101*(3), 6-13.

Zhao, X., Dichtl, F. F., & Foran, H. M. (2022). Predicting smoking behavior: intention and future self-continuity among Austrians. *Psychology, Health & Medicine*, *27*(5), 1042-1051.

Investigating the Effectiveness of Explanations in Active Video Watching

Raul Vincent LUMAPAS[a,b*], Antonija MITROVIC[a], Matthias GALSTER[a,c], Sanna MALINEN[a], Jay HOLLAND[a]
[a]*University of Canterbury, New Zealand*
[b]*Ateneo de Davao University, Philippines*
[c]*University of Bamberg, Germany*
*raulvincent.lumapas@pg.canterbury.ac.nz

Abstract: The use of video-based learning (VBL) has increased in recent years due to its effectiveness in distance learning. However, challenges in VBL, such as passive learning and low levels of engagement, continue to exist. Active Video Watching (AVW) addresses these challenges by supporting students' engagement and learning via features such as personalized nudges and visualizations. Furthermore, the AVW platform used in our research utilizes an AI model to classify comments students write on videos into several comment quality categories. Explanations of how comments are classified may enable students to understand the comment quality and, subsequently, write better comments. To determine whether explanations of comment quality would benefit students, we conducted a quasi-experimental study wherein we compared data collected from software engineering students at the Ateneo de Davao University (control group, 55 students using the AVW platform without explanations) to data collected from a comparable course at the University of Canterbury (experimental group, 45 students who received explanations). This study included a survey on students' perceptions on explanations. Results show that students in the experimental group, especially those who accessed explanations for multiple comments, wrote more higher-quality comments. Results of the survey show students are satisfied with explanations, particularly regarding the clarity, correctness, and the level of detail of explanations. This research contributes to using explanations in VBL platforms to increase learning and engagement.

Keywords: Video-Based Learning, Active Video Watching, Explanations, Explainable Artificial Intelligence

1. Introduction

Videos have become a popular and widely used method of learning (Fyfield et al., 2019). Video-based learning (VBL) is flexible, easy to use, provides opportunities for self-regulated learning, improves student attention, and increases motivation and engagement (Chatti et al., 2016, Dimitrova and Mitrovic, 2022, Seo et al., 2021). Furthermore, supplementing VBL with Artificial Intelligence (AI) can present more opportunities for engagement and learning (Seo et al., 2020). For example, Mohammadhassan et al. (2020) used a machine learning module to classify the quality of comments students write on videos and provide adaptive nudges to students, encouraging them to write better comments.

As with many VBL platforms, despite the perceived benefits of AI, the decision-making processes and the reasoning behind these AI predictions made are not visible to learners (Porayska-Pomsta et al., 2023, Edwards and Veale, 2017). These concerns have paved the way for explanations in AI in Education (AIED) systems and the adoption of explainable artificial intelligence (XAI) techniques (Farrow, 2023). Different studies have been conducted related to explanations and XAI in Education (XAI-Ed) (Türkmen, 2024), but explanations in AI-supported VBL have been scarce. With VBL being a popular method for teaching (Sablić et al., 2021), effective implementation of VBL, especially when AI is integrated, is vital (Seo et al., 2020). Explanations is a way to strengthen the use of VBL in learning. Moreover, it is also

important to evaluate explanations (Markus et al., 2021) and identify if these fit the needs of learners (Adadi and Berrada, 2018). This highlights the importance of exploring explanations of AI features in VBL.

This research investigates the effectiveness of integrating explanations of the AI-supported automatic assessment of comment quality in a platform for Active Video Watching (AVW). AVW-Space, the AVW platform we used for this experiment, automatically assesses the quality of comments students write. In previous studies with AVW-Space, the feedback from students included the need for explanations of the AI-generated comment quality. In this study, we added explanations and investigated their effects explanations on learning, engagement, and satisfaction. We aim to provide additional insights into the effects of explanations in VBL and contribute to the ongoing studies related to explanations in the area of AIED. We address the following research questions:

RQ1. Do explanations increase engagement and foster constructive behavior of learners? When explanations are available, we expect that students will engage with them, by accessing explanations of the quality of their comments, and also engage more with the platform (Hypothesis H1). Furthermore, while engaging with explanations, students would learn about how to write high-quality comments. With this, we expect more constructive learners in the experimental group (Hypothesis H2).

RQ2. Do explanations in AVW contribute to student learning? We anticipate that explanations of comment quality will have a positive effect on students' learning (Hypothesis H3). As mentioned, it is expected that explanations would clarify the reasons for comment quality and will prompt students to make more reflective or regulating comments.

RQ3. Are students satisfied with explanations in AVW? We will look into student feedback on explanations and identify if students have a positive perception of explanations to answer RQ3.

2. Related Work

2.1 Automatic Assessment of Comment Quality in AVW-Space

One effective form of video-based learning is Active Video Watching. AVW supports engagement in video-based learning by supporting self-regulated and reflective learning through a controlled video-watching environment (Mitrovic et al., 2016). AVW is scalable and provides a wide range of features to provide engagement when watching videos (Dimitrova and Mitrovic, 2022). AVW-Space is an AVW platform primarily used for soft-skills training (Mitrovic et al., 2019, Mitrovic et al., 2017). The platform allows teachers to create a space and integrate videos sourced from YouTube (Mitrovic et al., 2016). Using the space created by their teacher, students can watch videos and write comments.

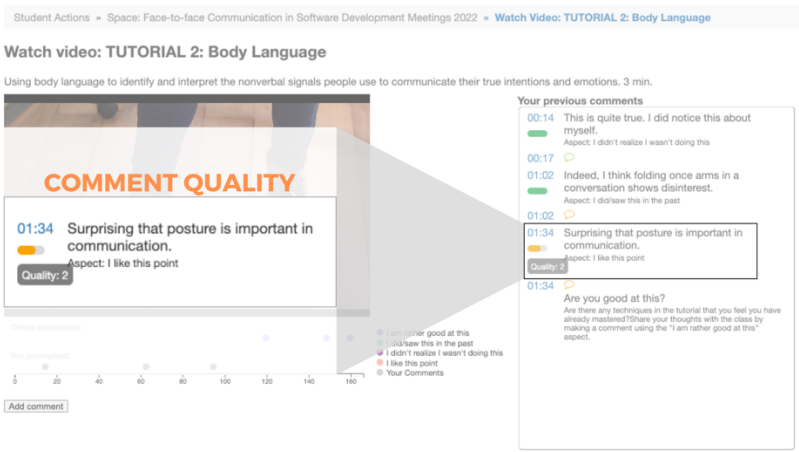


Figure 1. Comment quality feature in AVW-Space

The platform supports nudges (or personalized reminders) (Mitrovic et al., 2019), and includes automatic assessment of comment quality. This also is used for quality nudges (personalized reminders related to the quality of students' comments) (Mohammadhassan et al., 2022). Figure 1 shows a comment quality indicator, which is added to each comment to motivate students to write better comments (Mohammadhassan et al., 2020).

Mohammadhassan et al. (2020) proposed a quality scheme for comments on videos, with category one being of the lowest quality and category five as the highest. This labeling scheme is used in developing machine learning (ML) models to automatically assess the quality of students' comments (see Table 1). Comments in the first two categories are pedagogically undesirable and are considered low-quality comments as they do not convey deep thinking and reflection on the video content. The last three categories are considered high-quality, which convey critical thinking (category 3), reflection on past experiences or behaviors (category 4), or self-improvement (category 5) (Mohammadhassan et al., 2020).

Table 1. *Comment categories from Mohammadhassan et al. (Mohammadhassan et al., 2020)*

	Category	Definition
1	Affirmative, negative, off-topic	Comments which are irrelevant or merely affirmative/negative with no explanation
2	Repeating	Comments which only repeat the video content
3	Critical and analytical	Comments which mention points that are implicitly covered in the video or show critical thinking on the content of the video.
4	Reflective	Comments in which the learner reflects on their behaviour and previous experience or knowledge on giving presentations.
5	Regulating	Comments where the learner decides what they would do to improve themselves in future.

Machine learning is used to categorize user comments into quality categories. Currently, random forest is used for the assessment of comment quality as it was proposed by Mohammadhassan et al. (2020) as the best method among different classifiers when classifying comments from early studies in the same platform. Comments from previous experiments in AVW-Space were used to train and test the model. Predictions about the quality of comments from the model are used to generate quality nudges. Explaining this prediction can help students understand what good-quality comments are and help them formulate better comments.

2.2 Explanations in AIED and VBL

There have been several efforts to integrate explanations into AIED systems (Fiok et al., 2022). As AIED systems support different approaches to learning, explanations are personalized to the features of these systems (Zapata-Rivera and Arslan, 2024). Khosravi et al. (2022) presented a framework for XAI-ED, determined the application of XAI in different educational systems, and investigated four AIED systems integrating XAI. An example is RiPPLE by Darvishi et al. (2020), which uses explainable automated feedback on poor peer reviews that is supported by NLP models. Case studies like this aid the development of explanations in AIED systems, although there is need to explore further personalizing explanations (Khosravi et al., 2022). Personalizing explanation features, especially in educational systems, is highly recommended in literature (Conati et al., 2021, Kouki et al., 2019).

There is limited research on XAI in video-based learning. Although other XAI-ED frameworks or methods can be applied to AI-supported VBL systems, it is important to consider the personalization of explanations in VBL. Chen et al. (2023) proposed a similar recommendation to consider XAI in VBL to enhance social transparency and to develop a framework and expanded XAI framework for educational systems.

3. Explanations of Comment Quality in AVW-Space

We previously surveyed undergraduate students in order to identify the explanation requirements and the basis for the design, presentation, and timing of explanations in AVW (Lumapas et al., 2024). The findings showed that (1) students were primarily interested in explanations on how the comment quality is determined, (2) wanted to see explanations shown next to comments and (3) wanted explanations to be shown on-demand.

To generate explanations of comment quality, we use the comment text, its quality category, linguistic/textual features (e.g. nouns, pronouns, verbs, positive/negative words), and the selected *aspect* (teacher-defined micro-scaffolds for directing students to reflect on key points of the video or their experience (Mitrovic et al., 2017)). This information is then passed to a textual analysis feature, which checks for any contractions (e.g., "*I'm*", "*Don't*") or negative words or phrases (i.e., "*do not"*, "*cannot"*). This process is necessary to identify if the comment is positive or negative, as the explanations are also determined based on the sentiment. The text of the explanation is determined by the quality, the textual features, and the general sentiment (positive/negative) of the comment. For high-quality comments, as there is no need for further support, a simple explanation is added (e.g., "This comment is a quality 4 (high-quality) comment because you made a personal reflection related to empathy").

Explanations of low/medium quality comments (quality 1 or 2) provide more information. These explanations consist of three parts: (1) an explanation of the comment quality, (2) a recommendation on how to improve the comment, and (3) an example of a better-quality comment, as illustrated in Figure 2. In most cases, category 1 or 2 comments are short; generally, they simply agree/disagree (e.g., *"Good"*, *"I agree"*), are off-topic, vague, repeating video text or content, or repeat the aspect text. Depending on the issues identified regarding the comment, we add an appropriate explanation. For example, if a comment just simply agrees *"This is correct"*, we add this to the explanation text: *"This comment is a quality 1 (low-quality) comment because you made a short affirmative/negative or off-topic comment."* The recommendation in this case is *"You can make this comment better by relating it to your or your team's experience"*. The explanation also contains an example of a higher-quality comment, such as *"Our team takes input from everyone when assigning tasks and knowing if they are capable or comfortable."*

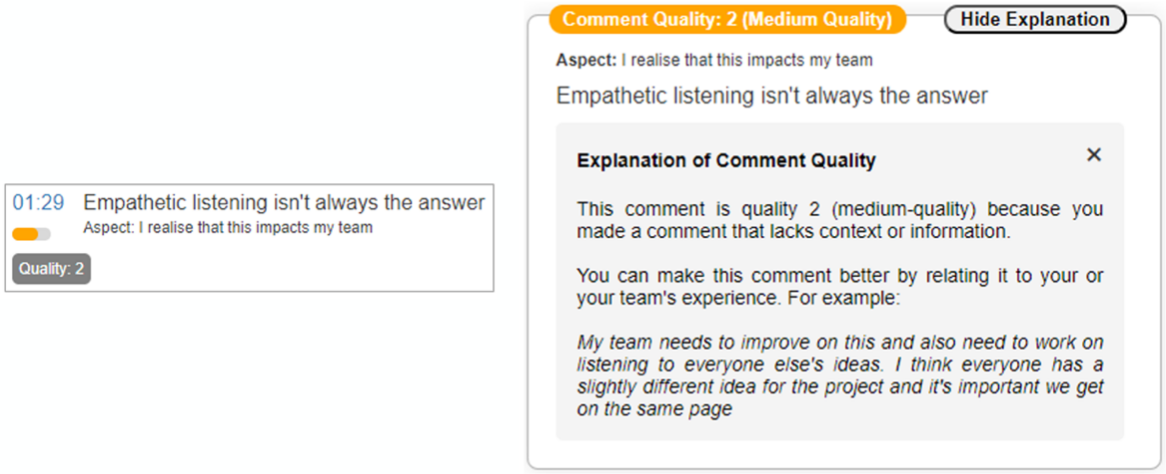


Figure 2. Comments without explanations (left) and comments with explanation (right)

Figure 2 shows the differences between comments without explanations and comments with explanations. This example shows a comment with a low quality with the commenter not relating to the aspect selected. The explanation shows comment quality by color (red for quality 1, yellow for quality 2, and green for high quality) and also text. Students can access and hide explanations.

4. Experiment Design

We conducted a quasi-experimental study with two groups of students at the Ateneo de Davao University (ADDU) in the Philippines, and at the University of Canterbury (UC) in New Zealand in 2024 (*Human Research Ethics Approval 2023/96/LR-PS and 2024/79/LR-PS*). Although students were from two different countries, they were enrolled in similar software engineering courses. Furthermore, although there are cultural differences between the two groups, both universities used English as the language of instruction. In both courses, students worked on a project in groups. Students who completed all training tasks were given 5% of the final course grade.

Explanations were introduced in AVW-Space used by UC students (experimental group) while the ADDU students did not have explanations (control group). Both groups used the platform for four weeks, and the learning materials and process were identical. The platform offered six five-to-ten-minute videos related to empathy in the context of software engineering. Students provided informed consent or opted out of participating in the study. In the latter case, students could complete the soft skills training and get course credit, but we did not collect data about them.

Two surveys were administered within AVW-Space, both of which contained a timed question (one minute) where students were asked to list all concepts they know about empathy. Responses were marked automatically using the ontology developed for empathy skills by the research team. The resulting Conceptual Knowledge score from Survey 1 (CK1) is the pre-test score, while CK2 from Survey 2 is the post-test score.

Survey 1 contained demographic questions. Survey 2 included seven questions about the explanation feature. Among these is an explanation satisfaction instrument derived from Hoffman et al.'s (2023) measures for XAI. There are seven statements in this instrument, and we used a Likert scale to identify if students agree/disagree with the statement (1 - Strongly Disagree to 5 - Strongly Agree). Another question asked students to select the reasons for seeking explanations. We also added open-ended questions on explanations, the purpose of re-reading/re-accessing explanations, and questions on their experience and suggestions. The platform also logged data about all actions students performed, such as log-ins, comments written and access to explanations.

We used the ICAP framework (Chi and Wylie, 2014), which classifies students based on their overt behaviors into Interactive, Constructive, Active, and Passive categories. Passive learners only watched the videos without writing comments. Active students wrote comments of low quality; they repeated the information provided in the videos without any reflection or elaboration. Constructive students reflected on the videos and added new information not explicitly taught. We do not consider the interactive category from ICAP (which promotes interaction between students), as interaction was not supported by the platform.

5. Results

We analyzed the data about the students who completed both surveys and watched at least one video (Table 2). There are 54 students in the control and 44 in the experimental group.

Table 2. *Number of responses to survey and completion of soft-skills training.*

Group	Survey 1	Watched Videos	Survey 2
Control	78 (69.6%)	**54 (48.2%)**	56 (50.0%)
Experimental	50 (94.3%)	49 (92.5%)	**44 (83.0%)**

5.1 Effects of Explanations on Engagement (RQ1)

Table 3 shows how students interacted with the platform. The experimental group accessed the platform more often than the control group, measured both by the number of days the platform was used and the number of sessions, which confirms our hypothesis H1. There are no significant differences between the groups in the number of nudges, comments made, and

the number of videos watched and commented on. However, there is a significant difference in the number of high-quality comments.

Table 3. *Engagement mean (STDEV)*

	Control (54)	**Experimental (44)**	**t-test**
Days on the Platform	4.22 (1.819)	8.44 (3.057)	t=8.507, p<.001
Sessions	4.78 (2.229)	9.89 (4.350)	t=7.534, p<.001
Nudges	26.94 (11.758)	27.82 (7.243)	t=.436, p=.332
Videos	5.91 (0.680)	6.00 (0.00)	t=.912, p=.182
Comments	10.72 (7.720)	10.87 (4.971)	t=.108, p=.457
Videos Commented	5.83 (0.855)	6.00 (0.00)	t=1.294, p=.099
High-Quality Comments	5.78 (5.62)	10.36 (4.75)	t=4.323, p<.001

We categorized the students post-hoc into the ICAP categories. We classified students who watched without making any comments as *"Passive"*. Consistent with findings from the previous studies with AVW-Space, interventions introduced to support student learning reduce the number of passive students (Mohammadhassan et al., 2020; 2022; Mitrovic et al., 2019). This is observed in our study too, as there were no passive students in either group. Meanwhile, to differentiate between the *"Constructive"* and *"Active"* students, we used the median number of high-quality comments made on the videos (Mohammadhassan et al., 2022). The median number of high-quality comments was seven, which means that those who commented eight or more high-quality comments were classified as *"Constructive"*. Those who wrote up to seven high-quality comments are considered as *"Active"*.

We performed a Chi-square test of homogeneity between the two groups and the ICAP categories. Results show a significant difference (Chi-square = 10.433, p=.001) with effect size (Phi) of .325 (p=.001). In the post-hoc analysis, we observed that there was an increase in the number of Constructive students in the experimental group in comparison to the control group (Table 4). The higher number of participants from the experimental group who are constructive learners confirms Hypothesis H2.

Table 4. *Numbers of constructive and active students in the two groups*

	Control (54)	**Experimental (44)**	**t-test**
Active	39 (86.7%)	18 (40.0%)	p<0.001
Constructive	15 (33.3%)	27 (60.0%)	p<0.001

We also investigated the access to explanations between the different ICAP categories in the experimental group. Table 5 shows that constructive students accessed more explanations in comparison to active students. Similarly, the number of unique comments for which the explanation was accessed is higher for constructive students than active students.

Table 5. *Comparing active and constructive students from the experimental group*

	Active (18)	**Constructive (27)**	**t-test**
Times Explanations is Accessed	1.72 (1.02)	4.63 (3.38)	t=3.536, p=.001
Unique Comments where Explanations was accessed	1.61 (0.85)	3.74 (2.40)	t=3.612, p=.001

5.2 Effects of Explanations on Learning (RQ2)

We first analysed whether the two groups are comparable in terms of their pre-existing knowledge of empathy. The Kolmogorov-Smirnov test confirmed that the CK1 scores were not distributed normally ($p < .05$). The comparison of CK1 scores for the two groups using the Mann-Whitney U-Test shows that the distribution of the CK1 scores is the same across the two groups of students ($U = 1397.5$, $p = .198$). Table 6 details the differences in the CK1 and

CK2 scores in the two groups, showing the mean scores and the standard deviation. In the control group, there is little difference between the CK1 and CK2. There is a higher CK2 score as compared to the CK1 score for the experimental group.

Table 6. *Conceptual Knowledge Scores - Mean and Standard Deviation*

Group	CK1	CK2
Control	9.13 (8.74)	9.20 (8.78)
Experimental	7.82 (7.82)	11.27 (6.81)

We also analyzed the number of comments, especially high-quality ones, as a potential indicator of learning. As mentioned earlier, most students who accessed explanations were constructive. Table 7 shows the average number of comments for each comment category (with categories 3-5 shown together as High-quality) by group. We observed that the experimental group had a higher average of high-quality comments than the control group ($t=4.323$, $p<0.001$). Comments from the control group, on the other hand, are divided between category two and high quality. There were also several category 1 comments in the control group, while the experimental group did not have any. Students from the experimental group consistently have high-quality comments, with 95.7% of the comments being quality 3-5. The control group, on the other hand, only 53.9% of the comments are high-quality.

Table 7. *Mean and Standard Deviation based on the comment categories for both groups*

Group	Category 1	Category 2	High Quality
Control	0.07 (0.26)	4.87 (4.93)	5.78 (5.62)
Experimental	0 (0.00)	0.47 (0.97)	10.36 (4.75)

As presented in Table 4, there are more constructive students in the experimental group than in the control group. We analysed whether there are differences as to the number of comments between the different ICAP groups in both groups (Table 8). The constructive students in both groups have the same average number of high-quality comments (13.07 and 13.00, respectively). This would mean that explanations of comment quality enable students to write comments that are critical, reflective, and regulating.

Table 8. *Comment Quality per ICAP categories Mean and Standard Deviation*

Group		Category 1	Category 2	High Quality
Control	Active	0.10 (0.31)	4.56 (4.91)	2.97 (2.22)
	Constructive	0.00 (0.00)	5.67 (5.05)	13.07 (5.24)
Experimental	Active	0.00 (0.00)	0.22 (0.73)	6.39 (0.61)
	Constructive	0.00 (0.00)	0.630 (1.079)	13.00 (4.455)

There are significantly more high-quality comments in the experimental group compared to the control group. Most students in the experimental group were constructive students, who and wrote critical, reflective, or regulating comments. As presented in Table 8, even active students in the experimental group wrote mostly high-quality comments. Furthermore, we can observe that in the experimental group, the mean for low-medium quality comments is very low (0 for category 1 and 0.22 and 0.630 for active and constructive students respectively). This confirms Hypothesis H3.

5.3 Explanation Satisfaction (RQ3)

Results of Survey 2 conducted with the experimental group show positive perceptions of explanations. Table 9 shows that most participants agreed that the explanations were presented clearly (mean=4.05), are correct (mean=3.89), and provide sufficient detail (mean=3.89). Participants also noted that the explanations are predictable (mean=3.83). Predictability is not necessarily negative (Chamola et al., 2023, Zhou et al., 202). Finding the

explanations as predictable can mean that the participants understood the content of the explanations well. We also looked into responses on whether explanations motivated students to write better comments. For that specific question, the mean score is 3.59. Students had varying opinions, which is reflected in the standard deviation value of 1.11, the highest in the explanation satisfaction questions. Most students, however, answered *"somewhat agree"* (25%) or *"strongly agree"* (27%) to the question on motivation.

Table 9. *Explanation Satisfaction Results*

Question	Mean (SD)
I understand how comment quality is determined based on the explanations	3.68 (1.02)
Explanations have sufficient detail	3.89 (0.80)
Explanations are predictable	3.82 (0.91)
Explanations are correct	3.89 (0.86)
Explanations are presented clearly	4.05 (0.74)
Explanations motivate me to write better comments	3.59 (1.11)
Explanations are useful	3.64 (0.86)

We examined the responses to the open-ended question, "How would you describe your experience with explanations?" Thirty-one (31) out of 44 participants (70%) had a positive perception towards explanations. Most students who accessed explanations found them beneficial. Explanations helped students to reflect on the task and comments made (e.g., *"Overall it was a positive experience. It helped give a brief period of reflection on each comment. It was nice to 'quantify' the quality of each comment, even if it wasn't fully clear as to why a rating was given"*). Some comments indicate self-regulation, where students mention that explanations help them improve their future comments (e.g., *"It was good! When my comment was of lower quality, I would always check the explanation and read the reason so I could improve my comments next time. One in particular that I remember is when I was commenting, often I would just write about how I agreed with what was being said, but the explanations helped to realize that a better comment would be self-reflective on my own experience"*). Other than improving on the tasks associated with the training, some students do mention the benefits of explanations in improving their understanding of the concepts mentioned in the videos (e.g., *"Helpful as I was able to gain a deeper understanding of what we are learning and how I can apply this to my personal participation in teamwork and in <Course Code>."*).

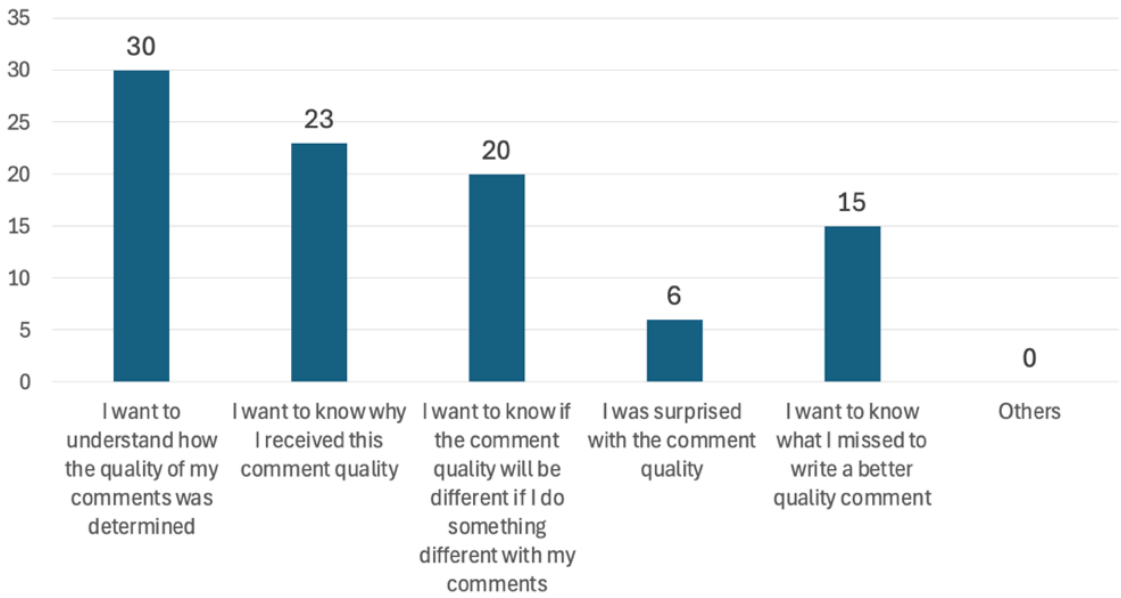


Figure 3. Reasons for seeking explanations

Interestingly, one participant mentioned the straightforward nature of the explanations is a reason to not access explanations for other comments anymore, mentioning, *"I think I quickly grasp onto explanation and don't need too much to rethink about it too much or have to go back to re-read an explanation".* This might mean that students might have understood the explanation when they saw it the first time and did not need to re-read explanations.

We also asked students in Survey 2 to select the reasons why they asked for explanations (Figure 3). Students could select multiple responses and could also select "Other" to specify a reason that was not listed. 29 out of the 44 students (65.9%) in the experimental group added more than one reason. Meanwhile, no students indicated their own reasons outside of the selection. Students mostly accessed explanations to understand how the comment quality was determined. However, some students also indicated that they accessed explanations to improve their comments.

6. Discussion and Conclusions

We proposed adding explanations of the quality of comments written for videos in VBL, in order to support learning and increase engagement. We found that explanations fostered constructive behavior, reflected in an increase in the number of students in the experimental group. We also identified that students who accessed explanations spent more time (i.e., days and sessions) on the platform and wrote more high-quality comments. It was observed that explanations affect engagement and are helpful for learners. Although we did see an increase in the number of high-quality comments in the experimental group, we plan to conduct more studies to further identify whether there is a direct effect of explanations on learning.

One limitation of our study is the small size of the experimental group, which is not sufficient to identify causation of explanations on learning. However, we can conclude that explanations can aid students in writing reflective and regulating (high-quality) comments and support constructive learning. Furthermore, students recognized the positive impact of explanations. Explanations promote reflection when commenting on the videos and help students improve their comments. To identify a potential causal relationship between our intervention and learning gain, more participants are needed. It is important to note that the current participants identified that explanations were presented clearly and had sufficient detail. Although, an analysis of the CK1 scores indicated that there is no significant difference in prior knowledge between the two groups, it is noted that differences in cultural and educational backgrounds may contribute to differences in the CK2 scores. Another method of measuring increase in learning should be used in future experiments.

As discussed earlier, the literature on explanations in VBL is limited. Our research contributes to exploring explanations as a way to support and boost learning and engagement in VBL. This research contributes to the expanding pool of research looking at explainability in learning systems, given the importance of explanations in promoting trust and transparency in a time when AI is more utilized in education.

References

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6:52138–52160.

Chamola, V., Hassija, V., Sulthana, A. R., Ghosh, D., Dhingra, D., and Sikdar, B. (2023). A review of trustworthy and explainable artificial intelligence. IEEE Access 11: 78994-79015.

Chatti, M. A., Marinov, M., Sabov, O., Laksono, R., Sofyan, Z., Fahmy Yousef, A. M., and Schroeder, U. (2016). Video annotation and analytics in CourseMapper. Smart Learning Environments 3: 1-21.

Chen, S., Situ, J., Cheng, H., Kirst, D., and Huang, Y. (2023). Mirrorus: Mirroring peers' affective cues to promote learner's metacognition in video-based learning. Proc. Human- Computer Interaction, 7(CSCW2):1–25.

Chi, M. T. and Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. Educational psychologist, 49(4):219–243.

Conati, C., Barral, O., Putnam, V., and Rieger, L. (2021). Toward personalized XAI: A case study in intelligent tutoring systems. Artificial intelligence, 298:103503.

Darvishi, A., Khosravi, H., and Sadiq, S. (2020). Utilising learner sourcing to inform design loop adaptivity. In European Conference on Technology-Enhanced Learning, pp. 332–346. Springer.
Dimitrova, V. and Mitrovic, A. (2022). Choice architecture for nudges to support constructive learning in active video watching. International Journal of Artificial Intelligence in Education, 32(4):892–930.
Edwards, L. and Veale, M. (2017). Slave to the algorithm? why a right to an explanation is probably not the remedy you are looking for. Duke L. & Tech. Rev., 16:18.
Farrow, R. (2023). The possibilities and limits of XAI in education: a socio-technical perspective. Learning, Media and Technology, 48(2):266–279.
Fiok, K., Farahani, F. V., Karwowski, W., and Ahram, T. (2022). Explainable artificial intelligence for education and training. The Journal of Defense Modeling and Simulation, 19(2):133–144.
Fyfield, M., Henderson, M., Heinrich, E., and Redmond, P. (2019). Videos in higher education: Making the most of a good thing. Australasian Journal of Educational Technology, 35(5):1–7.
Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2023). Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance. Frontiers in Computer Science, 5:1096257.
Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., and Gašević, D. (2022). Explainable artificial intelligence in education. Computers and Education: Artificial Intelligence, 3:100074.
Kouki, P., Schaffer, J., Pujara, J., O'Donovan, J., and Getoor, L. (2019). Personalized explanations for hybrid recommender systems. 24th int. conf. Intelligent User Interfaces, pp. 379–390.
Lumapas, R. V., Mitrovic, A., Galster, M., Malinen, S., Peiris, P., and Holland, J. (2024). Integrating explanations in active video watching. In International Conference on Computers in Education. Asia-Pacific Society for Computers in Education.
Markus, A. F., Kors, J. A., and Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. Journal of biomedical informatics, 113:103655.
Mitrovic, A., Dimitrova, V., Weerasinghe, A., and Lau, L. (2016). Reflective experiential learning: Using active video watching for soft skills training. In Proceedings of the 24th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education.
Mitrovic, A., Gordon, M., Piotrkowicz, A., and Dimitrova, V. (2019). Investigating the effect of adding nudges to increase engagement in active video watching. In Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part I 20, pages 320–332. Springer.
Mitrovic, A., Gostomski, P., Herritsch, A., and Dimitrova, V. (2017). Improving presentation skills of first-year engineering students using active video watching. In Annual Conference of the Australasian Association for Engineering Education (28th: 2017: Sydney), pages 809–816. Australasian Association for Engineering Education Sydney.
Mohammadhassan, N., Mitrovic, A., and Neshatian, K. (2022). Investigating the effect of nudges for improving comment quality in active video watching. Computers & Education, 176:104340.
Mohammadhassan, N., Mitrovic, A., Neshatian, K., and Jonathan, D. (2020). Automatic assessment of comment quality in active video watching. Int. Conf. on Computers in Education, pp. 1–10.
Porayska-Pomsta, K., Holmes, W., and Nemorin, S. (2023). The ethics of ai in education. In Handbook of Artificial Intelligence in Education, pp. 571–604. Edward Elgar Publishing.
Sablić, M., Mirosavljević, A., and Škugor, A. (2021). Video-based learning (vbl)—past, present and future: An overview of the research published from 2008 to 2019. Technology, Knowledge and Learning, 26(4):1061–1077.
Seo, K., Dodson, S., Harandi, N. M., Roberson, N., Fels, S., and Roll, I. (2021). Active learning with online video: The impact of learning context on engagement. Computers & Education, 165:104132.
Seo, K., Fels, S., Yoon, D., Roll, I., Dodson, S., and Fong, M. (2020). Artificial intelligence for video-based learning at scale. 7th ACM Conference on Learning@ Scale, pages 215–217.
Türkmen, G. (2024). The review of studies on explainable artificial intelligence in educational research. Educational Computing Research, 63(2): 277-310.
Zapata-Rivera, D. and Arslan, B. (2024). Learner modeling interpretability and explainability in intelligent adaptive systems. In Mind, Body, and Digital Brains, pp. 95–109. Springer.
Zhou, L., Moreno-Casares, P. A., Martínez-Plumed, F., Burden, J., Burnell, R., Cheke, L., Ferri, C., Marcoci, A., Mehrbakhsh, B., Moros-Daval, Y., et al. (2023). Predictable Artificial Intelligence. arXiv preprint, arXiv:2310.06167.

Applicability of Facial Video-Based Alertness Estimation Model in Real Online and In-Person Classrooms

Atsushi ASHIDA[a*], Ryosuke KAWAMURA[b], Hideaki HAYASHI[a] & Hajime NAGAHARA[a]
[a]*D3 Center, The University of Osaka, Japan*
[b]*Fujitsu Research of America, Inc., The United States*
*ashida@ids.osaka-u.ac.jp

Abstract: Accurately capturing learners' internal states is essential in modern educational environments to support effective teaching and the design of appropriate learning content. Various methods have been proposed for estimating such internal states, with recent approaches increasingly relying on machine learning techniques. However, models trained under specific conditions often fail to generalize to different instructional settings. To be practically useful, these models should be applicable across diverse learning environments, including e-learning, synchronous video-based instruction, and in-person classes. Nonetheless, few studies have evaluated internal state estimation models by transferring them from their training conditions to substantially different and operationally realistic educational settings. In this study, we investigate the effectiveness of our internal state estimation model by applying it in authentic classroom settings. The model estimates learners' alertness based on facial video, specifically focusing on the eye region, and was trained using data collected in a controlled e-learning environment. The evaluation was conducted using data obtained from real educational contexts, including both synchronous online classes and traditional in-person classroom sessions. This setting allowed us to assess the model's robustness across multiple instructional formats that reflect current hybrid learning environments. We compared the model's predictions to human-annotated labels indicating whether learners appeared to be asleep, using receiver operating characteristic (ROC) curves and area under the curve (AUC) scores. The results suggest that the model has the potential to function effectively even when applied to data collected in real-world instructional scenarios.

Keywords: Engagement, alertness, domain shift, computer vision, deep learning

1. Introduction

With societal shifts exemplified by the advent of Society 5.0, there is an increasing need for individualized learning environments that cater to learners' unique needs (Ghosh & Jermsittiparsert, 2024). To realize such personalized optimization, capturing learners' internal state with higher resolution than previously achieved is essential. Methods to assess these internal states typically involve analyzing behavioral and physiological data from learners.

In the field of learning analytics (LA), it is common practice to utilize behavioral data, such as learning logs derived from clickstreams, as they provide clear and quantifiable insights into cognitive aspects of learner behavior (Khor & Mutthulakshmi, 2023). However, such behavioral indicators primarily reflect cognitive dimensions of the learner's internal state. To enhance educational effectiveness, it is critical not only to address these cognitive dimensions but also to incorporate the learner's affective states. By integrating affective states alongside cognitive measures, more adaptive and individualized instructional strategies can be realized, ultimately leading to improved learning outcomes (e.g., D'Mello & Graesser, 2012).

Recent advancements in machine learning and deep learning have led to significant progress in developing technologies capable of capturing the learner's internal affective state. Such technologies hold promises for personalized education, offering deeper insights into

learner engagement and emotional responses. However, most existing approaches rely heavily on supervised learning, necessitating data collected from a specific, controlled environment. These supervised methods frequently encounter difficulties when deployed in real-world educational settings, as data obtained outside controlled conditions typically suffer from substantial variability due to external disturbances inherent to authentic educational settings. Consequently, models trained within a controlled environment often fail to generalize adequately when transferred directly to different, more dynamic educational contexts. A common approach to addressing this issue is to retrain models from scratch, requiring extensive data collection and annotation efforts in each new setting. This practice is not only resource-intensive but also exacerbates the cold-start problem, limiting the immediate applicability and scalability of affective state estimation technologies across diverse learning environments.

This paper aims to investigate the applicability of the learner's internal state estimation model, particularly focusing on the model's behavior when applied to data collected in actual classroom settings. Understanding which aspects of these models generalize effectively across varying educational contexts could provide critical insights for enhancing the reusability and applicability of machine learning models, potentially mitigating the challenges posed by the cold-start issue. The research question is "Can a learner's internal state estimation model trained in a controlled environment be feasibly applied and deployed in real-world educational settings?"

To address this research question, we first trained a model using data collected in a controlled e-learning environment. This model utilizes features extracted from the learner's eye region and estimates the learner's alertness levels. We then applied this model to data collected in an actual classroom setting. The data was collected from a synchronous online class conducted at a vocational school in collaboration with our research group. The model's predictions were evaluated by comparison with human annotations to determine its effectiveness in estimating the learner's alertness level in the wild. As an evaluation metric, we used the receiver operating characteristic (ROC) curve and the area under the curve (AUC) to assess the model's performance. The results indicate that the model can accurately estimate alertness levels, provided that the learner's face is properly captured by the camera. This suggests that the model has potential for real-world applications in educational settings.

Additionally, this study was approved by the ethics review committee of the University of Osaka. Data collection was conducted after explaining the purpose and methods to the participants and obtaining their informed consent.

2. Related Research

Prior research has mainly focused on cognitive aspects of learner modeling, such as knowledge representation or problem-solving behaviors. In personalized learning environments, considerable attention has been given to intelligent tutoring systems (ITSs) that monitor learners' states and adapt content accordingly (Wenger, 1987). These systems typically incorporate learner models such as the overlay model or buggy model (Carr et al., 1977; Greer et al., 2010), which aim to represent what the learner knows correctly (overlay model) or misunderstands (buggy model) about domain knowledge. These models are often domain-specific and not easily generalizable across subjects.

The field of learning analytics (LA) has offered more general, domain-independent insights by analyzing large-scale behavioral data. For example, Ma et al. (2022) investigated "jump back" behavior in digital reading, revealing patterns linked to cognitive engagement. However, the learner's internal state encompasses not only cognition but also affective responses, such as alertness or engagement (Kaur et al., 2021).

To address this, researchers have increasingly incorporated multimodal learning analytics (MMLA), combining video, audio, and physiological signals to capture emotional states (Blikstein & Worsley, 2016). Applications include EEG-based impasse detection (Yamamoto et al., 2021) and collaborative learning support (Sugimoto et al., 2017; Spikol et al., 2017), detection of comprehension or non-comprehension (Holmes et al., 2018), and identifying learners' thinking patterns for neurodivergent students. Our research group also

contributed to this area with models for alertness estimation based on facial video (Kawamura et al., 2021; Ashida et al., 2024). Despite their promise, MMLA methods often rely on data collected under controlled conditions for training data collection.

While MMLA methods have successfully captured the learner's internal states, most of these studies assume consistent training and deployment conditions. This assumption limits their ability to generalize across different instructional settings. More importantly, despite advancements in sensing and modeling technologies, relatively few studies have attempted to introduce such systems into actual classroom settings. This lack of in-situ implementation highlights a critical gap between controlled experimental validation and practical deployment. Our work aims to bridge this gap by applying a model trained in a controlled environment to real-world classroom data without retraining.

3. Alertness Estimation Model

3.1 Structure of The Model

In this section, the architecture of our developed alertness estimation model is described. This model focuses on learners' behavioral engagement, which is operationalized here as the degree of alertness (Fredricks et al., 2004). This model receives the facial video data of learners as input and outputs the estimated alertness level along with the time sequence. The alertness level is treated as a content-independent indicator of the learner's internal state, enabling its applicability across instructional contexts. Research studies revealed that the information from facial images is effective for recognizing the learner's inner state.

The structure of the alertness estimation model is illustrated in Figure 1. In our model, we utilize eye regions for estimating the alertness level of learners. In the preprocessing, first, the model extracts the frames from the input video data. The model estimates alertness levels based on 10 image frames extracted over a period of one second. Assuming an input video frame rate of 30 fps, frames are extracted from the video at intervals of three frames. After the frame extraction, the model extracts the eye region from the original image. Cropping the eye region consists of two phases. The former step is face detection using Dlib and OpenCV. The latter step is cropping the eye region based on facial landmarks by using Dlib. The eye region is cropped from the original image, and the cropped image is resized to a fixed size.

Let $S_{i:i+10}$ denote video segments from the i-th to $(i+10)$-th extracted frames, these 10 frames correspond to the one second of the input video and $y_{i:i+10}$be the alertness state during the period starting from the i-th frame and ending at $(i+10)$-th frame. In the recognition task, the alertness state during a period of the input sequence is the target of prediction. Therefore, we build a model that takes $S_{i:i+10}$ as input and predicts $y_{i:i+10}$.

The model focuses on the learner's alertness. The model consists of a convolutional neural network (CNN) for image feature extraction and a bidirectional gated recurrent unit (BiGRU) for learning time-directional dependencies. After preprocessing, the eye region images are then fed to a CNN-GRU model to extract features and estimate alertness state for each second. For the CNN component, we employed EfficientNet-B3 pre-trained on the ImageNet dataset (Tan & Le, 2019). Since the CNN-GRU processes ten frames simultaneously, the CNN features are concatenated and fed into fully connected (FC) layers consisting of 2560 and 256 units, respectively. The BiGRU module contains 256 hidden units. The final output is a probability score representing the likelihood of the learner alert.

3.2 Training Data Collection Environment

The training data were collected in a controlled e-learning environment. Data collection was conducted using class video lectures for undergraduate students. We used data from 53 undergraduate students, obtained with informed consent, while they were watching video lectures (Kawamura et al., 2021).

The built-in webcam of a laptop is used for face video data collection. The face video and screen capture are recorded by the software program Bandicam. The frame rate of the

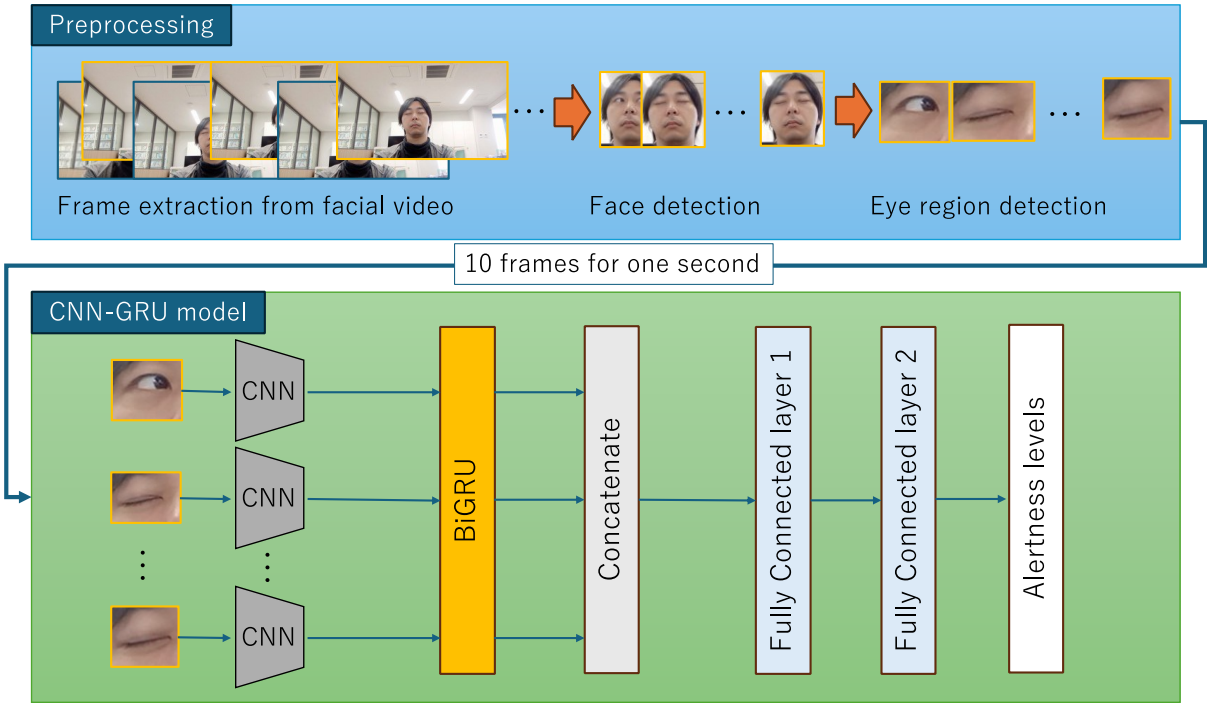


Figure 1. The structure of the model

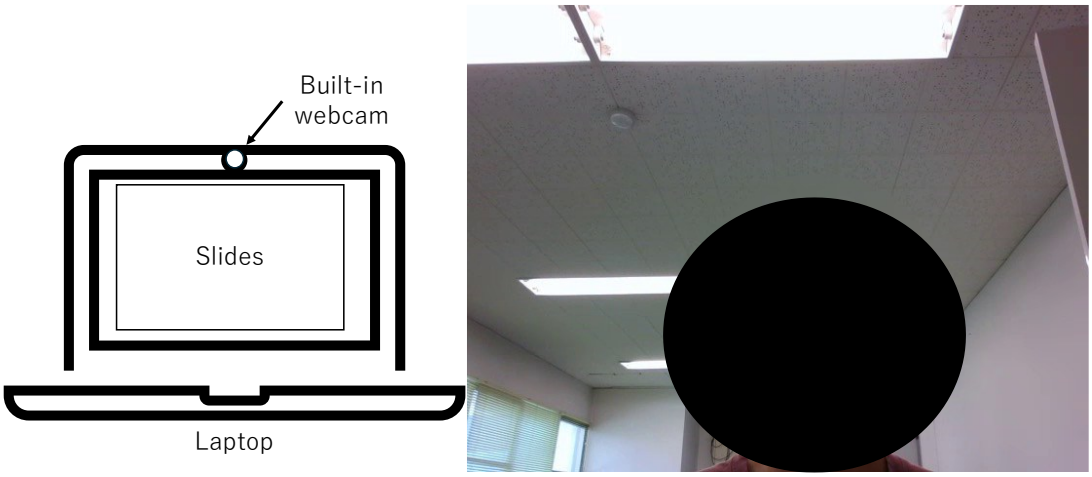


Figure 2. The image of face videos collection environment and actual collected data in an e-learning setting. The student's face is captured from the front.

facial video is 30 fps, and the resolution of the web-camera is 1280 x 720 pixels, and the screen capture is used to determine whether learners were engaged in learning. The collected facial video data were coded by an annotator. The code is the binary value (0: Asleep, 1: Awake) for every second through human observation. Data from four subjects were annotated by two additional annotators to investigate the reliability of annotations. The annotations among the three annotators matched in 93.8% of selected instances (Ashida et al., 2024).

Lighting conditions for capturing facial video data are bright, and the web camera built into the notebook computer used by the students to attend the video lectures was used, therefore, the students' faces were captured from the front. The sample of captured facial video data is shown in Figure 2. However, in consideration of privacy, individuals are not identified.

4. Data Collection in the Real Class Setting

In-the-wild data were collected at “OCA, Osaka College of Design and IT Technology,” a private vocational school. Informed consent was obtained from all participants. The target class name was “Outline of e-sports.” The class lasted 90 minutes. This class is a lecture-based session in which the instructor explains the current situation surrounding e-sports using slides. We conducted data collection four times. Three sessions were delivered online, and one session was delivered in person. The in-person session was conducted in a different physical classroom from that used for the online sessions. The images of the data collection environments are shown in Figures 3 and 4. The online setting environment is shown in Figure 3. The online lectures were conducted as synchronous classes using Zoom, with the instructor's video and slides displayed on the screen in front of the learners. Smartphones were set up below the display to record a video of the learners' faces. In the case of in-person classes (Figure 4), slides were projected onto a screen at the front of the classroom, and the instructor delivered the lecture next to it. Smartphones were placed on the desk in front of the learners to record a video of their faces.

Each setting of the lighting conditions differed from those in the training data (Figure 2). Table 1 shows the number of participants for each class. All sessions were delivered by the same instructor. The total number of participants was 17, and the total number of video data was 50.

We asked the learners who participated in the data collection to record facial videos during the classes using smartphones. Motorola G13 devices were employed as recording equipment. The videos were captured using the front-facing camera at a resolution of 1920 × 1080 pixels and an average frame rate of ~30 fps. The smartphones were mounted on clamp-style stands affixed to the desk in landscape orientation, and learners were instructed to adjust the position so that their faces remained within the field of view throughout the class (Figure 5). The orientation of the face and the angle of the camera differed from the training data collection situation, with the recording being taken from a lower oblique angle.

5. Experiment

5.1 Experimental Setting

As a preliminary evaluation of the model's applicability to real-class data, we employed the ROC curve and its AUC. To plot the ROC curve, ground-truth labels were required. After recording, the videos were annotated by human annotators who watched each recording and added alertness labels. The annotation task was divided among three annotators, and each video was annotated by a single annotator. The annotation format, consistent with that used for the training data, was binary (1 = awake, 0 = asleep). When an annotator was unable to determine the state, they marked the segment as “undecidable.”

We generated a dataset containing the model outputs and the corresponding human annotations and then created an ROC curve by plotting the true positive rate (TPR) against the false positive rate (FPR) while varying the threshold used to classify the model outputs as either awake or asleep. Data points for which the model failed to output an alertness level, owing to the absence of a detected face during the one-second interval, were excluded from the evaluation. Likewise, time-steps that annotators marked as “undecidable” were omitted from the ground-truth labels. Consequently, the ROC analysis was performed only on intervals for which both a model prediction and a definitive human annotation were available.

Table 1. The participants of the real data collection

Class	No. 1	No. 2	No. 3	No. 4	Total
Type	Online	Online	Online	In-person	-
Number of participants	12	14	12	12	50



Figure 3. The image and the photo of the facial video data collection environment in online settings. The learners attended the online classes and watched the display in front of them. Many of them set the smartphone under the monitor.

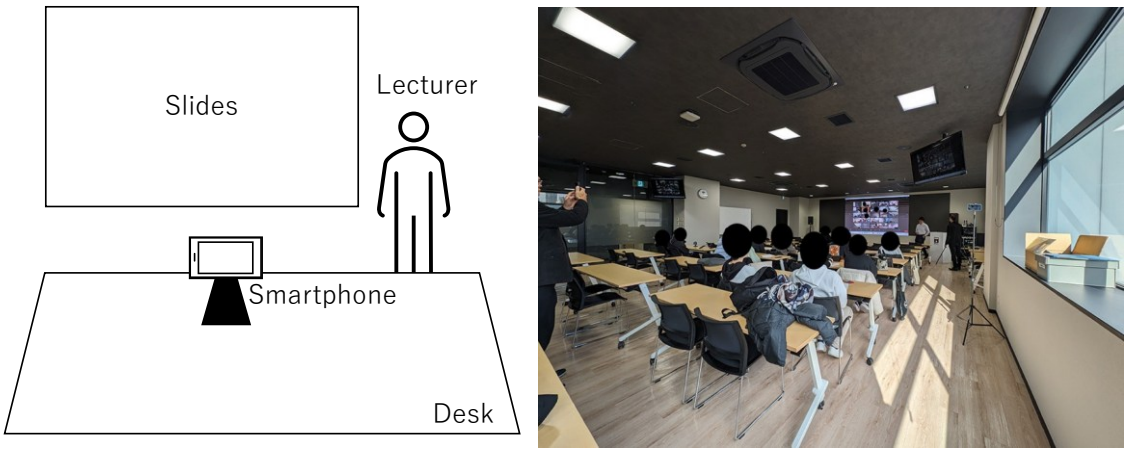


Figure 4. The image and the photo of the facial video data collection environment in an in-person setting. The classroom is different from Figure 3. Learners looked at the lecturer in front of them, and the angle of the face was also different from other settings.



Figure 5. The smartphone setting for data collection, the vertical angle is adjustable. The learners were asked to set the device in front of them and adjust the angle to be able to capture their faces.

5.2 Results and Discussion

We created ROC curves using the data from each of the four individual classes, as well as an ROC curve using the data combined across all classes. Figures 6 and 7 show the ROC curves for the overall dataset that includes data from all classes and each individual class. Corresponding AUC values are shown in Table 2.

In the figures, the orange continuous line indicates the ROC curve obtained from our alertness estimation model. The broken line indicates the 1:1 ratio between TPR and FPR. In other words, this line shows the result of the random classifier. Also, the closer the orange lines are to the top-left corner, the better the performance it indicates. The AUCs indicate the area under the ROC curves. The AUC takes a value between 0 and 1, with values closer to 1 indicating better classifier performance.

From figures 6 and 7, the AUCs were over 0.75 for all classes, and 0.84 for over all data. Various interpretations of the AUC exist (de Hond et al., 2022). In this paper, we adopted the criteria of Polo et al. They suggested that AUC values can be interpreted as follows: 0.5–0.6 (failed), 0.6–0.7 (worthless), 0.7–0.8 (poor), 0.8–0.9 (good), and > 0.9 (excellent) (Polo et. al., 2020). When applying this criterion to our results, the AUC for the overall dataset can be interpreted as good. These findings suggest that our model, which was trained on data collected in a controlled e-learning environment, has the potential to estimate learners' alertness levels in a real educational setting.

When evaluating the AUC for each class, Class No. 1 is categorized as "poor", Classes 2 and 3 as "excellent", and Class No. 4 as "good" based on the aforementioned AUC interpretation criteria. This indicates that the alertness estimation model has the potential to estimate the learner's alertness level regardless of the form of the class, online or in-person. Higher AUC values are expected for online classes compared to the in-person class, as the data collection conditions in the online setting are more similar to those used during training. Indeed, Classes 2 and 3, which were conducted online, show higher AUC values than Class 4, which was conducted in person. However, Class 1, also an online class, exhibits a lower AUC value than Class 4. Since Class 1 was the first dataset collected, it may have been influenced by factors such as participant unfamiliarity with the system or environmental inconsistencies, which could have affected model estimation accuracy.

Also, there are scenes in which the model failed to estimate the learner's alertness levels. For example, there were scenes where the learners concentrated on the class, but the model outputs a low alertness level. In this case, the model fails to estimate the correct alertness level. We need to investigate the model performance of the data in the e-learning

Table 2. The value of the AUCs

Class	No. 1	No. 2	No. 3	No. 4	Total
AUC	0.76	0.95	0.91	0.81	0.84

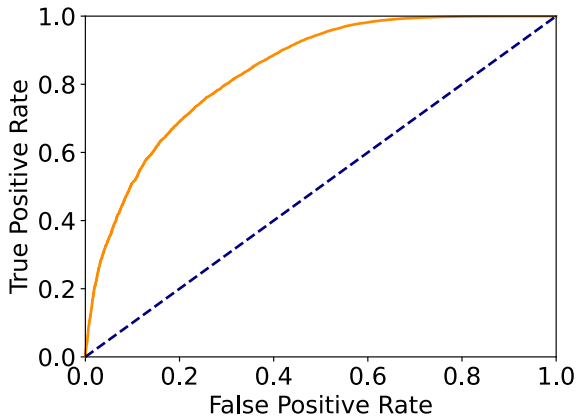


Figure 6. The ROC curve and AUC for overall data.

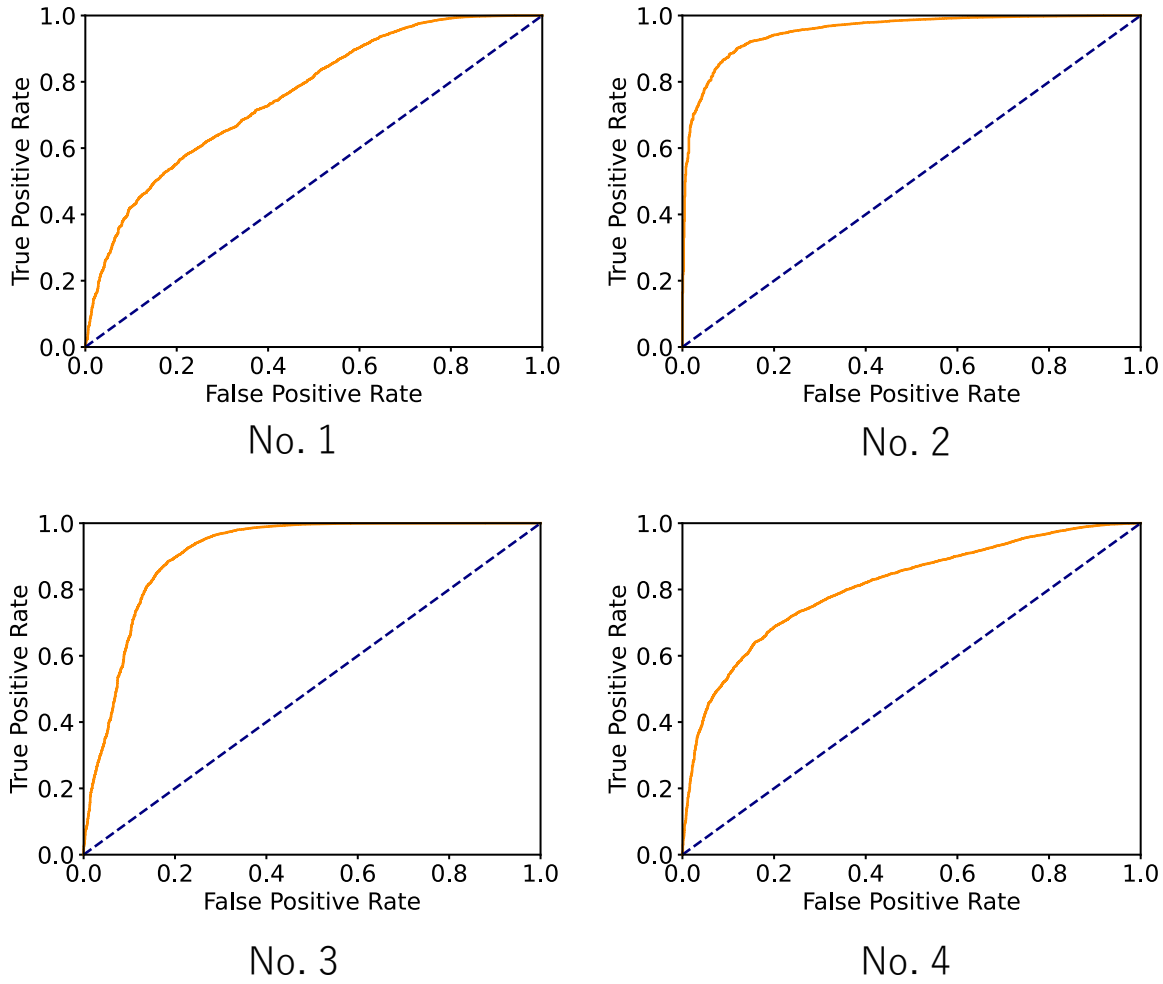


Figure 7. The ROC curves in class No. 1, No. 2, No. 3, and No. 4.

environment and compare it with data in the real educational environment to clarify whether the performance varies among environments. Also, we would like to clarify whether the additional training of the model using the data collected in this research improves the model's performance.

In the practical deployment of an alertness detection system, potential applications include promoting learners' metacognitive awareness when their alertness decreases or providing feedback to instructors by identifying students who require attention. In such cases, it is crucial to evaluate the system from the perspective of whether it can accurately detect learners with low alertness, which typically represents the minority class. Therefore, future evaluations should focus on metrics suitable for real-world implementation, such as plotting precision–recall (PR) curves and calculating PR-AUC values by treating the minority class, i.e., low-alertness learners, as the positive class.

6. Conclusion

This study investigated the applicability of a pre-trained alertness level estimation model to actual educational settings, which represent different environments from where the model was trained. The model used crop images of the area around the eyes from learners' facial videos as input and outputs the learners' alertness levels. It was trained on training data collected from learners taking e-learning courses in a controlled environment.

In the real-world environment, learners' facial videos were recorded during a total of four classes, including synchronous online classes and face-to-face classes. The alertness levels of the recorded videos were estimated by the alertness level estimation model, and the ground

truth was annotated by human annotators. For model evaluation, ROC-AUC was used as the evaluation metric. The evaluation results showed that the alertness estimation model achieved a reasonably good AUC score, suggesting that alertness estimation using images of the eye region, as employed by the model, has the potential to be used in actual classroom environments.

As future work, by fine-tuning the existing model with data collected in real-world environments and examining how much the model's performance improves, we can determine whether learning in a new environment is necessary. Furthermore, the usability of the data output by this model differs in actual classes. For instance, when identifying students with low alertness levels to provide feedback, higher performance in correctly classifying low-alertness data is preferable. Therefore, evaluating the model using metrics that emphasize the correct identification of positive instances, such as PR-AUC, can be considered. Additionally, as a future development, we aim to contribute to the improvement of actual learning environments by developing a dashboard that allows for the evaluation of learners and the classes themselves based on the metrics obtained from this model.

Acknowledgement

The authors would like to thank the staff and students of OCA, Osaka College of Design and IT Technology, for their participation and cooperation in the experiments.

This paper is partially supported by the Innovation Platform for Society 5.0 from the Japan Ministry of Education, Culture, Sports, Science and Technology, and JSPS KAKENHI Grant Number 24K16753.

References

Ghosh, U. K., & Jermsittiparsert, K. (2024). Personalised learning systems and the human touch in society 5.0. In Practice, Progress, and Proficiency in Sustainability, IGI Global, 211–232.

Khor, E. T., & Mutthulakshmi, K. (2023). A systematic review of the role of learning analytics in supporting personalized learning. *Education Sciences, 14*(1), 51.

D'Mello, S., & Graesser, A. (2012). AutoTutor and affective autotutor: learning by talking with cognitively and emotionally intelligent computers that talk back. ACM Transactions on Interactive Intelligent Systems, 2(4), 1–39.

Wenger, E. (1987). *Artificial intelligence and tutoring systems: Computational and cognitive approaches to the communication of knowledge.* Morgan Kaufmann Publishers Inc.

Carr, B., & Goldstein, I. P. (1977). *Overlays: A theory of modelling for computer aided instruction*.

Greer, J. E., & McCalla, G. I. (Eds.). (2010). Student modelling: The key to individualized knowledge-based instruction. Springer.

Ma, B., Lu, M., Taniguchi, Y., & Konomi, S. (2022). Exploring jump back behavior patterns and reasons in e-book system. Smart Learning Environments, 9(1), 1–23.

Kaur, P., Kumar, H., & Kaushal, S. (2021). Affective state and learning environment based analysis of students' performance in online assessment. *International Journal of Cognitive Computing in Engineering, 2*, 12–20.

Blikstein, P., & Worsley, M. (2016). Multimodal Learning Analytics and Education Data Mining: using computational technologies to measure complex learning tasks. Journal of Learning Analytics, 3(2), 220–238.

Yamamoto, S., Tobe, Y., Tawatsuji, Y., & Hirashima, T. (2021). In-process Feedback by Detecting Deadlock based on EEG Data in Exercise of Learning by Problem-posing. Proceedings of the 29th International Conference on Computers in Education, 21–30.

Sugimoto, A., Hayashi, Y., & Seta, K. (2017). Multimodal interaction aware platform for collaborative learning. Proceedings of the 25th International Conference on Computers in Education, 316–325

Spikol, D., Ruffaldi, E., Landolfi, L., & Cukurova, M. (2017). Estimation of success in collaborative learning based on multimodal learning analytics features. 2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT), 269–273.

Holmes, M., Latham, A., Crockett, K., & O'Shea, J. D. (2018). Near Real-Time Comprehension Classification with Artificial Neural Networks: Decoding e-Learner Non-Verbal Behavior. IEEE Transactions on Learning Technologies, 11(1), 5–12.

Wong, A. Y., Bryck, R. L., Baker, R. S., Hutt, S., & Mills, C. (2023). Using a Webcam Based Eye-tracker to Understand Students' Thought Patterns and Reading Behaviors in Neurodivergent Classrooms. LAK23: 13th International Learning Analytics and Knowledge Conference, 453–463.

Kawamura, R., Shirai, S., Takemura, N., Alizadeh, M., Cukurova, M., Takemura, H., & Nagahara, H. (2021). Detecting drowsy learners at the wheel of e-learning platforms with multimodal learning analytics. IEEE Access, 9, 115165–115174.

Ashida, A., Kawamura, R., Shirai, S., Takemura, N., Alizadeh, M., Hayashi, H., & Nagahara, H. (2024). Is internal state feedback in an e-learning environment acceptable to people? In *Proceedings of the 32nd International Conference on Computers in Education*, 89–94.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School Engagement: Potential of the Concept, State of the Evidence. Review of Educational Research, 74(1), 59-109.

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning, 6105–6114.

de Hond, A. A. H., Steyerberg, E. W., & van Calster, B. (2022). Interpreting area under the receiver operating characteristic curve. *The Lancet Digital Health*, 4(12), e853–e855.

Polo, T. C. F., & Miot, H. A. (2020). Use of ROC curves in clinical and experimental studies. Jornal Vascular Brasileiro, 19, e20200186.

Towards Scalable Annotation of Math Word Problems: Knowledge Component Tagging via LLMs and Sentence Embeddings

Chor Seng TAN[a*], Chengwei WEI[a], Jung-Jae KIM[a]
[a]*Institute for Infocomm Research (I^2R), A*STAR, Singapore*
*jjkim@i2r.a-star.edu.sg

Abstract: Mathematical word problems are a crucial component of mathematics education, requiring students to integrate multiple reasoning skills. Annotating these problems with knowledge components (KC) enables better personalized learning, adaptive tutoring, and AI-driven educational assessment. However, manual annotation is time-consuming and inconsistent, limiting the scalability of KC-based learning systems. In this work, we introduce a new labeled dataset of MWPs, derived from ASDiv and GSM8K, with KCs aligned to the Common Core mathematics framework. Using this dataset, we benchmark two different methods to perform automatic KC annotation without any labeled examples, namely LLM KC tagging and SBERT sentence embedding similarity scoring. Our results highlight key strengths and limitations of LLMs in this task, revealing challenges in consistency and reasoning alignment with human labels. We then show that SBERT-based similarity scoring underperforms LLM KC tagging, but can be significantly enhanced by combining the two methods, which addresses their respective limitations. This study provides critical insights into the feasibility of automated KC tagging, laying the foundation for future research in AI-assisted curriculum design and intelligent tutoring systems.

Keywords: Math word problems, automated knowledge component annotation, sentence embedding, large language model, natural language processing

1. Introduction

Mathematical word problems (MWPs) play a central role in mathematics education, requiring students to integrate numerical computation with problem-solving and logical reasoning. Unlike direct arithmetic exercises, MWPs demand an understanding of mathematical concepts, real-world application, and multi-step reasoning. As a result, MWPs are widely used in curriculum design, standardized testing, and adaptive learning platforms to assess students' problem-solving abilities (Verschaffel et al., 2020). However, understanding and solving MWPs effectively depends on mastering a set of underlying knowledge components (KCs) - fine-grained mathematical skills that correspond to cognitive subskills such as addition, multiplication, fraction manipulation, and proportional reasoning.

The concept of KCs originates from cognitive science and educational research, where learning is modeled as the progressive acquisition of fundamental skills (Corbett & Anderson, 1995). By tagging MWPs with appropriate KCs, educators and AI-driven tutoring systems can break down complex problems into their constituent skills, enabling personalized learning and targeted intervention (VanLehn, 2006). This approach is foundational to Intelligent Tutoring Systems (ITSs) - AI-driven platforms designed to provide real-time feedback, adaptive exercises, and automated assessments (Anderson et al., 1995; Pane et al., 2014). ITSs rely on KC-annotated datasets to track student progress, predict learning trajectories, and optimize instructional strategies. However, a key challenge in deploying such systems at scale is the manual effort required to annotate MWPs with KCs, which is often time-consuming and inconsistent.

To mitigate the challenges associated with mapping KCs to assessments, a variety of automated solutions employing machine learning and natural language processing (NLP) have been proposed. These approaches primarily use classification algorithms trained on annotated datasets to predict KC for math word problems. For example, Yaumauchi et al. (2023) used word n-gram features with a vector space model and random forest classifiers, while Shen et al. (2021) used a Task-adaptive Pre-trained BERT model. While these methods reported good results, they rely on substantial, high-quality labeled datasets, which are often scarce due to the labor-intensive nature of manual annotation and the need for domain-specific expertise. Moreover, these models are typically trained on a fixed set of curriculum standards – such as the Common Core – and thus exhibit limited generalizability. When applied to a different curriculum or an alternative set of KCs, their performance may degrade significantly, necessitating retraining or reannotation, which limits their scalability and practical utility.

Recent advances in natural language processing, particularly in sentence embedding models and large language models (LLMs), offer promising pathways to automate the KC annotation process. Sentence-BERT (SBERT) (Reimers et al., 2019) for example allows for semantic comparison between math problems and skill descriptions using dense vector representations without the need for labelled data. Meanwhile, LLMs like the GPT-family of models from OpenAI and Llama models from Meta exhibit strong reasoning capabilities and internalized knowledge of curriculum frameworks, enabling them to infer education concepts directly from problem content. These tools may significantly reduce human effort and enhance the scalability of ITSs. However, their ability to consistently identify and map MWPs to the correct underlying KCs has not been fully explored.

To address this gap, we introduce a new dataset of MWPs from ASDiv (Miao et al., 2020) and GSM8K (Cobbe et al., 2021), annotated with knowledge components aligned to the Common Core mathematics framework. Using this dataset, we benchmark several automated KC tagging approaches. First, we use cosine similarity between SBERT-encoded problem-solution pairs and KC descriptions to score and rank candidate KCs. Second, we evaluate GPT-4o mini, a lightweight yet high-performance LLM from OpenAI (2024), under multiple prompting strategies, to assess its ability to predict the appropriate KC directly, relying solely on its internalized knowledge of curriculum standards without external KC descriptions. Lastly, we introduce a hybrid method that prompts GPT-4o mini to generate KC descriptions from the MWP and then match these generated descriptions to KCs using SBERT similarity scoring.

The main contributions of this work are as follows: 1) we present two datasets based on problems from ASDiv and GSM8K with manual KC annotations from experts based on the Common Core State Standards (CCSS), 2) we benchmark the performance of different NLP models, namely sentence embedding using SBERT and LLM annotation using OpenAI's GPT-4o mini, to perform automated KC annotation without any labelled examples, 3) we show that the two methods of SBERT and LLM annotation, including reasoning-based LLM annotation, show high degree of inconsistency across the two datasets, and thus are unreliable to be used as input to manual validation, and 4) we demonstrate that by combining the two methods we can address the issues of the two methods.

Our study provides a comprehensive evaluation of the strengths and limitations of both embedding-based and generative approaches to KC tagging. Our results show that each method has complementary advantages and combining their outputs yields improved performance. This work contributes to AI-driven education research and the broader goal of automating curriculum-aware reasoning tasks, paving the way for more effective AI-powered learning systems.

2. Background and Related Work

Automating the tagging of Knowledge Components (KCs) for math word problems using natural language processing (NLP) methods has been an ongoing area of research for over a decade. One of the earliest efforts was from Cetintas et al. (2009), who trained a support vector machine (SVM) classifier to differentiate Multiplicative Compare and Equal Group

problems from other types of math problems drawn from a fourth-grade math textbook. Subsequent studies expanded the classification scope to a broader range of KCs. For instance, Karlovcec et al. (2012) demonstrated a text mining approach using a SVM classifier as well as a search engine-based approach with a k-nearest neighbors (KNN) classifier to tag math problems with labels drawn from a pool of 106 KCs. Meanwhile, Pardos et al. (2017) used a skip-gram approach to match math problems with labels from a pool of 198 KCs. While these early works reported promising results, later analysis revealed that the datasets used, which were from the ASSISTments platform, were based on templates and vulnerable to overfitting. Patikorn et al. (2019) pointed out that the models were exploiting spurious textual cues – such as keywords and formatting patterns – rather than learning mathematical concepts, thus limiting generalizable to more diverse datasets.

More recent works leverage language models such as BERT and GPT that are based on the Transformer architecture, which allows for modelling bidirectional dependencies and generating context-sensitive representations of texts. (Vaswani et al., 2017). Tan et al. (2024) showed that training a RoBERTa classifier model to predict KC can achieve strong performance on a diversified math word problem dataset. However, this approach still requires annotated data for training and may not generalize well across different curricula. Moore et al. (2024) employed GPT-4 to generate KCs for multiple-choice questions in Chemistry and E-learning. They developed an ontology to cluster questions that assess similar KCs based on their content. However, their method does not rely on any existing set of KCs but instead uses LLMs to generate a new set of KCs based on the problem texts and compares the outputs to KCs generated by humans. Li et al. (2024) also explored the use of LLMs for KC tagging, particularly focusing on prompt engineering, zero-shot and few-shot settings, and comparing different LLM models. They managed to achieve high accuracies with their LLM annotations. Nevertheless, their dataset was limited to just 12 distinctive KCs, which were also not aligned with any established curriculum. Shan et al. (2024) proposed a prompt-based approach where LLMs are guided to produce annotations in a key-value format, with mathematical terms as keys and their corresponding annotations as values. However, their approach focuses on tagging KCs to mathematical expressions rather than word problems and the respective solutions. Their approach is also not aligned with any established curriculum.

Despite all the advances in automated KC-tagging of math problems, there is still a lack of generalizable methods to annotate MWPs with curriculum-aligned KCs that do not require labelled training data. Our work aims to address this gap by evaluating unsupervised approaches for tagging MWPs with curriculum-aligned KCs.

3. Methodology

3.1 Data Construction

To study the effectiveness of automated KC annotation, we utilize two widely used MWP datasets: ASDiv and GSM8K, which are selected due to their complementary characteristics. ASDiv is a diverse dataset for MWP solving comprising 2,305 elementary-level problems collected from sources like textbooks and educational websites, and includes topics like arithmetic, algebra, and geometry. GSM8K comprises 8,500 arithmetic MWPs from elementary and middle school level that require multi-step reasoning. Topics include basic arithmetic up to rate problems and basic probability.

A critical challenge in mathematical learning systems lies in understanding the fundamental skills required to solve different MWPs. To address this, we engaged external annotators with expertise in grade to middle school level math education to manually annotate each problem in our selected ASDiv and GSM8K datasets with its corresponding knowledge component. The goal was to create a high-quality, human-annotated dataset that can serve as a ground truth for training and evaluating automated KC-tagging models.

The annotators were instructed to use a structured framework based on the Common Core State Standards (CCSS), a set of mathematics benchmarks widely adopted in U.S.

states, aligning each problem with fine-grained skill categories. Each problem was assigned one or more KCs, based on not only the problem text but also the solution as provided in the original dataset. For solutions with multiple expressions, each expression is taken as a step. As both ASDiv and GSM8K comprise only math word problems, we emphasize annotating with KCs that involve solving word problems over KCs that are only related to calculations. For multi-step problems like in GSM8K, annotators are instructed to consider both at the step-level as well as the problem-level before deciding on the appropriate annotation. Annotators were given guidance to select the fewest number of KCs and using the earliest grade KCs that best match the mathematical skill required to solve the problem. For example, a problem that requires two calculation steps with different operations to solve may involve addition/subtraction as a step (2.OA.A.1) and multiplication/division as a step (3.OA.A.3), but we can use a single KC that covers two-step word problems with all four operations (3.OA.D.8) to cover the problem.

To ensure consistency in the annotation among the annotators, we conduct several rounds of calibration where each problem was independently annotated by two annotators until an Inter Annotator Agreement (IAA) of at least 80% is achieved. The resulting dataset provides a rich resource for training and evaluating AI models to automatically infer knowledge components, enabling personalized learning, automated curriculum alignment, and improved feedback mechanisms in ITSs. By benchmarking LLM-based approaches on this dataset, we aim to assess whether AI models can effectively replace or augment human annotators in this critical task.

Table 1. Examples of KC-annotated problems from ASDiv

Problem	Solution	Answer	KC
There are 5 birds in a tree. How many bird legs do you see?	5*2=10	10 (bird legs)	3.OA.A.3
Dave bought a new flat screen TV. The screen was 2 feet wide and 4 feet tall. What is the area of the screen?	2*4=8	8 (square feet)	4.MD.A.3
The sum of three consecutive odd numbers is one hundred twenty-three. What is the smallest of the three numbers ?	x:The first number; x+(x+2)+(x+4) =123	39	6.EE.B.6

Table 2. Examples of KC-annotated problems from GSM8K

Problem	Solution	Step	KC
A chef bought 4 bags of onions. Each bag weighs 50 pounds. A pound of onions cost $1.50. How much did the chef spend?	A bag of onions cost $1.50 x 50 = <<1.5*50>>=75.Therefore, the chef spent $75 x 4 = <<75*4>> = 300 for the four bags of onions. #### 300	1.5*50=75	4.MD.A.2
		75*4=300	4.MD.A.2
John buys 2 packs of index cards for all his students. He has 6 classes and 30 students in each class. How many packs did he buy?	John has 6*30=<<6*30=180>>180 students. So he bought 180*2 = <<180*2=360>>360 packs #### 360	6*30=180	3.OA.D.8
		180*2=360	4.NBT.B.5

The final annotated ASDiv dataset comprises 2,000 problems randomly selected from the original ASDiv problem pool, with 49 unique KCs and 107 unique combinations of KCs, demonstrating the diversity of the dataset. On the other hand, the annotated GSM8K dataset comprises 1,426 randomly selected problems with 4,868 total steps, with the number of steps for the problems ranging from 1 to as high as 8, and an average of 3.36 steps/problem. There were a total of 29 unique KCs and 42 unique KC combinations. Figure 1 shows the breakdown of the grades and domains for the annotations of both datasets. While the grades for both datasets span from kindergarten (K) to Grade 8, the ASDiv has a better distribution of grades across the problems in contrast to GSM8K, which sees a majority of problems concentrated

around Grade 4. In terms of the KC domains, both datasets have majority of the KCs under Operations and Algebraic Thinking (OA) and Measurement and Data (MD). Other common domains include Numbers and Operations in Base Ten (NBT), Ratios and Proportional Distribution (RP), as well as Expressions and Equations (EE). Tables 1 and 2 illustrate some example problems from the two datasets, along with their solutions and the corresponding KC annotations.

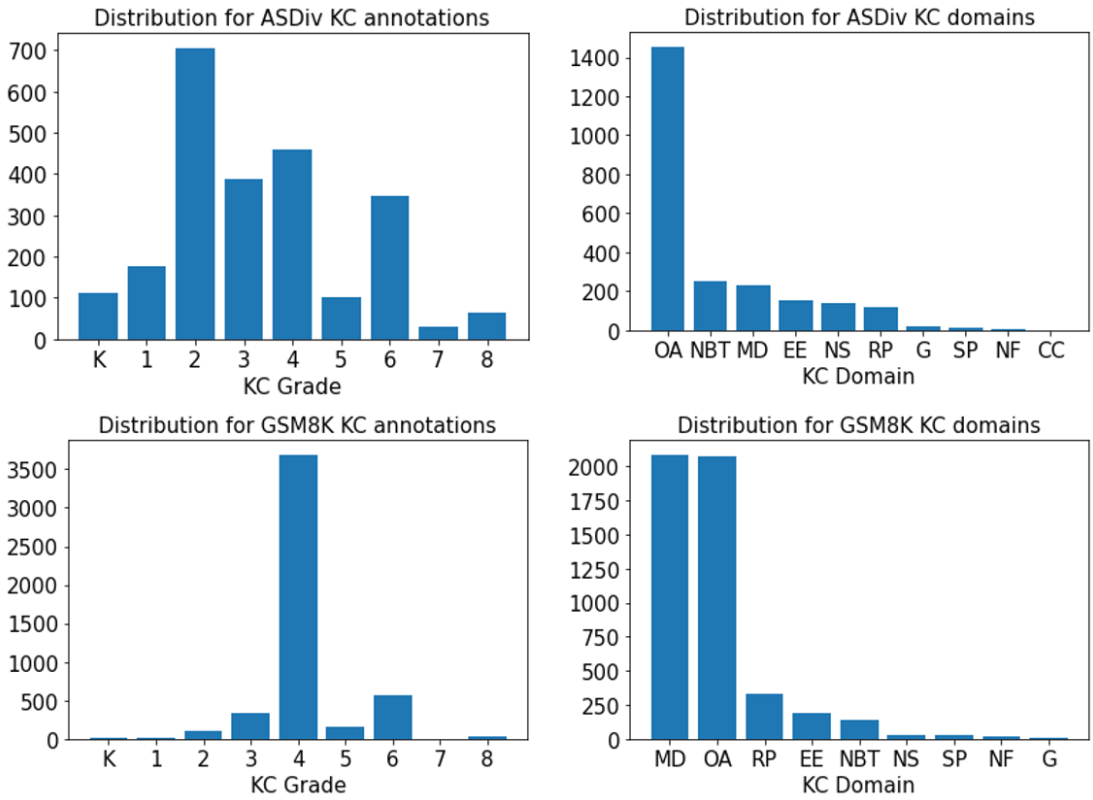


Figure 1. Distribution of grades and domains for ASDiv and GSM8K KC annotations

3.2 Models

3.2.1 – SBERT-based Sentence Embedding and Similarity Ranking

For the first approach, we adopt a retrieval-style strategy using sentence-level embeddings. We use a pre-trained Sentence-BERT (SBERT) model (all-MiniLM-L6-v2) to generate dense vector representations for each problem as well as each candidate KC description from CCSS. For each problem, we concatenate the problem text along with the solution and answer as such: "*Problem/Word problem: <problem text> Solution <solution> Answer: <answer>*". During our experiments with problems from ASDiv, we found that using "Word problem" performed much better than "Problem" as it biases the embedding towards word problem KCs. For GSM8K, we use "Word problem" exclusively for the problem embeddings. We also experimented with using only the problem text without the solution in the problem embedding to compare the performance.

For each problem, we compute the cosine similarity between its embedding and the embeddings of all available KC descriptions. The KC with the highest similarity score is selected as the predicted label. We also evaluate the results using recall@k, such that the top-k ranked KCs can be analyzed to evaluate near-miss performance. We also calculate the mean reciprocal rank (MRR), which captures how close the ground truth is to the top rank. This approach offers a lightweight and transparent baseline for semantic matching between instructional content and curricular standards without requiring any task-specific training.

3.2.2 – LLM-based Labelling with GPT-4o mini

We also leverage on a large language model, GPT-4o mini, to perform automated KC tagging. We prompt the model with the math word problem and the ground truth solution and ask it to select the most appropriate KC from the CCSS. Because there are over 200 KCs in the CCSS just from Grades 1 to 8 alone, we do not provide the model with the descriptions of all the possible KCs. Instead, we rely on the LLM's internal knowledge of the CCSS taxonomy to assign the most appropriate label. This method simulates how an LLM might be used in automated curriculum alignment or tutoring systems, where it must understand both the semantics of the math word problem and the pedagogical scope of various standards without additional input.

We experiment with four prompting strategies:

- **Zero-shot:** A basic prompt that instructs the model to output only the relevant CCSS KC
- **Few-shot**: The prompt includes a few example problems paired with gold KC labels before presenting the test example, the only output is the predicted KC
- **Chain-of-thought (CoT):** The model is prompted to explain its reasoning step-by-step before producing a label KC
- **Few-shot + CoT**: Combines examples with intermediate reasoning steps for a few example problems before prompting the model to reason step-by-step and produce a label KC

These variations allow us to explore the effects of guidance and reasoning on LLM performance in aligning curriculum standards with problem content. We use GPT-4o mini via the OpenAI API, with temperature set at 0 to ensure deterministic output. The outputs produced by GPT-4o mini are parsed using string matching and compared against the ground truth annotations to assess the accuracy. For the chain-of-thought prompts, we also consider the qualitative alignment with human reasoning.

3.2.3 – LLM + SBERT Embedding-based Matching

In addition to the direct SBERT similarity ranking method and the prompting-based LLM classification, we introduce a third hybrid method that leverages the generative capabilities of LLMs together with semantic similarity scoring. This approach is designed to mimic a more human-like reasoning pipeline: first summarizing the mathematical knowledge applied in the problem, then aligning it with an existing curriculum taxonomy.

In this method, we prompt the LLM (GPT-4o mini) with the full math problem text and corresponding solution steps, and instruct it to generate a natural language description of the knowledge components or mathematical skills involved in solving the problem. Once generated, each description is encoded using the SBERT model to produce a vector representation and the cosine similarities between this embedding and those of all the available CCSS KC descriptions are computed and ranked. This hybrid method thus separates the understanding of applied mathematical knowledge, handled by the LLM, from the alignment with standard curriculum concepts, handled by SBERT embeddings, offering a flexible alternative to direct classification or retrieval-based methods.

4. Results

4.1 SBERT sentence embedding results

The results for the SBERT-embedding method on the ASDiv dataset are summarized in Table 3. We see better recall@k across all values for k as well as better MRR when both problem text and the solution are used for the problem embedding. We also observe significant improvements across the board when we use "Word problem" as part of the input for the problem embedding. We observe a similar trend for the GSM8K dataset (Table 4). However,

the overall performance is lower, with a best recall@10 of 51.85% for GSM8K vs. 79.22% for ASDiv, and a MRR of 0.1592 vs. 0.3374. One possible reason for this is that the solution for ASDiv problems only includes mathematic expressions (see Table 1) while for GSM8K there are much longer explanations in natural language along with the mathematical expressions (see Table 2), which dilute the core mathematical concepts applied. As a result, the embedding becomes an averaged representation, and the truly discriminative parts end up being underweighted, adversely affecting the similarity scoring.

Table 3. *SBERT sentence embedding similarity scoring results for ASDiv*

ASDiv	recall@1	recall@3	recall@5	recall@10	MRR
Problem only ("Problem")	10.39%	25.92%	37.91%	58.79%	0.2419
Problem only ("Word problem")	13.19%	33.27%	49.85%	74.08%	0.2964
Problem + solution ("Problem")	16.08%	37.41%	51.20%	68.33%	0.3196
Problem + Solution ("Word problem")	15.53%	40.01%	59.09%	79.22%	0.3374

Table 4. *SBERT sentence embedding similarity scoring results for GSM8K*

GSM8K	recall@1	recall@3	recall@5	recall@10	MRR
Problem only ("Word Problem")	3.04%	8.63%	16.52%	46.24%	0.1395
Problem + Solution ("Word problem")	3.51%	11.50%	24.03%	51.85%	0.1592

4.2 LLM Annotation Results with GPT-4o mini

The results for GPT-4o mini tagging are shown in Tables 5 and 6 for ASDiv and GSM8K problems respectively. For ASDiv, we observe an improvement in accuracy from zero-shot prompting (13.39%) to chain-of-thought (15.68%) as well as few-shot prompting (22.13%). The best performance is achieved by combining few-shot and chain-of-thought prompting, with an accuracy of 27.57%. This compares very favorably with the best recall@1 score for the SBERT-embedding method which achieved 16.08%, and is likely due to the LLM's ability to perform contextual reasoning. One common downside that we observe in LLMs is the tendency to hallucinate, which in this case is usually when it generates KC annotations that do not exist in the CCSS. The rate of hallucination is highest for the zero-shot setting at 6.5%, and is greatly reduced in the few-shot setting at 0.7%. The hallucination rate is moderate at 2.7% and 2.2% for the chain-of-thought and few-shot + chain-of-thought settings respectively.

Table 5. *ASDiv LLM Results with GPT-4o mini*

	Zero-shot	Few-shot	CoT	Few-shot CoT
Accuracy	13.39%	22.13%	15.68%	27.57%
Hallucinations	6.5%	0.7%	2.7%	2.2%

Table 6. *GSM8K LLM Results with GPT-4o mini*

	Zero-shot	Few-shot	CoT	Few-shot CoT
Accuracy	53.51%	54.81%	46.77%	52.59%
Hallucinations	0.2%	0.2%	0.4%	0.2%

In contrast to the SBERT-embedding approach, we observe better results for GSM8K using LLM-tagging compared to ASDiv. The accuracy for few-shot prompting performed the best at 54.81% compared to 53.51% for zero-shot prompting. Unlike for ASDiv, the LLM-

tagging performance is worse with chain-of-thought reasoning, with accuracy of 46.77% for chain-of-thought and 52.59% for few-shot + chain-of-thought. Upon examining the chain-of-thought output provided by the LLM, we observed that the reasoning is usually sound, but the final decision is a similar but incorrect KC. For example, a common error is between 4.OA.A.3 and 4.MD.A.2, which are both related to solving word problems with the four operations, with the former emphasizing multistep problem solving while the latter involving measurement quantities like distances, intervals of time, liquid volumes, and money. However, we note that the LLM accuracy is high for the GSM8K problems compared to ASDiv even for the worse-performing chain-of-thought prompting. The rate of hallucination by the LLM is generally low for GSM8K, ranging from 0.2% to 0.4%.

4.3 LLM KC generation + SBERT sentence embedding

The results for the combined LLM + SBERT-embedding approach are shown in Tables 7 and 8 along with the best results from the other two approaches. The accuracy/recall@1 for ASDiv dataset is much improved from 15.53% using only sentence-embedding similarity scoring and from 27.6% using LLM few-shot + CoT to 39.36% with the hybrid approach. We also see improvements across the board for higher values of k, and the overall MRR achieves 0.5352, meaning that the correct KC is typically within the top 2 ranked KCs on average.

For the GSM8K dataset, the hybrid approach significantly improves the recall@k and MRR over the SBERT-only approach, with recall@1 improving from 3.51% to 16.17%, and MRR improving from 0.1592 to 0.3820. However, the hybrid approach still underperforms the best LLM annotation result, which achieves a recall@1 of 54.81% with few-shot prompting. Upon analyzing the model outputs, we believe this is due to many CCSS KCs having highly similar descriptions, differing only in subtle differences. For example, the description for 2.OA.A.1 and 2.MD.B.5 are very similar in that both are related to solving word problems using addition and subtraction within 100, with the main differences being that 2.OA.A.1 emphasizes that the solution can be up to two steps, while 2.MD.B.5 does not specify the step number but instead focuses on problems involving length with consistent units. When prompted to describe the knowledge components applied in the problem, the LLM may not go into sufficient detail to capture these fine distinctions between similar KCs. As a result, the hybrid approach can retrieve closely related KCs, and perform significantly better than the SBERT-only approach, but struggles to precisely identify the correct one among several similar candidates.

Table 7. *ASDiv results comparison*

ASDiv	recall@1	recall@3	recall@5	recall@10	MRR
LLM Few-shot + CoT	27.6%	-	-	-	-
SBERT Problem + Solution (“Word problem”)	15.53%	40.01%	59.09%	79.22%	0.3374
LLM + SBERT	39.36%	58.74%	72.78%	87.31%	0.5352

Table 8. *GSM8K results comparison*

GSM8K	recall@1	recall@3	recall@5	recall@10	MRR
LLM Few-shot	54.81%	-	-	-	-
SBERT Problem + Solution (“Word problem”)	3.51%	11.50%	24.03%	51.85%	0.1592
LLM + SBERT	16.17%	53.23%	69.78%	82.00%	0.3820

Note that for both datasets, the hybrid approach shows above 50% recall@3 and consistently high performance of recall@k (k=5,10). This is important for manual validation, as the automatic annotations by the hybrid approach provide reliable inputs for manual validation, unlike the two methods of SBERT and LLM annotation. Furthermore, the hybrid

method achieves above 80% recall@10, similar to IAA of human annotators. This would indicate that top-10 results of the hybrid method can be used for manual validation by a single human annotator effectively.

4.4 Discussion

From the results, we observe that while SBERT sentence embeddings can serve as a useful tool for narrowing down candidate KCs for MWPs, their annotation accuracy remains relatively low compared to that of GPT-4o mini, especially for the GSM8K dataset. One likely reason for this performance gap for GSM8K is the difference in solution formats. GSM8K solutions include lengthy, natural language explanations (see Table 2) compared to the more concise, math expressions-based solution format for ASDiv (see Table 1). Preprocessing of the text solutions to extract only the mathematic steps before computing sentence embeddings may yield better results for GSM8K. Additionally, GSM8K emphasizes multistep problems, which might not be fully captured in the embeddings of the problem text and solution. This makes it challenging for similarity-based methods to accurately distinguish between KCs with subtle differences and consistently align problems with the most appropriate KC.

When annotating MWPs using GPT-4o mini, we found that it shows high degree of inconsistency across the two datasets and that the chain-of-thought prompting has opposite effects between the datasets, while few-shot prompting leads to performance improvements for both datasets. The few-shot prompting helps to guide the model towards producing responses that align more closely with the desired output format. Since MWP annotation is inherently subjective and context-dependent, it is crucial to craft clear, targeted system prompts and carefully selected few-shot examples when using LLMs for annotation. These prompt strategies help steer the model towards consistent and curriculum-aligned outputs, thereby improving annotation quality and reliability.

The hybrid LLM+SBERT approach demonstrates promising performance, particularly on the ASDiv dataset. By combining the LLM's ability to abstract and articulate the concepts involved in a MWP with SBERT's semantic similarity matching, this method outperforms both the SBERT-only and LLM-only approaches on ASDiv. This is likely due to ASDiv's well-structured problems and concise solution formats, which make it easier for the LLM to generate focused descriptions and for the similarity scoring to correctly identify the matching KC. In contrast, on GSM8K, the hybrid method underperforms the LLM-only approach. The longer, multi-step reasoning required in GSM8K appears to make the LLM's generated KC descriptions either too verbose or too vague, resulting in less precise matches during similarity scoring.

5. Conclusion

Our experiments demonstrate that automated KC tagging for math word problems by combining SBERT embeddings and LLMs can produce promising results, offering valuable support in reducing manual effort and increasing scalability. Importantly, our findings suggest that further refinement—such as improved prompt design and more carefully curated few-shot examples—has the potential to significantly boost performance. With continued progress, these methods could eventually approach human-level reliability, enabling more efficient and consistent KC annotation at scale.

Acknowledgements

This research is supported by the Ministry of Education, Singapore, under its Science of Learning Grant (award ID: MOE-MOESOL2021-0006). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Education, Singapore.

References

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. The journal of the learning sciences, 4(2), 167-207.

Cetintas, S., Si, L., Xin, Y. P., Zhang, D., & Park, J. Y. (2009). Automatic Text Categorization of Mathematical Word Problems. In FLAIRS.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., Schulman, J. (2021). Training verifiers to solve math word problems. arXiv:2110.14168.

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. User modeling and user-adapted interaction, 4, 253-278.

Karlovčec, M., Córdova-Sánchez, M., & Pardos, Z. A. (2012). Knowledge component suggestion for untagged content in an intelligent tutoring system. In Proceedings of 11th International Conference on Intelligent Tutoring Systems, pp. 195-200.

Li, H., Xu, T., Tang, J., & Wen, Q. (2024). Automate knowledge concept tagging on math questions with LLMs. arXiv:2403.17281.

Miao, S. Y., Liang, C. C., & Su, K. Y. (2020). A diverse corpus for evaluating and developing English math word problem solvers. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 975–984.

Moore, S., Schmucker, R., Mitchell, T., & Stamper, J. (2024). Automated generation and tagging of knowledge components from multiple-choice questions. In Proceedings of the eleventh ACM conference on learning@ scale, pp. 122-133.

OpenAI. (2024). GPT-4o Mini: Advancing cost-efficient intelligence. OpenAI. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

Pane, J. F., Griffin, B. A., McCaffrey, D. F., & Karam, R. (2014). Effectiveness of cognitive tutor algebra I at scale. Educational Evaluation and Policy Analysis, 36(2), 127-144.

Pardos, Z. A., & Dadu, A. (2017). Imputing KCs with representations of problem content and context. In Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization, pp. 148-155.

Patikorn, T., Deisadze, D., Grande, L., Yu, Z., & Heffernan, N. (2019). Generalizability of methods for imputing mathematical skills needed to solve problems from texts. In Proceedings of 20th International Conference on Artificial Intelligence in Education (AIED), pp. 396-405.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3982–3992.

Shan, R., & Youssef, A. (2024). Using large language models to automate annotation and part-of-math tagging of math equations. In International Conference on Intelligent Computer Mathematics, pp. 3-20.

Shen, J. T., Yamashita, M., Prihar, E., Heffernan, N., Wu, X., McGrew, S., & Lee, D. (2021). Classifying math knowledge components via task-adaptive pre-trained BERT. In Proceedings of 22nd International Conference on Artificial Intelligence in Education (AIED), pp. 408-419.

Tan, C. S., & Kim, J. J. (2024). Automated Math Word Problem Knowledge Component Labeling and Recommendation. In International Conference in Methodologies and Intelligent Systems for Technology Enhanced Learning, pp. 338-348.

VanLehn, K. (2006). The behavior of tutoring systems. International journal of artificial intelligence in education, 16(3), 227-265.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Verschaffel, L., Schukajlow, S., Star, J., & Van Dooren, W. (2020). Word problems in mathematics education: A survey. ZDM Mathematics Education, 52, 1-16.

Yamauchi, T., Flanagan, B., Nakamoto, R., Dai, Y., Takami, K., & Ogata, H. (2023). Automated labeling of PDF mathematical exercises with word N-grams VSM classification. Smart Learning Environments, 10(1), 51.

Identifying Cognitive and Metacognitive Processes Using Retrospective and Concurrent Think-Aloud Protocols

Rumana PATHAN*, Sahana MURTHY & Ramkumar RAJENDRAN
Centre for Educational Technology, Indian Institute of Technology Bombay, India
*rumana_pathan@iitb.ac.in

Abstract: Think-aloud protocols such as Retrospective Think-Aloud (RTA) and Concurrent Think-Aloud (CTA) are widely used in educational research to examine cognitive and metacognitive processes. RTA captures reflective verbalizations after task completion, often yielding elaborate explanations, while CTA records real-time thoughts, offering direct insights into learners' in-task thinking. Although both protocols uncover complex learning processes, CTA is more commonly used to model metacognition. This study compares RTA and CTA in identifying cognitive and metacognitive processes during problem-solving in MEttLE (Modelling-based Estimation Learning Environment), a computer-based, open-ended learning environment for complex, estimation-based problem solving. We conducted parallel studies using both protocols as learners engaged in estimation problem-solving tasks. The analysis focused on: (1) the overall frequency of verbalizations, (2) the frequency of cognitive and metacognitive processes, and (3) a temporal analysis of metacognitive activity using CTA. Findings show that CTA yielded more on-task verbalizations than RTA. RTA captured high-level control and reflective processes such as model building, estimation reasoning, and adaptive attributions. In contrast, CTA revealed procedural strategies like selecting and adapting learning approaches and gathering context-specific knowledge. CTA also enabled temporal tracking of metacognitive shifts throughout the task. These findings highlight the complementary strengths of both protocols and offer methodological insights for researchers seeking to analyze learners' thinking in technology-enhanced learning environments.

Keywords: Retrospective think-aloud; Concurrent think-aloud; Metacognitive processes; Self-regulated Learning

1. Introduction

Novice learners in computer-based environments engage in cognitive and metacognitive processes such as activating prior knowledge, setting objectives, using strategies, monitoring comprehension, and responding to system-generated prompts to solve complex problems (Azevedo & Wiedbusch, 2023). However, research indicates they often struggle to regulate and monitor learning effectively (Zimmerman, 2013). Identifying these processes is crucial to supporting learners. Among various methods such as human observation, discourse analysis, self-reports, trace data, and neuroimaging, think-aloud protocols remain a preferred approach due to their ability to capture learners' thought processes in real time, offering a rich and nuanced understanding (Vaccaro & Fleming, 2018).

Think-aloud is a simple yet powerful method to reveal complex cognitive processes (Greene et al., 2017; Ericsson, 2006). It captures real-time verbalizations, enabling researchers to interpret learners' cognitive activity as it unfolds. Studies using think-aloud have uncovered processes used by experts like chess masters and doctors (Ericsson, 2006), as well as usability challenges in system design (Alhadreti & Mayhew, 2018). Arguably, it provides the closest representation of spontaneous thought (Greene et al., 2017). Two main think-aloud approaches exist: retrospective (RTA), in which learners verbalize after the task with cues like video playback, and concurrent (CTA), where learners speak within seconds of

the thought occurring during the task (Ericsson, 2006). RTA reduces cognitive load during performance but may result in incomplete recall (Birns et al., 2002), while CTA provides real-time data but may strain learners (Greene et al., 2017). Both approaches are widely used in studies on cognitive and metacognitive processes.

The motivation for this research stems from the need to support novice learners who often face challenges regulating cognition during complex, open-ended problem-solving tasks in computer-based environments. While think-aloud protocols are common in metacognitive research, most studies have focused on CTA in simpler tasks like reading, with limited work exploring RTA or applying these methods in complex, dynamic contexts. A critical gap remains in understanding how RTA and CTA differ in capturing the richness, frequency, and temporal dynamics of metacognitive activity (Hertzum, 2024). This study addresses the gap by comparing RTA and CTA in MEttLE (Modelling-based Estimation Learning Environment), a complex, open-ended learning environment in electrical engineering. Using RTA (N=10) and CTA (N=4), we analyzed learner verbalizations coded via Pintrich's SRL framework and MEttLE's pedagogical design across four phases and 17 processes. We compared protocols based on verbalization frequency and the distribution of metacognitive processes. Additionally, we used CTA to demonstrate the temporal flow of metacognitive processes.

2. Literature Review

Metacognition is defined as "knowledge concerning one's own cognitive processes, products, and anything related to them" (Flavell, 1976). It is broadly categorized into two components: knowledge of cognition, i.e., awareness of strategies and cognitive processes, and regulation of cognition, which involves planning, monitoring, and control (Efklides, 2011). While its value in complex problem-solving is well-established, limited research explores how learners actively regulate cognition during such tasks.

The regulation of cognition is closely associated with Self-Regulated Learning (SRL), a framework that encapsulates how learners plan, monitor, and reflect on their learning. SRL is influenced by cognitive, motivational, affective, and metacognitive processes (Panadero, 2017). Several SRL models (e.g., Pintrich, 2000,) detail these processes to varying degrees. Within these frameworks, regulation of cognition is central to achieving learning goals, particularly in complex and open-ended problem-solving scenarios (Krieger et al., 2022).

To understand SRL processes, researchers have used various methods including questionnaires, interviews, observations, eye-tracking, log data, and think-aloud protocols (Greene et al., 2017). Among these, think-aloud protocols are the most widely used, offering real-time access to cognitive and metacognitive processes (Fan et al., 2022). These are typically implemented in two forms: retrospective think-aloud (RTA), where learners reflect post-task, often using screen recordings; and concurrent think-aloud (CTA), which captures verbalizations during task performance (Ericsson & Simon, 1993). Both methods can be analyzed using SRL frameworks, such as Pintrich's (2000), and adaptations suited to digital learning environments (Greene & Azevedo, 2010; Pathan, Murthy, & Rajendran, 2021).

Numerous studies have applied think-aloud protocols to explore SRL. CTA has been used in digital reading (Coiro et al., 2014) and biology learning (Greene & Azevedo, 2010), while Trevors et al. (2014) combined RTA and CTA to identify belief structures. Each protocol has strengths and limitations. CTA provides immediate, unfiltered access to thought processes but may miss unconscious reasoning and impose cognitive load (Jääskeläinen, 2010). RTA offers richer reflection post-task but risks omissions and rationalizations (Birns et al., 2002). To maximize the use of short-term working memory during RTA, use of techniques such as video-stimulated recalls are suggested (Alhadreti & Mayhew, 2018).

To summarize, although both RTA and CTA are valid methods for identifying metacognitive processes, CTA is more commonly used in studies. Each method has context-specific strengths depending on task type and research objective. Most CTA-based research focuses on reading and comprehension tasks, with limited work exploring their use in complex problem-solving and the temporal dynamics of metacognition. Understanding these processes is especially important in open-ended, complex problem-solving tasks where learners need to plan, monitor, and adjust their thinking. This study addresses that gap by comparing RTA and

CTA in an open-ended learning environment (OELE) setting, aiming to inform future research design and protocol selection. The findings will benefit researchers by helping them choose the most appropriate verbal protocol method based on their study goals and task contexts.

3. Research Method

This comparative study investigates how RTA and CTA protocols differ in identifying cognitive and metacognitive processes in learners using a complex problem-solving OELE.

3.1 Participants and Setting

Fifteen undergraduate engineering students (ages 20–21) from a private institute in a tier-2 city in India participated in this study. Seven were from mechanical engineering and eight from electronics and telecommunication, with a gender distribution of five females and ten males. Five students (1 female, 4 males) were assigned to the concurrent think-aloud (CTA) group, and ten (4 females, 6 males) to the retrospective think-aloud (RTA) group; however, one male CTA participant's data was excluded due to technical issues. All participants provided informed consent and received cash compensation for their participation in the study. The study was reviewed and approved by the Institute Ethics Committee at IIT Bombay, India.

Participants interacted with MEttLE on desktop computers with 23-inch screens. MEttLE was locally hosted on a machine via port 8080, with learners accessing it through a static IP over LAN. Interaction data including timestamp, object ID, log data, page ID, and session ID was recorded on a local MongoDB server (Pathan, Shaikh & Rajendran, 2019). Verbal and screen data were captured using OBS Studio for both CTA and RTA sessions. Additionally, handheld recorders were used during RTA interviews to capture audio.

3.2 Learning Environment and Procedure

MEttLE (Modelling-based Estimation Learning Environment) is an OELE designed to support novice learners in estimation-based problem-solving tasks (Kothiyal & Murthy, 2010). Learners solve problems such as estimating the motor power of a racing car using parameters like wheel diameter, distance, and weight. The system scaffolds both cognitive and metacognitive processes through structured tasks, expert guidance, and reflective prompts.

Problem-solving in MEttLE involves five interconnected tasks: functional, qualitative, and quantitative modelling, followed by calculation and evaluation. Learners define system operations, identify relationships among entities, develop equations, make assumptions, perform calculations, and evaluate their solutions. Being open-ended, MEttLE allows flexible sequencing and iterative exploration. Key tools include a simulator, calculator, info center, scribble pad, causal map builder, equation builder, and problem map, along with hints and prompts to support strategic thinking.

In the RTA protocol, learners engaged in a 60–90 minute problem-solving session, after which they participated in a stimulated recall interview. Using their session recordings, they were asked open-ended questions such as "What were you thinking while watching the tutorial video?" to prompt reflection on their cognitive processes. In the CTA protocol, learners first practiced thinking aloud with a warm-up task (e.g., long multiplication) before solving the same estimation problem in MEttLE while verbalizing their thoughts in real time. Researcher interaction was minimized, limited to neutral prompts like "keep talking." The think-aloud verbalizations and interaction with MEttLE were recorded in a video format and transcribed.

4. Coding and Data Analysis

4.1 Coding of Think Aloud Protocols

The transcribed retrospective and concurrent think-aloud protocols were analyzed using a coding mechanism that links learner verbalizations to observable indicators, each

corresponding to a specific metacognitive process and phase (Pathan, Singh, Murthy & Rajendran, 2022). This scheme draws on Pintrich's (2000) SRL framework and the pedagogical design of MEttLE (Kothiyal & Murthy, 2018). It comprises four phases - Planning & Activation, Monitoring, Control & Regulation, and Reaction & Reflection, encompassing 17 metacognitive processes (e.g., target goal setting, feeling of knowing), and their indicators. Each phase represents a distinct dimension of metacognitive regulation. Phase 1 (Planning & activation) focuses on goal setting and activating prior knowledge; Phase 2 (Monitoring) captures metacognitive awareness such as the feeling of knowing; Phase 3 (Control & regulation) involves selecting and adapting cognitive strategies; and Phase 4 (Reaction and reflection) includes learners' evaluations and reflections on their task performance.

For example, when a learner articulates a plan or goal such as deciding to complete a task first, it is coded under the indicator "Learner may begin a task by setting specific goals for learning," which falls under the Target Goal Setting process within the Planning & Activation phase. Similarly, if a learner expresses uncertainty but believes they know the answer (e.g., *"I can't remember the exact value, but I know it was discussed"*), this maps to the indicator "Learner cannot recall something... but has strong feelings they know it," representing the Feeling of Knowing process in the Monitoring phase. Another example is when a learner uses the scribble pad to visually represent a model, which corresponds to the indicator "Learner uses scribble pad" under the External Representation process in the Control & Regulation phase. These examples illustrate how specific learner behaviors are systematically categorized to reveal the metacognitive processes and phases involved in problem-solving.

4.2 Data Coding and Analysis

Table 1 presents example excerpts from both RTA and CTA protocols alongside their corresponding codes. One RTA example, *"I selected mass and increased the parameters with velocity, and then I compared where it is varying and found the maximum change"*—was interpreted using the screen recording, which showed the learner interacting with the simulator. Here, the learner held the mass constant and varied the velocity to observe its effect on power, thereby employing a one-variable-at-a-time strategy to explore system dynamics. This behaviour was mapped to the process of 'model building' under the phase 'control and regulation'. Similarly, a CTA example, *"I do not understand this. Does it mean I can create a causal map of my own?"* was observed after the learner read a prompt related to causal map creation within the qualitative model-building subgoal. This statement reflects metacognitive monitoring, as the learner questions their understanding, aligning with the process of 'judgement of learning and comprehension monitoring' under the 'monitoring' phase.

Table 1. *Example excerpts from RTA and CTA protocol, the context obtained from the screen recordings, and the codes.*

	Excerpt	Context	Indicator	Process	Phase
RTA	*Okay, now I will have to find power*	Read problem statement	Learner begins a task by setting a specific goal.	Target goal setting	Planning & activation
	The idea was not so clear; given a question, I could not get the exact idea about the approach.	Problem map	Learner becomes aware that he does not understand something	Judgement of learning	Monitoring
	I selected mass and increased the parameters with velocity, and then I compared where it is varying and then found the maximum change	Simulator	Learner uses variable manipulation, with implicit guidance to incorporate problem context in the estimation process	Model building	Control & regulation

	The question forced me to think whether the option we selected was correct or not.	Qualitative model evaluation	Learner uses question prompts to reflect on (problem-solving) process	Productive reflection on process	Reaction and reflection.
CTA	*How should I begin? Okay, let me calculate the maximum value first.*	Functional model	Learner begins a task by setting a specific goal.	Target goal setting	Planning & activation
	I do not understand this. Does it mean I can create a causal map of my own?	Causal map builder	Learner becomes aware that he does not understand something	Judgement of learning	Monitoring
	I think Pramod[1] is wrong because certainly, for a 10 kg toy car, it requires 10W, but we do not know the specifications.	Evaluation	Learners use question prompts to do estimation reasoning	Estimation reasoning	Control & regulation
	I think for now I am going in the right direction	Qualitative model evaluation	Learner evaluates his performance.	Cognitive judgement	Reaction & reflection.

5. Results

We coded 1698 RTA (N=10) and 2011 CTA (N=4) verbalizations, averaging ~170 and ~503 per learner, respectively. Analyses focused on (1) frequency of verbalizations and phases, (2) frequency of key processes (excluding those with average <2), and (3) temporal patterns of phases in CTA.

5.1 Frequency of Verbalizations and Phases

The average frequency of verbalizations was higher in CTA (503) compared to RTA (170). In RTA, 57% of the verbalizations (approximately 97 per learner) were on-task and mapped to one of the four metacognitive phases, while the remaining were not on-task. In contrast, 92% of CTA verbalizations (about 463 per learner) were on-task. These findings indicate that CTA not only yields a greater volume of verbal data but also a higher proportion of relevant, on-task verbalizations compared to RTA, aligning with existing literature discussed earlier.

Figure 1 illustrates the average frequency of phases (namely, planning & activation, monitoring, control & regulation, and reaction & reflection) per learner in both RTA and CTA protocol analyses, along with the percentage distribution of each phase in the two protocols. For instance, in the RTA protocol, learners deployed the planning and activation phase approximately 30 times, and this phase accounted for 30.6% of all phases. We found that, the phase of 'control & regulation' was the most frequently used phase in both CTA (78.8%) and RTA protocol (31.3%) analyses, surpassing the other phases in terms of frequency. However, both RTA and CTA exhibit a nearly identical occurrence of verbalizations in the planning & activation, monitoring, and reaction & reflection phases.

[1] Pramod is the name of a fictitious person used in a question prompt, specifically designed to help learners compare and evaluate their estimated value of power.

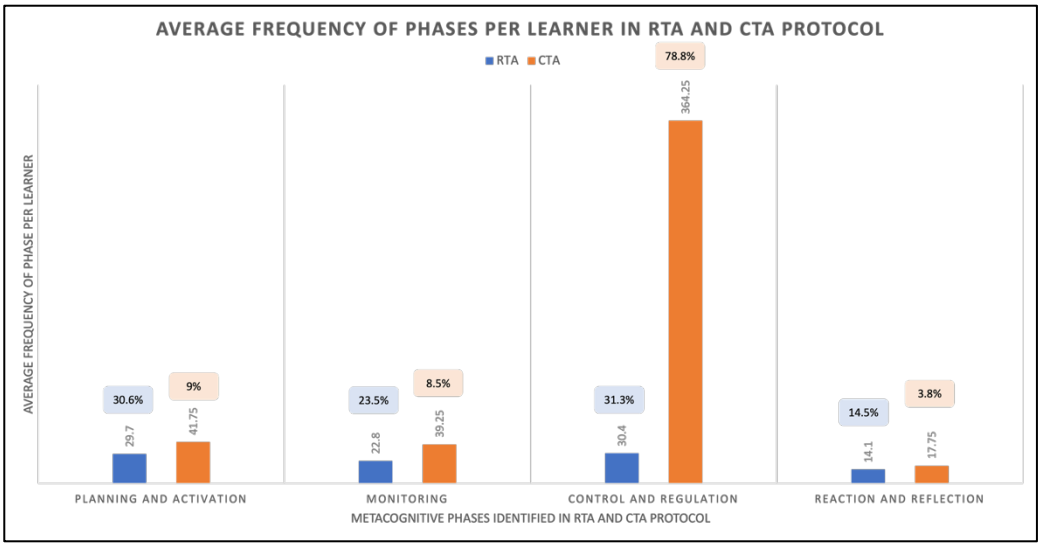


Figure 1. The average frequency of phases per learner in RTA and CTA protocol analysis.

5.2 Frequency of Processes

Due to the disparity in frequencies between the control and regulation phase and the other phases, we conducted a frequency analysis of the processes and divided the results into two subsections. 1) the difference between the average frequency of processes in the 'control & regulation' phase and 2) the difference between the average frequency of the processes in the 'planning & activation', 'monitoring', and 'reaction & reflection' phases.

5.2.1 Processes in the phase control and regulation

Figure 2 gives us a deeper insight into the type of specific processes represented in Control & Regulation phase in the two protocols. The average frequency of all processes per learner is higher in the CTA protocol than in the RTA protocol. However, the relative percentage distribution of some processes is higher in the RTA protocol. For example, learners in RTA verbalize statements that indicate problem-solving processes such as 'model building' (52.7%) and 'estimation reasoning' (28.4%) more frequently than other processes. Conversely, CTA verbalizations indicate a comparatively higher frequency of procedural processes such as 'selection and adaptation of control strategies' (30.5%) and 'gather context-specific knowledge' (44.7%) as compared to 'model building' (16.3%) and 'estimation reasoning' (8.4%).

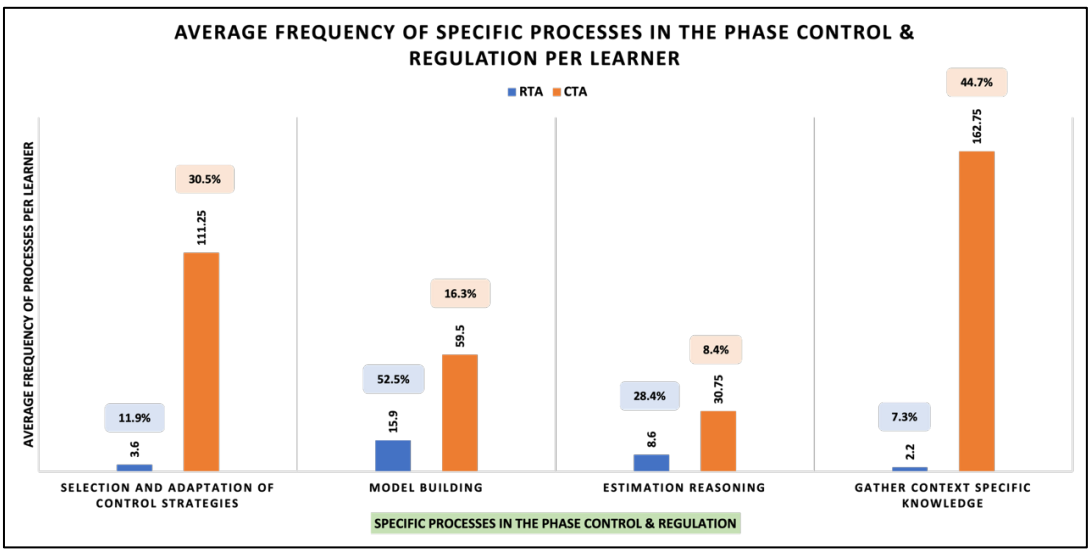


Figure 2. The average frequency of processes in the phase control and regulation in RTA and CTA protocol analysis

The procedural processes of selecting and adapting control strategies and gathering context-specific knowledge showed distinct patterns across the two protocols. In CTA, commonly observed indicators for strategy adaptation included re-reading, paraphrasing, and summarizing text—actions primarily aimed at enhancing comprehension and retention. Additionally, learners employed specific problem-solving strategies such as sequencing subgoals, identifying relevant information sources, and interpreting visual inputs like animations and graphs. In contrast, RTA captured only these problem-solving strategies, without the broader range of comprehension-related behaviors seen in CTA. For the process of gathering context-specific knowledge, CTA learners frequently accessed information embedded in prompts and hints, while RTA learners primarily referred to the infocenter.

The problem-solving processes of model building and estimation reasoning exhibited similar indicators in both RTA and CTA protocols. In both cases, learners engaged in model building by manipulating variables in simulations, using fictive motion language, constructing causal maps, and arranging relevant parameters and equations. Likewise, estimation reasoning was evident as learners in both protocols verbalized their use of question prompts and hints to support their estimations.

5.2.2 Processes in the phase – planning & activation, monitoring, and reaction & reflection

Figure 3 shows the average frequency of all processes within the phases of (a) planning & activation, (b) monitoring, and (c) reaction & reflection. Across these phases, most processes appeared more frequently in CTA than in RTA, with the exception of 'adaptive attributions', which occurred about five times per learner in RTA but was absent in CTA. Conversely, processes such as 'target goal setting', 'activation of relevant prior content knowledge', 'judgement of learning and comprehension monitoring', and 'cognitive judgements' were slightly more frequent in the CTA protocol.

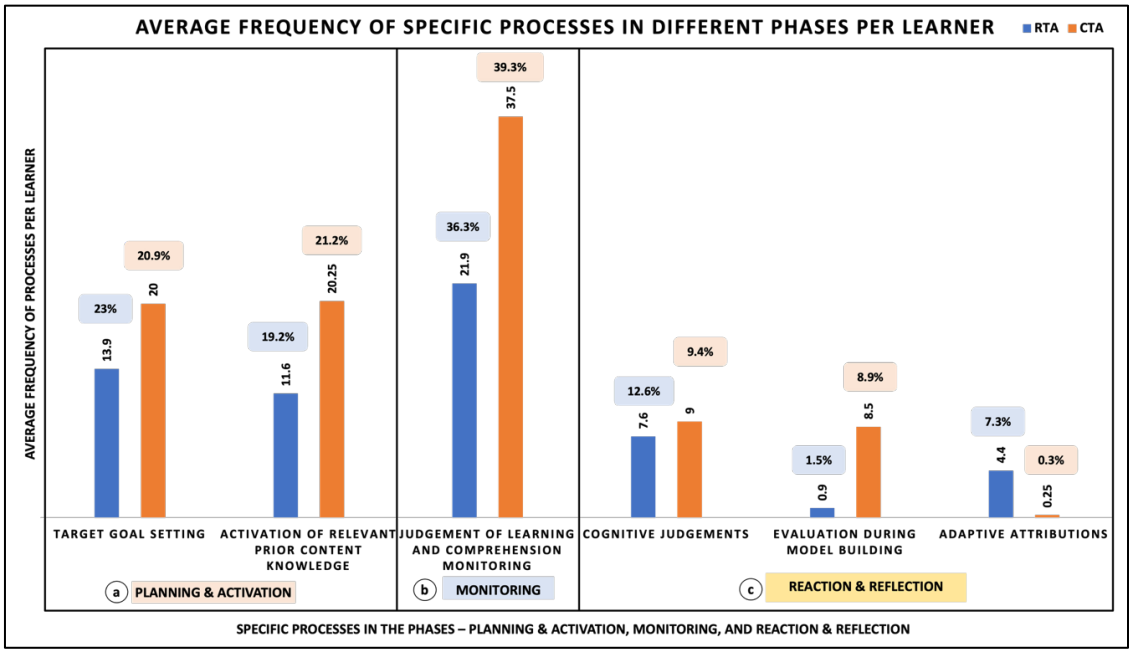


Figure 3. The average frequency of processes in the phases a) planning & activation, b) monitoring, and c) reaction & reflection phase in RTA and CTA protocol analysis

In the Planning & Activation phase, learners in both RTA and CTA protocols demonstrated the process of target goal setting by articulating specific goals related to learning, time management, and performance expectations. These goals were not fixed; learners revised them as they progressed through the task and expressed their evolving intentions using planning-related prompts. Both groups also engaged in activating relevant prior content knowledge, but they did so differently. RTA learners typically accessed prior knowledge by recalling information from memory. In contrast, CTA learners adopted a more

deliberate approach, actively using prompts and self-questioning strategies to refine their understanding and construct a more accurate representation of the problem.

Within the Monitoring phase, the process of judgement of learning and comprehension monitoring was evident in both protocols. Learners identified moments of clarity or confusion in their understanding, sometimes questioning the task (e.g., *"If I increase the acceleration at this point, what will be the velocity?"*) to check comprehension. In both RTA and CTA, learners used verbal statements triggered by question prompts to monitor and evaluate their ongoing understanding of the task.

In the Reaction & Reflection phase, learners engaged in cognitive judgments by evaluating their own performance. The process of evaluation during model building was observed when learners reflected on their models using evaluation and contextualization prompts. While both protocols captured this behaviour, it appeared more frequently in CTA. In contrast, adaptive attributions—where learners attributed success or failure to internal factors (e.g., effort) or external influences (e.g., poor strategies)—were more commonly verbalized in RTA, suggesting a reflective, post-task orientation more suited to retrospective reporting.

5.3 Temporal Patterns of Phases in CTA.

One significant methodological advantage of the CTA protocol over RTA is its ability to capture the temporal nature of metacognitive activity. Because CTA involves immediate verbalization of thoughts during task performance, it allows researchers to track the timing and sequence of metacognitive processes—something not possible with RTA. Figure 4 illustrates the average frequency of four metacognitive phases—planning & activation, monitoring, control & regulation, and reaction & reflection—over a 120-minute problem-solving session in an OELE. To analyze trends over time, the session was divided into 12 intervals of 10 minutes each. For example, in the first interval (0–10 minutes), learners engaged in planning & activation 11 times out of 170 total metacognitive events (6.47%) and in control & regulation 132 times (77.65%). Overall, learners used control & regulation strategies most frequently—about three times per minute—while planning & activation and monitoring occurred roughly once every three minutes, and reaction & reflection once every seven minutes.

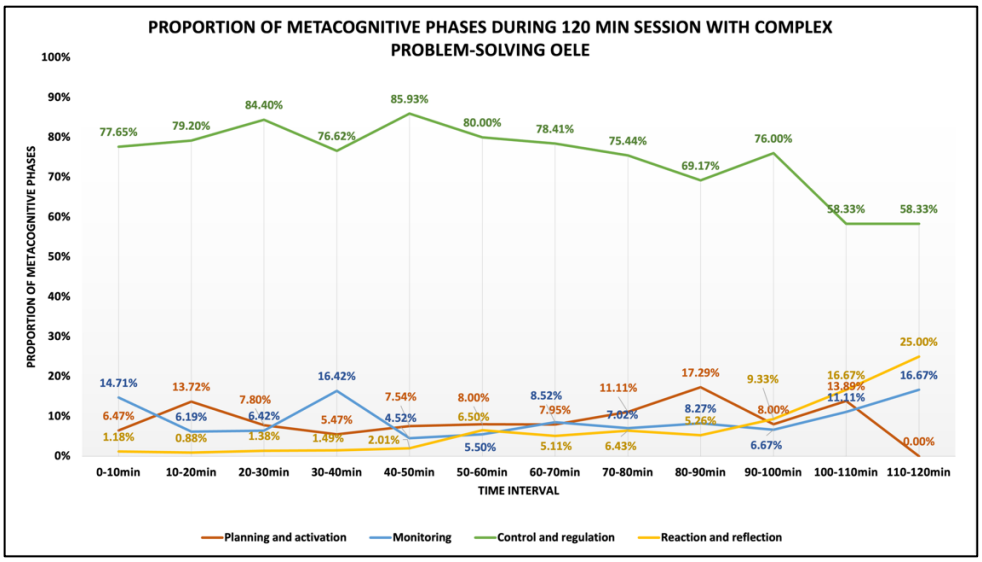


Figure 4. The proportion of metacognitive phases every 10 minutes in CTA protocol analysis

The temporal distribution of phases reveals several patterns. Control & regulation dominated the early to mid-stages of the session, peaking at 85.93% between 40–50 minutes and tapering to 58.33% by the session's end. Planning & activation was used consistently, with a peak of 17.29% at 80–90 minutes, but dropped to 0% in the final 30 minutes. Monitoring was most visible at the beginning (0–10 mins and 30–40 mins) and showed a gradual increase, peaking at 16.67% near the end. Reaction & reflection was the least used initially (below 2%

in the first 50 minutes) but rose steadily, reaching a high of 25% in the final 10 minutes. These patterns suggest a shift in learners' metacognitive engagement—from strategic planning and control in the early stages to increased evaluation and reflection as the session progressed.

6. Discussion and Conclusion

The analysis reveals that learners using the CTA protocol verbalized their thoughts more frequently than those using RTA. CTA participants often reported all thoughts, including procedural actions like re-reading and paraphrasing, while RTA learners tended to omit such details, focusing instead on strategies they perceived as significant—such as interpreting simulator graphs. This aligns with earlier comparisons of RTA and CTA in usability studies (Birns et al., 2002). Our study contributes to the limited body of research comparing these protocols in the context of metacognition, especially in complex, open-ended learning tasks.

Both protocols were effective in identifying cognitive and metacognitive processes within the phases of planning & activation, monitoring, and reaction & reflection. The key difference was in the control & regulation phase, which appeared more frequently in CTA. This is likely because CTA captures in-the-moment actions without filtering, including unproductive steps, while RTA, being retrospective, leads learners to report only those actions they found meaningful. Additionally, CTA and RTA differed in the indicators of processes such as control strategy adaptation, context-specific knowledge gathering, prior knowledge activation, and adaptive attributions—the latter more common in RTA due to its reflective nature.

By capturing these distinctions, the study addresses a critical gap in the literature: understanding how RTA and CTA differ in capturing the richness, frequency, and temporal characteristics of metacognitive activity during complex problem-solving. The temporal data from CTA revealed how metacognitive engagement evolved across the session—control & regulation dominated early stages, while reaction & reflection peaked near the end. These patterns support SRL theory (Greene & Azevedo, 2010) and demonstrate CTA's value in uncovering temporal shifts that are otherwise inaccessible through RTA.

To conclude, this study investigated the differences between applying RTA and CTA protocols to analyze learners' cognitive and metacognitive processes while interacting with a complex problem-solving learning environment. Data were collected from 10 RTA and 4 CTA participants using MEttLE, an OELE designed for engineering estimation tasks. The verbalizations were coded using a mechanism grounded in Pintrich's SRL framework and MEttLE's pedagogical design, and the protocols were compared based on the frequency of metacognitive phases—planning & activation, monitoring, control & regulation, and reaction & reflection—and their associated processes. However, several limitations must be acknowledged. The small sample size limits the generalizability of our findings, although approximately 3700 verbalizations were collected, providing rich data for analysis. We did not explore the influence of learner personality, and the temporal characteristics of metacognitive phases may vary across different learning contexts. Some metacognitive processes, like model building, are specific to estimation tasks in MEttLE and may not transfer to other domains. Furthermore, the lab-based setup may have influenced learner behaviour, and the study focuses solely on overt indicators of metacognition, potentially missing more implicit cognitive activity. To address these limitations, future work will increase the sample size, consider learner characteristics, and incorporate tools like eye-tracking to enhance retrospective protocols beyond screen recordings (Paikrao et al., 2025).

Acknowledgements

We gratefully acknowledge Dr. Aditi Kothiyal, the participants, and the research team at IIT Bombay for their invaluable support and contributions to this study.

References

Alhadreti, O., & Mayhew, P. (2018, April). Rethinking thinking aloud: A comparison of three think-aloud protocols. In Proceedings of the 2018 CHI conference on human factors in computing systems.

Azevedo, R., & Wiedbusch, M. (2023). Theories of metacognition and pedagogy applied to AIED systems. In Handbook of artificial intelligence in education (pp. 45-67). Edward Elgar Publishing.
Birns, J. H., Joffre, K. A., Leclerc, J. F., & Paulsen, C. A. (2002, July). Getting the whole picture: Collecting usability data using two methods—-concurrent think aloud and retrospective probing. In Proceedings of UPA Conference (pp. 8-12).
Coiro, J., Knobel, M., Lankshear, C., & Leu, D. J. (Eds.). (2014). Handbook of research on new literacies. Routledge.
Efklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. Educational psychologist, 46(1), 6-25.
Ericsson, K. A. (2006). Protocol analysis and expert thought: Concurrent verbalisations of thinking during experts' performance on representative tasks. The Cambridge handbook of expertise and expert performance, 223-241.
Ericsson, K. A., & Simon, H. A. (1993). Protocol analysis (revised edition). Overview of methodology of protocol analysis.
Fan, Y., van der Graaf, J., Lim, L., Raković, M., Singh, S., Kilgour, J., ... & Gašević, D. (2022). Towards investigating the validity of measurement of self-regulated learning based on trace data. Metacognition and Learning, 17(3), 949-987.
Flavell, J. H. (1976). Metacognitive aspects of problem solving. The nature of intelligence.
Greene, J. A., Deekens, V. M., Copeland, D. Z., & Yu, S. (2017). Capturing and modeling self-regulated learning using think-aloud protocols. In Handbook of self-regulation of learning and performance (pp. 323-337). Routledge.
Greene, J. A., & Azevedo, R. (2010). The measurement of learners' self-regulated cognitive and metacognitive processes while using computer-based learning environments. Educational psychologist, 45(4), 203-209.
Hertzum, M. (2024). Concurrent or Retrospective Thinking Aloud in Usability Tests: A Meta-Analytic Review. ACM Transactions on Computer-Human Interaction, 31(3), 1-29.
Jääskeläinen, R. (2010). Think-aloud protocol. Handbook of translation studies, 1, 371-374.
Kothiyal, A., & Murthy, S. (2018). MEttLE: a modelling-based learning environment for undergraduate engineering estimation problem solving. Research and practice in technology enhanced learning.
Krieger, F., Azevedo, R., Graesser, A. C., & Greiff, S. (2022). Introduction to the special issue: the role of metacognition in complex skills-spotlights on problem solving, collaboration, and self-regulated learning. Metacognition and Learning, 1-8.
Munshi, A., & Biswas, G. (2019, June). Personalisation in OELEs: developing a data-driven framework to model and scaffold SRL processes. In International Conference on Artificial Intelligence in Education (pp. 354-358). Springer, Cham.
Paikrao, A., Nath, D., Rajendran, R., & Mitra, R., (2025). Enhancing Retrospective Think-Aloud Protocols with Eye-Gaze data. In B. Jiang et al. (Eds.), Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education.
Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. Frontiers in psychology, 8, 422.
Pathan, R., Murthy, S., & Rajendran, R. (2021, November). A coding mechanism for analysis of SRL processes in an open-ended learning environment. In International Conference on Computers in Education.
Pathan, R., Shaikh, U., & Rajendran, R. (2019, December). Capturing learner interaction in computer-based learning environment: design and application. In 2019 IEEE Tenth International Conference on Technology for Education (T4E) (pp. 146-153). IEEE.
Pathan, R., Singh, D., Murthy, S., & Rajendran, R. (2022, June). Identifying Metacognitive Processes Using Trace Data in an Open-Ended Problem-Solving Learning Environment. In International Conference on Intelligent Tutoring Systems (pp. 213-226). Cham: Springer International Publishing.
Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In Handbook of self-regulation (pp. 451-502). Academic Press.
Trevors, G., Duffy, M., & Azevedo, R. (2014). Note-taking within MetaTutor: Interactions between an intelligent tutoring system and prior knowledge on note-taking and learning. Educational Technology Research and Development, 62(5), 507-528.
Vaccaro, A. G., & Fleming, S. M. (2018). Thinking about thinking: A coordinate-based meta-analysis of neuroimaging studies of metacognitive judgements. Brain and neuroscience advances, 2, 2398212818810591.
Zimmerman, B. J. (2013). Theories of self-regulated learning and academic achievement: An overview and analysis. Self-regulated learning and academic achievement, 1-36.

Did You Mispronounce or Did I Mishear? – Detecting Kanji Mispronunciations in Children's Oral Reading

Takuya MATSUZAKI* & Arata SAITO
Graduate School of Science, Tokyo University of Science, Japan
*matuzaki@rs.tus.ac.jp

Abstract: Oral reading of textbooks is a common practice in Japanese elementary education, yet it poses a unique challenge: kanji characters often have multiple readings, and without a listener to provide feedback, students may unknowingly reinforce incorrect pronunciations. To address this, we propose a novel algorithm for detecting kanji mispronunciations in children's oral reading. Our method combines a deep learning-based automatic speech recognition (ASR) system with two probabilistic models: one modeling plausible kanji mispronunciations and the other modeling typical ASR errors. By aligning phoneme sequences generated from both the original text and the ASR output, the algorithm distinguishes genuine mispronunciations from transcription errors due to ASR. Experimental evaluation on speech data from children aged 6–9 shows that the proposed method successfully detects 84.6% of kanji mispronunciations that are included in the mispronunciation candidate dictionary of the probabilistic mispronunciation model.

Keywords: Kanji pronunciation, oral reading, corrective feedback, speech recognition

1. Introduction

Reading a textbook aloud is a traditional classroom activity in Japanese elementary and middle schools and still a common practice in first and foreign language classes as well as classes of other subjects. The fluency of oral reading has been identified as a key component of reading proficiency (Adams, 1990; Fuchs et al., 2001) and indicator of reading comprehension ability (Good et al., 2001; Roberts et al., 2005; Roehrig et al., 2008; Kim et al., 2010). Although there are some controversies about the directionality between reading fluency and reading comprehension (e.g., Little et al., 2017), we may at least expect that an oral reader can check her/his understanding about a text either by hearing her/him read it aloud or by feedback from a listener and may hopefully improve the understanding about the text.

Reading aloud from a textbook is thus a good candidate for a homework assignment. When doing it as homework, however, not every student has someone to listen to it. It is problematic since the execution of the homework is not guaranteed without a listener and engagement in the reading activity is naturally motivated by the existence of a listener, even if it is a robot (Nakadai et al., 2015) or a dog (Le Roux et al., 2014).

The absence of a listener causes another problem peculiar to Japanese. Kanji characters, i.e., Chinese characters in the Japanese orthographic system, generally have more than one pronunciation. For example, "上" (up) has at least the following seven pronunciations: *jou*, *shou*, *ue*, *uwa*, *kami*, *a* (as in "上げる," *a-ge-ru*), and *nobo* (as in "上る," *nobo-ru*). An oral reader has to select one of them according to the word that includes the kanji and sometimes by considering the context around the word. As a result, Japanese children, and even adults, often mispronounce a kanji in reading a text aloud. However, without a listener's corrective feedback, it is difficult for a reader to be aware of such mispronunciation and it may even strengthen the wrong association between a pronunciation and a word.

In this paper, we propose an algorithm for detecting mispronunciation of kanji in oral reading by children. Our long-term goal is to develop an AI-powered reading assistant that provides corrective feedback to children about the accuracy and fluency of their reading and motivates children's engagement in the oral reading homework. The kanji mispronunciation detection technique will be a key component of the system that helps foster the ability of choosing a proper pronunciation for kanji, which is indispensable to read Japanese text fluently.

Our kanji mispronunciation detection algorithm owes to recent advances in automatic speech recognition (ASR) based on deep learning. However, to detect mispronunciations accurately, it is not enough to simply compare the ASR result, i.e., an automatically generated transcript, against the text that was read aloud. This is because of the inherent difficulty of ASR. For instance, the word error rate (WER) of Whisper-medium model (Radford et al., 2023), which we used in the experiment, is reported to be 10.5% on a Japanese speech dataset. It means that one out of ten words is somehow misrecognized by ASR. Hence, if we report all discrepancy between the transcript and original text as potential mispronunciation, it yields too many false negatives.

To address this problem, we combine two probabilistic models: one for mispronunciation generation and the other for ASR errors. The mispronunciation generation model quantifies the plausibility of a wrong pronunciation for a word including specific kanji characters. The ASR error model captures the tendency of a phoneme to be recognized as another phoneme. By integrating them with a speech-to-text alignment algorithm, we discriminate between genuine mispronunciations and wrongly transcribed pronunciations. We evaluated the algorithm on oral reading of textbooks by children aged from 6 to 9. Experimental results show that our algorithm detects 84.6% of kanji mispronunciations included in the candidate dictionary of the probabilistic mispronunciation model.

The rest of the paper is organized as follows. Section 2 provides a summary of related work. Section 3 details the kanji mispronunciation detection algorithm. Section 4 presents experimental results. Section 5 concludes the paper.

2. Related Work

Several attempts have been made to measure the oral reading fluency and/or accuracy of students utilizing ASR (see, e.g., (Nese & Kamata, 2021) and references therein). Some recent works pursue the same goal while taking advantage of the progress in ASR by deep learning (Yıldız et al., 2024; Vaidya et al., 2024; Henkel et al., 2025; da Silva et al., 2025). While such automated assessment involves detection of mispronunciations in nature, it is not necessarily within their scope to accurately identify where and how a pronunciation error occurs in students' speech. This is because their goal is to achieve a high correlation between the fluency score obtained by ASR-based automatic method and human raters' scoring.

Another line of research has aimed at developing an automatic reading tutor based on ASR (Mostow et al., 1993; Duchateau et al., 2009; Bai et al., 2021). These systems provide feedback about missing word and pronunciation errors in oral reading in addition to the assessment of reading fluency. The main targets of these systems have however been English and other European languages. We are not aware of any previous studies that addressed the issue of identifying pronunciation errors due to logographic nature of a script system such as Chinese and Japanese, especially under the existence of ASR errors.

3. Method

Figure 1 shows an overview of our mispronunciation detection algorithm. Suppose a user read aloud a Japanese text "彼は約束を反故にした" (*kare wa yakusoku o hogo ni shita*; He broke his promise) and mispronounced the word "反故" (*ho-go*; break one's promise) as "*han ko*" because "反" may be read as "*han*" as in "反対" (*han-tai*; oppose) and "故" may be read as

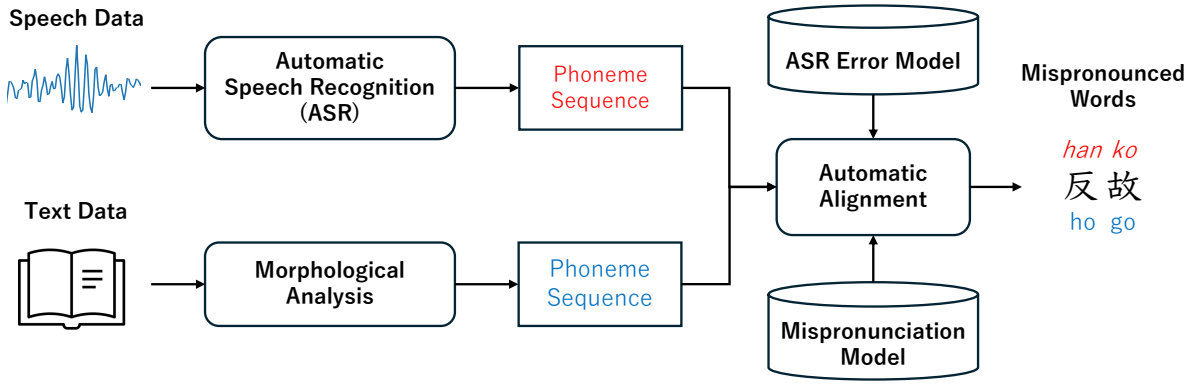


Figure 1. Overview of Mispronunciation Detection Algorithm.

"ko" as in "故郷" (*ko-kyo*; hometown). The speech data is then processed with an automatic speech recognition (ASR) system. Suppose the ASR system made a recognition error at the word "彼" (*kare*; he) and produced "*kore*" as its transcript but recognized other part of the speech correctly. We thus have the following phoneme sequence as the result of ASR:

k o r e w a y a k u s o k u o h a n k o n i sh i t a

Meanwhile, the same text is processed with a morphological analyzer. Suppose it produced a correct phoneme sequence as its reading:

k a r e w a y a k u s o k u o h o g o n i sh i t a

By comparing these two phoneme sequences, we want to detect the user's mispronunciation of the word "反故" (*han ko* / ho go) while silently ignoring the mis-recognition of the word "彼" (*kore* / kare) since it is due to an error of ASR. To this end, we utilize two probabilistic models: one for mispronunciations and the other for ASR errors. The mispronunciation model provides the (estimated) probability of the wrong pronunciations, P(*han* | 反) and P(*ko* | 故), while the ASR error model provides the probability of the mis-recognition, P(*o* | a), for "彼" (*kore* / kare). By combining these probabilities and finding the most plausible alignment between the two phoneme sequences, the system reports the mispronunciation of "反故" while ignoring the ASR error at "彼". The rest of this section explains the details of the algorithm.

3.1 Fine-tuning and Adaptation of ASR model

We used Whisper-medium (Radford et al., 2023) model for the ASR. Whisper is a neural ASR model trained on 680,000 hours of multilingual speech data collected from the internet. We further trained it for two purposes: to produce phoneme sequence instead of orthographic text (i.e., one including kanji and other Japanese characters) and to adapt it to children's speech.

To produce phoneme sequences as the ASR results, we fine-tuned Whisper using approximately 100 hours of speech with transcripts taken from the Corpus of Spontaneous Japanese (CSJ) (Maekawa, 2003). The CSJ corpus is a collection of spontaneous speech by adults, such as conference talks.

The fine-tuned model's accuracy (phoneme recall) was 95.5% on adults' speech but it dropped to 90.0% on children's speech. To obtain better ASR results on children's speech, we further trained the model using 100.5 hours of speech data by 160 children (age 6 – age 8) taken from ATR's Japanese Elementary School Children Speech Database.[1] The adapted model recognized children's speech with the accuracy of 95.9%.

[1] https://www.atr-p.com/products/dbpdf/ATR_JuniorDB.pdf

3.2 Speech Recognition Error Model

To discern a mispronunciation from ASR errors, we need to quantify the plausibility of each ASR error type. Specifically, we estimated the following three types of probabilities by counting our ASR model's errors made on ATR's children speech database:

- $P(\text{ASR} = q \mid \text{pron} = p)$: the probability of outputting, either correctly or wrongly, a phoneme q for a pronounced phoneme p.
- $P(\text{ASR} = q \mid \text{pron} = \phi)$: the probability of outputting a phoneme q despite no corresponding phoneme was pronounced (i.e., phoneme q is inserted to the output).
- $P(\text{ASR} = \phi \mid \text{pron} = p)$: the probability of outputting nothing for a pronounced phoneme p (i.e., phoneme p is missing in the output).

3.3 Mispronunciation Model

Our probabilistic mispronunciation model has two components:

- $P(\text{mp} = 1 \mid \text{char} = c)$: the probability that a kanji character c is mispronounced by a child.
- $P(\text{pron} = p_1 p_2 \cdots p_n \mid \text{char} = c, \text{mp} = 1)$: the conditional probability that a kanji c is pronounced as $p_1 p_2 \cdots p_n$ given that c is mispronounced.

By estimating these two types of probabilities, we can quantify, for instance, the probability of the mispronunciation of "反故" as *han ko* as follows:

$$P(\text{mp} = 1 \mid \text{char} = 反)\, P(\text{pron} = han \mid \text{mp} = 1, \text{char} = 反)$$
$$\times\ P(\text{mp} = 1 \mid \text{char} = 故)\, P(\text{pron} = ko \mid \text{mp} = 1, \text{char} = 故).$$

Our current estimation of the probability $P(\text{mp} = 1 \mid \text{char} = c)$ is rather crude. We simply assume it does not depend on the pronounced character c:

$$P(\text{mp} = 1 \mid \text{char} = c) \ = \ \varepsilon.$$

The small constant probability ε was set to 0.05. Once we obtain a larger dataset of children's speech, we may be able to obtain a better estimate of this probability, which depends on the pronounced character as well as the traits of the reader (e.g., her/his age).

To estimate the probability of mispronouncing a kanji character c as phoneme sequence $p_1 p_2 \cdots p_n$, we counted the relative frequency of that reading in newspaper articles. Our assumption is that children tend to mispronounce a kanji character with a pronunciation that she/he hears often for that character. Since it is difficult to collect children's daily oral communication data, we instead use newspaper articles as a rough approximation of it. Specifically, we processed all the Mainichi Newspaper articles published between 1991 to 2018 by a morphological analyzer and estimated the pronunciation of each occurrence of kanji characters in the articles. By counting the number of occurrences of character c and its pronunciation $p_1 p_2 \cdots p_n$, we estimated the mispronunciation probability as follows:

$$P(\text{pron} = p_1 p_2 \cdots p_n \mid \text{char} = c, \text{mp} = 1) = \frac{\text{Count}(c \text{ with pronunciation } \ p_1 p_2 \cdots p_n)}{\text{Count}(c)}.$$

In addition to the mispronunciation probabilities, we estimated the probability that a phoneme p appearing in a text is not pronounced, denoted $P(\text{pron} = \phi \mid \text{phoneme} = p)$, and the probability that a phoneme p not appearing in a text is pronounced, denoted $P(\text{pron} = p \mid \text{phoneme} = \phi)$. We obtained these estimates by comparing the transcript and the source text in ATR's children speech database.

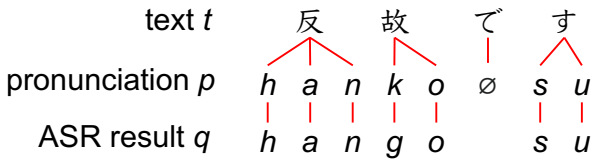


Figure 2. Alignment among Text, Pronunciation, and ASR Result.

3.4 Alignment between Speech and Text

Based on the probability models of ASR errors and children's mispronunciations, we model the generation process of children's oral reading and its speech recognition result as follows:

1. A child reads a text aloud. In reading it, she/he may mispronounce some of the kanji characters in the text, insert some unnecessary phonemes, and/or skip some phonemes, with the probabilities according to the mispronunciation model.

2. Given the recorded oral reading as input, an ASR system produces its transcript as a phoneme sequence. During it, some of the phonemes may be mis-recognized or totally missed (skipped), and some unnecessary phonemes may be inserted to the transcript, with the probabilities according to the ASR error model.

Let t denote the text that is read, $p_1 p_2 \cdots p_n$ denote the child's pronunciation of the text expressed as a phoneme sequence, and $q_1 q_2 \cdots q_m$ denote the result of ASR. The text t and the ASR result $q_1 q_2 \cdots q_m$ are observable as symbol sequences, but the true phoneme sequence $p_1 p_2 \cdots p_n$ pronounced by the child is only recorded as audio data and not directly observable as a symbol sequence. Our aim is to recover the child's true pronunciation from the observable data. We do so by finding the most probable phoneme sequence $\hat{p}_1 \hat{p}_2 \cdots \hat{p}_n$ given the text t and the ASR result $q_1 q_2 \cdots q_m$:

$$\begin{aligned} \hat{p}_1 \hat{p}_2 \ldots \hat{p}_n &= \underset{p_1 p_2 \cdots p_{n'}}{\operatorname{argmax}} P(\text{pron} = p_1 p_2 \cdots p_{n'} \mid \text{text} = t, \text{ASR} = q_1 q_2 \cdots q_m) \\ &= \underset{p_1 p_2 \cdots p_{n'}}{\operatorname{argmax}} P(\text{pron} = p_1 p_2 \cdots p_{n'}, \text{ASR} = q_1 q_2 \cdots q_m \mid \text{text} = t) . \end{aligned} \tag{1}$$

The mispronunciation model and the ASR error model assume an alignment among the characters in the text, the pronounced phonemes, and the phonemes in an ASR result. We denote an alignment by a. Figure 2 shows, by red lines, an example of alignment among text $t =$ "反故です" (*ho-go de su*; the promise was broken), pronunciation $p =$ "h a n k o s u", and ASR result $q =$ "h a n g o s u". Since the alignment is not observable, the rightmost-hand side of Eq. (1) takes the form of a summation over all possible alignments, but we approximate it with the most plausible alignment (i.e., so-called Viterbi-approximation):

$$\begin{aligned} \hat{p}_1 \hat{p}_2 \ldots \hat{p}_n &= \underset{p_1 p_2 \cdots p_{n'}}{\operatorname{argmax}} \sum_a P(\text{pron} = p_1 p_2 \cdots p_{n'}, \text{ASR} = q_1 q_2 \cdots q_m, \text{align} = a \mid \text{text} = t) \\ &\approx \underset{p_1 p_2 \cdots p_{n'}}{\operatorname{argmax}} \max_a P(\text{pron} = p_1 p_2 \cdots p_{n'}, \text{ASR} = q_1 q_2 \cdots q_m, \text{align} = a \mid \text{text} = t) . \end{aligned} \tag{2}$$

Assuming the probabilistic independence between the pronunciation of the characters in the text, we decompose the last probability as follows:

$$\begin{aligned} &P(\text{pron} = p_1 p_2 \cdots p_{n'}, \text{ASR} = q_1 q_2 \cdots q_m, \text{align} = a \mid \text{text} = t) \\ &= \prod_{i=1}^{N} P(\text{pron} = p_1^{(i)} p_2^{(i)} \cdots p_{n_i}^{(i)}, \text{ASR} = q_1^{(i)} q_2^{(i)} \cdots q_{m_i'}^{(i)}, \text{align} = a^{(i)} \mid \text{char} = c_i) \end{aligned}$$

$$= \prod_{i=1}^{N} \left\{ \begin{array}{l} P\left(\mathrm{ASR} = q_1^{(i)} \cdots q_{m_i}^{(i)} \mid \mathrm{pron} = p_1^{(i)} \cdots p_{n_i}^{(i)}, \mathrm{align} = a^{(i)}\right) \\ \quad \times P(\mathrm{pron} = p_1^{(i)} \cdots p_{n_i}^{(i)}, \mathrm{align} = a^{(i)} \mid \mathrm{char} = c_i) \end{array} \right\}, \quad (3)$$

where we assume the text t is $c_1 c_2 \cdots c_N$ as a character sequence, $p_1 p_2 \cdots p_{n'} = p_1^{(1)} \cdots p_{n_N}^{(N)}$, $q_1 q_2 \cdots q_m = q_1^{(1)} \cdots q_{m_N}^{(N)}$, and $a = a^{(1)} \cdots a^{(N)}$. The two sorts of probabilities, $P(\mathrm{ASR} \mid \mathrm{pron}, \mathrm{align})$ and $P(\mathrm{pron}, \mathrm{align} \mid \mathrm{char})$, in the rightmost-hand side of Eq. (3) can further be decomposed using the ASR error model and the mispronunciation model's component probabilities.

For instance, suppose that, as shown in Figure 2, the text is "反故です" and the reader mispronounced the first character "反" as *han* and skip pronouncing "で" (*de*). Also suppose that the ASR system misrecognized the (correct) pronunciation for "故" (*go*) as *ko*. The probability of this process is calculated as follows:

$P(\mathrm{pron} = \mathrm{h\ a\ n\ g\ o\ s\ u}, \mathrm{ASR} = \mathrm{h\ a\ n\ k\ o\ s\ u}, \mathrm{align} = a \mid \mathrm{text} = \text{反故です})$
$= P(\mathrm{ASR} = \mathrm{h\ a\ n} \mid \mathrm{pron} = \mathrm{h\ a\ n})\, P(\mathrm{pron} = \mathrm{h\ a\ n} \mid \mathrm{mp} = 1, \mathrm{char} = \text{反})\, P(\mathrm{mp} = 1 \mid \mathrm{char} = \text{反})$
$\times P(\mathrm{ASR} = \mathrm{k\ o} \mid \mathrm{pron} = \mathrm{g\ o})\, P(\mathrm{mp} = 0 \mid \mathrm{char} = \text{故})$
$\times P(\mathrm{pron} = \phi \mid \mathrm{char} = \text{で})$
$\times P(\mathrm{ASR} = \mathrm{s\ u} \mid \mathrm{pron} = \mathrm{s\ u})\, P(\mathrm{mp} = 0 \mid \mathrm{char} = \text{す}).$

By further decomposing the ASR recognition probabilities such like

$$P(\mathrm{ASR} = \mathrm{k\ o} \mid \mathrm{pron} = \mathrm{g\ o}) = P(\mathrm{ASR} = \mathrm{k} \mid \mathrm{pron} = \mathrm{g}) P(\mathrm{ASR} = \mathrm{o} \mid \mathrm{pron} = \mathrm{o}),$$

and substituting the estimated probabilities according to the component models, we have a probability score for the above hypothetical process.

The number of possible pronunciations $p_1 p_2 \cdots p_{n'}$ and alignments a considered in Eq. (1) is exponentially large in the length of the text. It is however possible to find the most probable pronunciation efficiently by using a dynamic programming algorithm that extends minimal edit distance calculation between two sequences (Kaji, 2023).

4. Experiment

4.1 Evaluation Data

For the evaluation of the proposed method, we recorded oral reading of paragraphs taken from textbooks by 249 elementary school students of grades 1 through 3 (ages 6 through 9). Some students provided more than one recorded file, and 370 files (6.8 hours in total) were collected. The audio speech files were manually transcribed and annotated with alignment to the text as well as mispronunciations, fillers, and repetition in the speech. The audio data and the transcription were processed with Julius (Lee et al., 2001) to obtain phoneme-level timestamps (i.e., forced-alignments) and segmented with pauses into utterances shorter than 30 seconds, which is the maximum duration of Whisper's input. On some of the speech data, the forced-alignment by Julius was failed presumably due to high noise-level of the recordings. The evaluation hereafter is based on 122 utterances (2 hours in total) thus obtained.

4.2 Mispronunciation Detection Results and Error Analysis

There were 13 occurrences of kanji mispronunciation in the evaluation data. The proposed method successfully detected 11 cases (84%) of them. Table 1 shows the mispronounced words, the speaker's mispronunciations, and the number of correctly detected ones.

Meanwhile, there were 35 cases where the system wrongly detected mispronunciation despite the reader correctly pronounced the text. We analyzed the causes of the misdetections

by inspecting the ASR output and the probability of mispronunciation and ASR errors according to the models. Table 2 provides the summary of the reason for the misdetections.

The most frequent cause of misdetection was that a word (or a part of a word) skipped by the speaker tend to be aligned with a mispronunciation that is shorter than the correct pronunciation. In our mispronunciation model, a phoneme p is assumed to be skipped by a speaker with probability $P(\text{pron} = \phi \mid \text{phoneme} = p)$, independently of its context. Thus, the probability for a word being skipped is exponentially small in the length of the word and often becomes smaller than the product of *P*(a word is mispronounced as a shorter word) and *P*(the mispronunciation is skipped). In actuality, a whole word, rather than a phoneme in it, is often skipped, and the true probability of such a skip of a whole word is much larger than one assumed in the current model. We expect that this type of error can be remedied by refining the mispronunciation model and relaxing the false independence assumption.

The second frequent cause of misdetection was due to ASR errors where the output of ASR is very close to or identical to a mispronunciation of a word despite that the speaker correctly pronounces it. In other words, a detection error of this type happens when the probability of ASR error is under-estimated. Upon inspection, we found that such problematic ASR errors often include phoneme-level errors typically made on young children's speech, such as a misrecognition between /j i/ and /ch i/. Thus, this type of error can be remedied either by a better adaptation of the ASR model to children's speech or by a refinement of the ASR error model, possibly by incorporating a speaker's trait, such as age, into the input.

Table 1. *Kanji Mispronunciations in Evaluation Data and Detection Results*

Word	Correct Pronunciation	Speaker's Pronunciation	Occurrences	Correctly Detected
大かい (athletic meet)	t a i k a i	o o k a i	1	1
金 (gold)	k i n	k a n e	1	1
せん手 (athlete)	s e n sy u	s e n t e	1	1
国 (nation)	k u n i	k o k u	1	1
年 (year)	t o sh i	n e n	1	1
大そう (very)	t a i s o o	o o s o o	1	1
見回した (look around)	m i m a w a sh i t a	m i k a i sh i t a	1	1
上あご (upper jaw)	u w a a g o	u e a g o	3	1
入れて (put in)	i r e t e	h a i cl t e	3	3

Table 2. *Reasons for Misdetections of Kanji Mispronunciation*

Cause of Misdetection	Number of Misdetections
Speaker skipped (a part of) a word, and system detected it as skipping (a part of) a mispronounced word	18
ASR system wrongly outputs a phoneme sequence that is close or identical to a mispronunciation of a word	14
Speaker repeated a word, and system detected it as a mispronunciation of a word nearby	3

5. Conclusion

This study introduced a method for detecting kanji mispronunciations in children's oral reading by integrating a fine-tuned ASR system with probabilistic models for both pronunciation errors and ASR recognition errors. The experimental results, achieving 84.6% detection accuracy, demonstrated the potential of this approach in supporting automated reading feedback. However, the analysis also revealed limitations, such as model sensitivity to phoneme skipping and underestimation of specific ASR error types. Addressing these issues through improved modeling of pronunciation context and speaker traits—such as age—could further enhance

detection accuracy. Ultimately, our method provides a core component for AI-based reading support tools aimed at improving children's reading fluency and comprehension of Japanese texts through individualized, automated corrective feedback.

Acknowledgements

This study was supported by JSPS KAKENHI Grant Number JP21H04416.

References

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. MIT Press.

Bai, Y., Tejedor-García, C., Hubers, F., Cucchiarini, C., & Strik, H. (2021). An ASR-Based Tutor for Learning to Read: How to Optimize Feedback to First Graders. *Proceedings of Speech and Computer: 23rd International Conference (SPECOM 2021)* (pp. 58–69), Springer-Verlag, Berlin. https://doi.org/10.1007/978-3-030-87802-3_6

Duchateau, J., Kong, Y. O., Cleuren, L., Latacz, L., Roelens, J., Samir, A., Demuynck, K., Ghesquière, P., Verhelst, W., & Van hamme, H. (2009). Developing a reading tutor: Design and evaluation of dedicated speech recognition and synthesis modules. *Speech Communication*, 51(10), 985-994.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading*, 5(3), 239–256. https://doi.org/10.1207/S1532799XSSR0503_3

Good, R. H. III, Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utilit y of a continuum of fluency-based indicators of foundational reading skills for third-grade high-sta kes outcomes. *Scientific Studies of Reading*, 5(3), 257–288. https://doi.org/10.1207/S1532799XS SR0503_4

Henkel, O., Horne-Robinson, H., Hills, L., Roberts, B., & McGrane, J. (2025). Supporting Literacy Assessment in West Africa: Using State-of-the-Art Speech Models to Assess Oral Reading Fluency. *International Journal of Artificial Intelligence in Education*, 35, 282–303. https://doi.org/10.1007/s40593-024-00435-9

Kaji, N. (2023). Lattice path edit distance: A Romanization-aware edit distance for extracting misspelling-correction pairs from Japanese search query logs. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track* (pp. 233–242). Association for Computational Linguistics.

Kim, Y.-S., Petscher, Y., Schatschneider, C., & Foorman, B. (2010). Does growth rate in oral reading fluency matter in predicting reading comprehension achievement? *Journal of Educational Psychology*, 102(3), 652–667. https://doi.org/10.1037/a0019643

Lee, A., Kawahara, T., & Shikano, K. (2001) Julius – An Open Source Real-Time Large Vocabulary Recognition Engine. *Proceedings of 7th European Conference on Speech Communication and Technology* (pp.1691–1694). International Speech Communication Association.

Little, C. W., Hart, S. A., Quinn, J. M., Tucker-Drob, E. M., Taylor, J., & Schatschneider, C. (2017). Exploring the co-development of reading fluency and reading comprehension: A twin study. *Child Development*, 88(3), 934–945. https://doi.org/10.1111/cdev.12670

Maekawa, K. (2003). Corpus of Spontaneous Japanese: its design and evaluation. *Proceedings of The ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition* (pp. 7-12).

Mostow, J., Hauptmann, A. G., Chase, L. L., & Roth, S. (1993). Towards a reading coach that listens: automated detection of oral reading errors. *Proceedings of the eleventh national conference on Artificial intelligence* (pp. 392–397). AAAI Press.

Nakadai, H., Hee, L. S., Kitajima, M., Hoshino, J. (2015). KINJIRO: Animatronics for Children's Reading Aloud Training. *Entertainment Computing - ICEC 2015* (pp. 252-260). Springer-Verlag. https://doi.org/10.1007/978-3-319-24589-8_19

Nese, J. F. T., & Kamata, A. (2021). Evidence for automated scoring and shorter passages of CBM-R in early elementary school. *School Psychology*, 36(1), 47–59. https://doi.org/10.1037/spq0000415

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *Proceedings of the International Conference on Machine Learning*, (pp. 28492-28518). JMLR.org.

Roberts, G., Good, R., & Corcoran, S. (2005). Story retell: A fluency-based indicator of reading compr ehension. *School Psychology Quarterly*, 20(3), 304–317. https://doi.org/10.1521/scpq.2005.20.3. 304

Roehrig, A. D., Petscher, Y., Nettles, S. M., Hudson, R. F., & Torgesen, J. K. (2008). Accuracy of the DIBELS oral reading fluency measure for predicting third grade reading comprehension outcomes. *Journal of School Psychology*, 46(3), 343–366. https://doi.org/10.1016/j.jsp.2007.06.006

Le Roux, M. C., Swartz, L., & Swart, E. (2014). The effect of an animal-assisted reading program on the reading rate, accuracy and comprehension of grade 3 students: A randomized control study. *Child & Youth Care Forum*, 43(6), 655-673.

da Silva, G. C., Rodrigues, R. L., Amorim, A. N., Jeon, L., Albuquerque, E. X. S., Silva, V. C., da Silva, V. F., Pinheiro, A. L. A., Nunes, J. P. J. R., de Souza, S. X. M. G., Silva, M. S., Mauro, I., & Maciel, A. M. A. (2025). Assessing reading fluency in elementary grades: A machine learning approach, *Computers and Education: Artificial Intelligence*, 8, 100411.

Vaidya, M., Sahoo, B. K., & Rao, P. (2024). Deep Learning for Assessment of Oral Reading Fluency. arXiv. https://doi.org/10.48550/arXiv.2405.19426

Yıldız, M., Keskin, H. K., Oyucu, S., Hartman, D. K., Temur, M., & Aydoğmuş, M. (2024). Can Artificial Intelligence Identify Reading Fluency and Level? Comparison of Human and Machine Performance. *Reading & Writing Quarterly*, 41(1), 66–83.

GAI Pedagogical Agents and Teacher-Guided Prompt Use in EFL Debate Training

Farhad A.K. CASSIM[a, b*] & Jie Chi YANG[a]
[a]*Graduate Institute of Network Learning Technology, National Central University, Taiwan*
[b]*General Education Center, National Yunlin University of Science and Technology, Taiwan*
*cassim@yuntech.edu.tw

Abstract: This study investigated the pedagogical functions and learner uptake associated with two generative AI approaches for training EFL students' debating skills. Conducted within the context of a Model United Nations (MUN) activity in two university EFL classes, 16 students from class A engaged with a generative AI pedagogical agent (GAI-PA), while another 16 from class B used teacher-guided ChatGPT prompts to enhance their debating over an 8-week period. The instructions embedded in the two approaches were analyzed using a constructivist grounded theory approach to determine the specific pedagogical functions emerging from each system. After exposure, the learners' uptake of the functions was examined through a deductive analysis and rating of their responses recorded in their conversation sessions. Results show the GAI-PA enforced a highly structured and pedagogically driven design, producing consistently high learner uptake across most functions. In contrast, the teacher-guided prompt system relied on greater learner initiative and yielded more variable uptake, producing particularly low learner uptake in cognitively demanding areas such as discourse construction, revision, and ethical AI usage.

Keywords: Debating, English as a foreign language (EFL), generative AI pedagogical agent (GAI-PA), pedagogical functions, teacher-guided prompt use

1. Introduction

Debating enhances EFL learners' communication skills in academic and real-world settings (García-Sánchez, 2020), but it also presents challenges. Learners struggle with fear of mistakes and negative evaluation (Gregersen & Horwitz, 2002), and also low self-confidence (Aslan & Thompson, 2021). Cultural norms emphasizing politeness and group harmony further reduce assertiveness in adversarial formats (Nursanti et al., 2023).

Generative AI (GAI) tools like ChatGPT, however, offer new potential for debate training through interactive, human-like responses. Despite their growing use, the literature lacks studies comparing the pedagogical effectiveness of the two main approaches— GAI pedagogical agents (GAI-PA) and teacher-guided prompt use— through which GAI can be employed for EFL debate training. Specifically, the literature does not identify the pedagogical functions each supports and how learners engage with them. Yet, understanding this can guide educators in choosing strategies to optimize GAI-assisted debate learning. Hence, this study sought to address these research questions (RQs):

1. What pedagogical functions for EFL debate training emerge from the GAI-PA and how are they different from those emerging from teacher-guided prompt use?
2. How do learners take up these functions during their debate practice?

2. Literature Review

2.1 Debating and the Model United Nations

Debating is a structured form of argumentation involving arguments, counterarguments, and rebuttals to persuade an audience or judges (El Majidi et al., 2021; Savitz et al., 2021).

For upper-intermediate EFL learners, it is an important but cognitively demanding goal due to the need for research, organization, and autonomy (Council of Europe, 2020).

One popular format for encouraging student engagement in debating is the Model United Nations (MUN), where students act as country delegates debating global issues (Calossi & Coticchia, 2018; Nfor, 2023). MUN activities include research, position paper writing, speech delivery, and debates and diplomacy—requiring high levels of student-centered learning (Hazen, 2019). Through offering learners dynamic role-play and a structured learning path from preparation to live debating, however, MUN fosters engagement and builds key skills like negotiation and public speaking (McIntosh, 2001; O'Dell et al., 2024).

Being a face-to-face format of debating, though, MUN can trigger anxiety and limits revision opportunities, especially sentence-level revisions, due to its real-time pace (Hirci & Peterlin, 2020; Ho, 2015; Liu & Sadler, 2003). Pre-debate practice with GAI tools like ChatGPT, on the other hand, can offer a low-pressure environment to develop linguistic and strategic competence ahead of classroom debates.

2.2 Pedagogical Functions of Teacher-Guided Prompt Use and GAI-PAs

Prompt engineering involves training students or teachers to craft effective prompts that guide tools like ChatGPT toward high-quality, pedagogically relevant responses. As a teacher-guided use of AI, it enhances learning by improving instructional precision and adaptability. It supports pedagogical functions— specific instructional operations such as guidance, emotional support, and reflective learning—that structure the learning process (Lee & Lee, 2024; Yusuf et al., 2025). Wang et al. (2023) found that effective prompt engineering improved information quality and efficiency in flipped classrooms. Darmawansah et al. (2025) showed that strong prompts enhanced ChatGPT's adaptability and students' argumentation. Ghafouri et al. (2024) reported that prompt quality improved instructional clarity and boosted both teacher self-efficacy and student writing. These studies demonstrate how teacher-guided prompting enables AI to deliver targeted instructional support.

Despite the positive results reported, teacher-guided prompt use may not always lead to high quality prompts in students. Darmawansah et al. (2025), for example, assert that despite students significantly higher argumentative speaking performance as a result of their learning in ChatGPT, prompt quality was still a significant issue and therefore encouraged educators to pay more attention to this. Woo et al. (2024) also found that learners experienced high cognitive overload due to the prompt engineering required for complex language learning tasks. Finally, an issue that cannot be overlooked is whether students have the self-regulated learning strategies needed for recognizing problems and designing better approaches, which prompt engineering for complex tasks like debating would require. In many EFL contexts, students have been exposed to long and sustained usage of overly-strict teacher-centered instruction (Yuan, 2024). Under the influence of such backgrounds, students exhibit amotivation, fear of failure, and avoidance of challenges (Bartholomew et al., 2018).

An alternative to prompt engineering is to employ GAI-PAs, that is GAIs that purposefully function as AI educators. In ChatGPT, this option is available in its MyGPT area that allows creators to customize their own GPT. Custom GPTs are transformed into GAI-PAs by uploading domain-specific resources (e.g. textbooks) and providing explicit pedagogical instructions to guide GIA-user interactions (Collins et al., 2024; Sevgi et al., 2024).

Recent studies have begun to report how custom GPTs, when embedded with pedagogical frameworks, can deliver targeted instructional functions often difficult to achieve solely with teacher-guided prompt use. For example, *AnatomyGPT* was developed as a domain-specific tutor. It's ability to provide rationales and citations points based on its knowledge base and pedagogical guidance made it a superior source of transparent and source-based academic learning for students (Collins et al., 2024). *GamiflcA Edu*, designed to support educators in implementing gamification and serious games, also achieved similar results. By means of the instructional frameworks embedded in its system, it was able to make use of the pedagogical function of scaffolded design to generate content tailored to diverse instructional contexts (López-Galisteo et al., 2025). Like other systems, both systems demonstrate that GAI-PAs achieve pedagogical functions that exhibit strong user uptake.

2.3 Guidelines for Implementing AI-Assisted Language Learning

While both forms of GAI-assisted learning can deliver pedagogical value, differences in how they embed and execute instructional functions raises critical questions about general guidelines both should follow to ensure best practices. Drawing on the literature in AI-assisted learning and the learning needs of students pursuing upper intermediate English proficiency, 5 guidelines emerge for implementing GAI-assisted language learning in an MUN context. GAI should enable 1) *adaptive learning paths*, allowing learners to select topics and adjust complexity for skill progression, 2) *real-time feedback and scaffolding* to guide self-correction in vocabulary, grammar, diplomatic language, and argument structure, 3) *engagement and interactive learning* through dynamic simulations and role-playing to improve students communication, 4) *self-regulated learning* to encourage goal-setting, performance monitoring, and reflection, and 5) *AI literacy and ethics* to ensure learners critically assess AI-generated arguments, detect biases, and maintain diplomatic accuracy (Chang et al., 2023; Khalil et al., 2024; Ng et al., 2024; Pan et al., 2024; Qadhi, et al., 2024; Shi & Aryadoust 2024; Wei, 2023).

The instructional architecture of both teacher-guided prompts and GAI-PAs should reflect these principles, yet it is important to note that their designs differ in assumptions about learner needs and instructional control. These differences affect the type of support provided, cognitive demands, learner autonomy, and the consistency of pedagogical delivery. Understanding their distinct functions and how learners respond is thus crucial for designing effective, context-sensitive AI learning systems.

3. Methods

3.1 Research Approach

This study adopted a qualitative research design. Students from two classes with comparable levels and characteristics were assigned to either the experimental (GAI-PA) or control system (teacher-guided prompt use) for 8 weeks of debate training. After the intervention, a deductive analysis and rating of the learners' responses on their respective systems were conducted to determine learner uptake of each system's pedagogical functions. These pedagogical functions had been previously identified using a constructivist grounded theory analysis of each system's embedded instructions.

3.2 Participants

The participants were drawn from two intact sophomore English Presentation and Writing classes in the College of Management, where two EFL instructors were jointly implementing a semester-long Model United Nations project using comparable instruction, goals, and content. With their consent, an email invitation was sent and 56 students enrolled in the supplementary debate training program. Following a purposeful selection procedure, 32 of the 56 students were selected and assigned to the experimental (n = 16, Class A) and control group (n = 16, Class B). The rest received the same intervention as their class peers but were excluded from the study due to incomplete sampling. All participants had been placed by the university in advanced-level English classes based on their TOEIC scores (M = 652 for experimental; M = 667 for control). For our study, their scores were deemed suitable for debate training (Council of Europe, 2020). The groups were also balanced in gender (experimental: M = 7, F = 9; control: M = 8, F = 8) and shared similar profiles, including limited debate and argumentative writing experience, high interest in GAI, and little prior exposure to teacher-guided prompt use. Overall, the final sample of 32 students (16 per group) reflected both the voluntary nature of the participation and purposeful sampling to ensure group equivalence by reducing confounding variables, though this reduced overall sample size.

3.3 Procedure

3.3.1 Exposure to the Intervention

All the participants from the two classes who enrolled were informed of the study's purpose—to explore the use of GAI tools like ChatGPT in debate training—and signed a consent form. They then completed a participant profile. Based on their participant profile, 16 students from class A and 16 students from class B were chosen to form the experimental and control groups and were then assigned to their respective classes to be taught by experienced EFL instructors familiar with teacher-guided prompt use. The intervention spanned weekly onsite sessions of 2 hours for 8 weeks and proceeded in four phases: 1 week of training, 3 weeks of adaptive practice, 3 weeks of unguided practice, and 1 week of review and reflection. After each practice with their respective systems, students submitted links of their conversation session in ChatGPT for feedback to guide subsequent practice. The study concluded with appreciation expressed to the participants and instructors, who were invited to a forum to receive the study's results and share reflections.

3.3.2 System Design

A GAI-PA, *MUN Debate Coach*, was developed using ChatGPT's MyGPT tool. Table 1 shows the system's two modules, illustrating how it supported both scaffolded practice and independent performance, while figure 1 shows its general pedagogical pathway.

Table 1. *Modules Installed in MUN Debate Coach*

Modules	Stages	Description
Neonate	1. Debater's Profile	Student prompted to complete profile
	2. Sub-Issue & Country	Student and ChatGPT decide roles and topic
	3. Guided Debating	ChatGPT provides scaffolded learning and feedback to enable guided debating
	4. Reflection	Student reflects on original debater's profile
Tribune	1. Debater's Profile	Student completes debater's profile with prompts from ChatGPT
	2. Sub-issue & Country	Student and ChatGPT decide roles and topic
	3. Debating	Student and ChatGPT present and defend arguments, with learner independently constructing his parts and ChatGPT providing feedback
	4. Reflection	Student reflects on original debater's profile

The core of the system was the instructional files in its backend containing pre-installed prompts designed using a chain-of-reasoning approach. To execute the core instructions, the system first put itself into a predetermined context defining its desired mode of operation. The system's entire instructional pathway depended on knowing the previous and anticipating the next step, preventing unscripted reasoning paths, thus creating interlocked steps. Prompts were also carefully crafted with consistent terminology, clear phrasing, and transition markers to guide the system. They also combined procedures with examples, included scripts and directives outlining what the system should and should not do, and instructed the GAI-PA to "speak out" the steps it would follow, reinforcing instructional awareness. Once the system was completed, it underwent iterative refinement through pilot testing with students similar to the target group until it achieved the consistency required for experimental use.

The control group engaged with ChatGPT through a teacher-guided, prompt-based model simulating structured debate preparation. Unlike the experimental group's closed, scaffolded AI system, this semi-structured design used teacher-curated prompts and

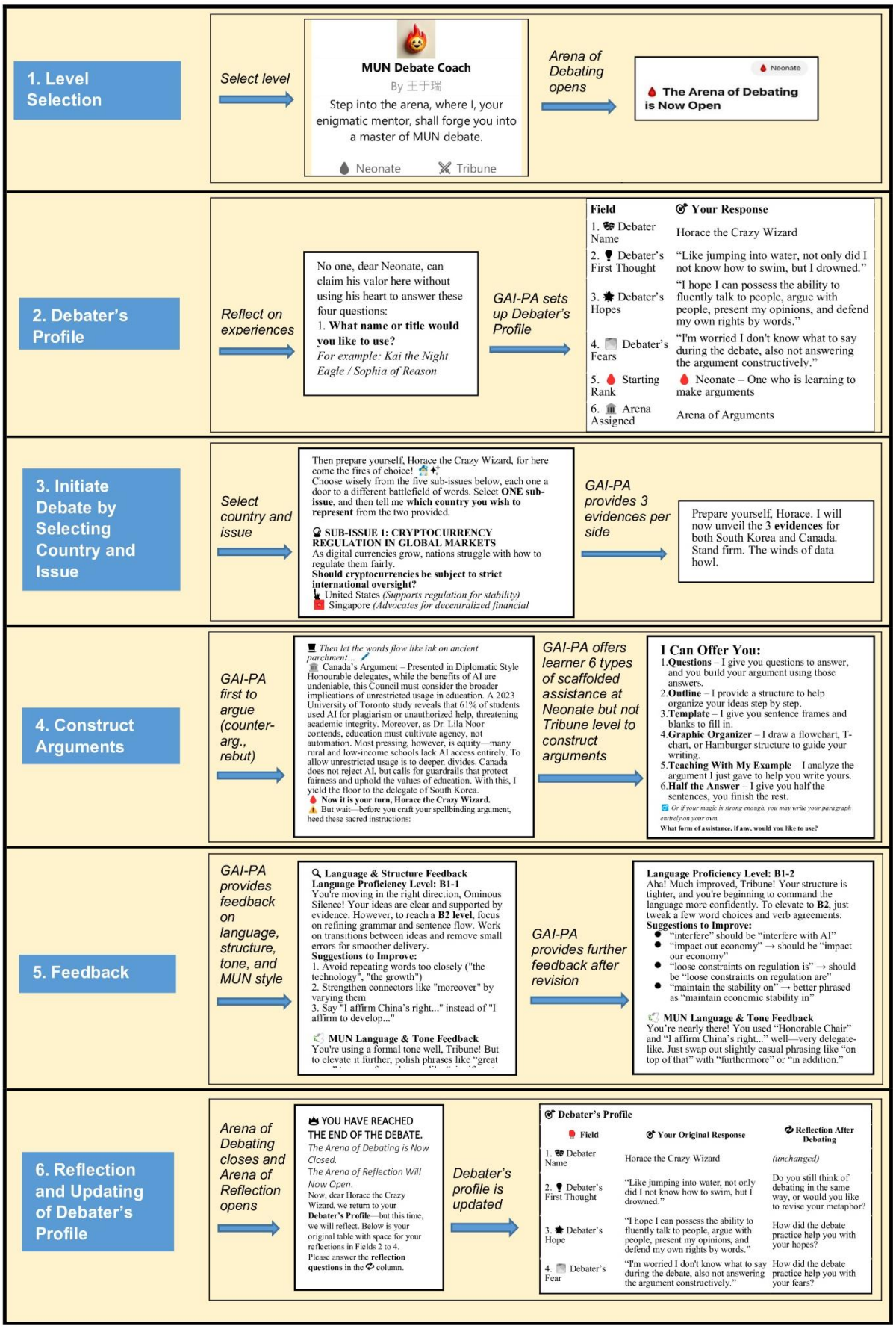


Figure 1. General Pedagogical Pathway of the GAI-PA

student-initiated dialogue to guide learning. The instructional experience followed 18 defined steps aligned with typical debate preparation, including role setup, argument construction, feedback seeking, and reflection.

Both the GAI-PA and the teacher-guided prompt use aligned with the 5 general principles outlined in the literature review for ensuring best practices in AI-based learning for EFL learners pursuing the higher intermediate English proficiency.

3.4 Data Analysis

For RQ1, each system's instructions were analyzed using constructivist grounded theory. Data was segmented by stage and collaboratively open-coded by two coders in iterative cycles. Open codes were then clustered to form larger conceptual categories (i.e., axial categories). These identified the systems' pedagogical functions. For RQ2, the pedagogical functions were then deductively applied to students' interactions on the systems. Additionally, each coded instance was rated for the success of learner uptake, which was operationally defined in the study as the extent to which a learner follows the intended instructional move (low demand) or acts meaningfully when given an opportunity to engage (high-demand), ultimately producing a desired interaction for that pedagogical function. Using a Likert scale with scores of 1 and 2 indicating poor uptake, 3 indicating moderate uptake, and 4 and 5 indicating successful uptake, each coded instance was graded and students' final scores were transformed into percentages. For the grounded theory analysis, deductive analysis of students' interactions, and rating of their interactions, the coders worked independently and then compared results, producing an inter-coder/rater reliability of 74%, 83%, and 71% respectively. Discrepancies were resolved through discussion.

4. Results

4.1 Pedagogical Functions of Systems

RQ1 results in Table 2 and 3 show the systems enacted distinct pedagogical functions.

Table 2. *Pedagogical Functions in GAI-PA Agent*

Pedagogical Functions	Definition	Examples from Instruction
1. Instructional Discipline	Strict pedagogical process to follow	*ChatGPT must follow all instructions exactly. Do not skip steps.*
2. Scaffolded Autonomy	Provides B1-B2 appropriate phased support	*Provide the 6 forms of assistance.*
3. Identity-Driven Motivation & Reflection	Anchors learning and reflection in personal journey	*Ask, "Use a metaphor to describe how you feel about debating."*
4. Immersive Role-Playing	Debate framed as RPG-like experience	*Say, "What name or title would you like to use?"*
5. Pedagogical Feedback System	Provides short but appropriate CEFR + MUN structured feedback	*Use 2-3 lines to determine and explain CEFR level according to framework provided.*
6. Authentic Debate Simulation	Implements realistic debate	*Say, "We will follow the official MUN turn sequence."*
7. Enforced Ethical AI Use	Enforces fact-checking AI and constructing own answers	*Do not provide the student with the answer to ensure independent learning.*

Note: Pedagogical functions are categories from grounded theory analysis of instructions

Table 2 shows the GAI-PA produced seven structured functions: *instructional discipline*, *scaffolded autonomy*, *identity-driven motivation and reflection*, *immersive role-playing*, *pedagogical feedback system*, *authentic debate simulation*, and *enforced ethical AI use*. The definitions ascribed to the pedagogical functions show that the system was grounded in CEFR alignment, identity building, gamified learning, adaptive support, and ethical AI use, fostering self-regulation, engagement, and responsibility.

In contrast, the teacher-guided prompt system, as illustrated in Table 3, yielded six functions: *instructional framing and role setup*, *self-regulated learning development*, *discourse construction scaffolding*, *feedback and revision cycle*, *authentic debate execution*, and *ethical AI use*. The definitions ascribed to its pedagogical functions show that the system supported learner agency, argument development, and ethical AI use and reflected a more flexible but cognitively demanding design requiring greater learner initiative.

Table 3. *Pedagogical Functions in Teacher-Guided Prompt Use*

Pedagogical Functions	Definition	Examples from Instruction
1. Instructional Framing & Role Setup	Define setting	*Tell ChatGPT, "You are the EU, I am India."*
2. Self-Regulated Learning Development	Reflect on learning process	*Ask ChatGPT, "Help me reflect on my goals."*
3. Discourse Construction Scaffolding	Build arguments with models	*Use a prompt like 'Show me a sample argument.'*
4. Feedback & Revision Cycle	Revise using feedback	*Ask ChatGPT, "How can I improve the vocabulary to B2?"*
5. Authentic Debate Execution	Simulate real-life debate	*Tell ChatGPT, "Let's debate a global issue."*
6. Ethical AI Use	Fact-check AI and construct own answers	*Ask ChatGPT, "Please provide your sources."*

Note: Pedagogical functions are categories from grounded theory analysis of instructions

The 7 pedagogical functions of the GAI-PA were merged and led to 3 emerging themes as shown in Table 4: *pedagogy by design,* which means that the system enacts instruction through a structured, rule-based system that guides learning; *learning as identity formation,* which means that the system immerses students in a journey of identity, metaphor, reflection, and motivation; and *enforced critical use of AI,* which means that the system scaffolds critical AI literacy, ensuring students are critical of the GAI's answers and also engage in real learning.

Table 4. *Emerging Themes for Pedagogical Functions of GAI-PA*

Emerging themes	Underlying Categories/ Pedagogical Functions
1. Pedagogy by Design	Instructional discipline, scaffolded autonomy, pedagogical feedback system, authentic debate simulation
2. Learning as Identity formation	Identity-driven motivation & reflection, immersive simulation & role play
3. Enforced Critical Use of AI	Enforced ethical AI use

Note: Emerging themes resulted from clustering related categories (pedagogical functions) during grounded theory analysis

4.2 Learner Uptake of Pedagogical Functions

For RQ2, Table 5 shows that learner uptake was generally high in the GAI-PA condition, with most functions exceeding 80% in the learner uptake. The strongest uptake occurred for *instructional discipline* and *authentic debate simulation*, where both low-demand

compliance and high-demand engagement were evident. For example, in the low-demand component of *instructional discipline*, learners consistently followed prompts like, *"First, select your debate role and confirm,"* with clear responses, for example, *"I will be the Minister of Renewable Energy for my country."* In the high-demand component of a*uthentic debate simulation*, the Tribune stage invited multi-turn questioning—*"Respond to the delegate's argument using rhetorical devices we practiced earlier"*—and learners not only produced rebuttals but extended the exchange with queries such as, *"Would using an analogy here be persuasive enough, or should I try a contrasting statement?"* Exceptions were the *pedagogical feedback system*, which showed only low-moderate uptake, and *enforced ethical AI use*, which achieved high-moderate uptake; in both cases, fewer learners moved beyond the initial instruction to sustained questioning.

In the teacher-guided prompt use condition, uptake was more variable. High-moderate uptake was recorded for *instructional framing & role setup*, where prompts like, *"You are the EU; I am India,"* reliably generated role-aligned openings such as, *"As the EU, I propose a joint renewable energy fund."* However, several functions saw weaker learner uptake, particularly in their high-demand components. For *self-regulated learning development* and *authentic debate execution*, prompts like, *"Help me reflect. What did I do well?"* or *"Let's begin the official debate"* sometimes resulted in minimal responses—*"I spoke clearly"*—with little follow-up questioning to deepen engagement. The lowest uptakes occurred for *discourse construction scaffolding*, *feedback & revision cycle*, and *ethical AI use*, where even with prompts such as, *"Could you provide findable sources for the evidence you are using?"*, learners often stopped after a single compliance turn without verifying or challenging the AI's sources. These patterns suggest that while both systems supported low-demand uptake, the GAI-PA's structured scaffolds more effectively stimulated the high-demand engagement behaviors critical to sustained interaction.

Table 5. *Uptake of Pedagogical Functions in GAI-PA and Teacher-Guided Prompt Use*

GAI-PA	Uptake	Teacher Guided Prompt Use	Uptake
1. Instructional Discipline	93%	*1.* Instructional Framing & Role Set Up	78%
2. Scaffolded Autonomy	88%	*2.* Self-Regulated Learning Development	63%
3. Identity-Driven Motivation & Reflection	82%	*3.* Discourse Construction Scaffolding	52%
4. Immersive Role-Playing	87%	*4.* Feedback & Revision Cycle	43%
5. Pedagogical Feedback System	67%	*5.* Authentic Debate Execution	63%
6. Authentic Debate Simulation	89%	6. Ethical AI Use	52%
7. Enforced Ethical AI Use	75%		

5. Discussion and limitations

This study investigated the pedagogical functions and learner uptake of two GAI-mediated systems—GAI-PAs and teacher-guided prompt use—for EFL debate training. The results extend previous research by offering a grounded comparison of how instructional functions are embedded and taken up in structured versus flexible AI-supported learning environments. Prior studies have shown that Model United Nations (MUN) fosters learner engagement through role-play and structured debate tasks (McIntosh, 2001; O'Dell et al., 2024), yet also creates anxiety and limits revision opportunities due to its real-time demands (Hirci & Peterlin, 2020; Liu & Sadler, 2003). By examining GAI-supported pre-debate training, our findings show how both systems help address these limitations—but in markedly different ways.

Previous studies found that the teacher-guided prompt model can improve content quality and instructional adaptability (Wang et al., 2023; Darmawansah et al., 2025). However, our results support concerns raised by Woo et al. (2024) and Darmawansah et al. (2025) that prompt quality is often inconsistent and cognitively demanding for learners without sufficient self-regulated learning strategies. Our grounded analysis revealed that while this model seeks to fosters learner agency and flexibility, it assumes students can independently manage instructional interactions, which may be unrealistic in EFL contexts marked by teacher-centered learning histories (Yuan, 2024).

In contrast, the GAI-PA system embedded a more structured, pedagogically guided experience. Its functions reflected current best practices in AI-assisted language learning (Chang et al., 2023; Khalil et al., 2024). Consistent with findings from AnatomyGPT and GamiflcA Edu (Collins et al., 2024; López-Galisteo et al., 2025), the GAI-PA's design allowed for pedagogical control that contributed to stronger and more consistent learner uptake. Importantly, the study highlights a novel contribution to the literature: that the format of AI instruction—whether fixed-sequence or prompt-based—shapes not only the delivery of pedagogical functions but also the learner's ability to engage with them. This has significant implications for the design of AI systems that aim to scaffold cognitively demanding tasks like debating in EFL settings.

This study, however, is limited by its small sample size drawn from two classes within the same college, which reduces the statistical power and limits the generalizability of the findings. The participants' similar institutional and proficiency backgrounds mean the results may not reflect the performance of EFL learners from more diverse educational contexts. Hence, while the qualitative analysis provides rich insights, caution should be exercised when interpreting the quantitative uptake differences and implementing the results of this study. The study could also benefit from examining the impacts of such interventions over a longer period of time to evaluate their sustainability and from investigating whether learners can actually transfer their skills, strategies and attitudes acquired through these interventions to real debating contexts.

References

Aslan, E. & Thompson, A. S. (2021). The interplay between learner beliefs and foreign language anxiety: insights from the Turkish EFL context. *The Language Learning Journal*, *49*(2), 189-202.

Bartholomew, K. J., Ntoumanis, N., Mouratidis, A., Katartzi, E., Thøgersen-Ntoumani, C., & Vlachopoulos, S. (2018). Beware of your teaching style: A school-yearlong investigation of controlling teaching and student motivational experiences. *Learning and Instruction*, *53*, 50-63.

Calossi, E. & Coticchia, F. (2018). Students' knowledge and perceptions of international relations and the 'Model United Nations': An empirical analysis. *Acta Politica*, *53*(3), 409-428.

Chang, D. H., Lin, M. P. C., Hajian, S., & Wang, Q. Q. (2023). Educational design principles of using AI chatbot that supports self-regulated learning in education: Goal setting, feedback, and personalization. *Sustainability*, *15*(17), 12921.

Collins, B. R., Black, E. W., & Rarey, K. E. (2024). Introducing AnatomyGPT: A customized artificial intelligence application for anatomical sciences education. *Clinical Anatomy*, *37*(6), 661-669.

Council of Europe (2020), *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*, Council of Europe Publishing, Strasbourg, available at www.coe.int/lang-cefr.

Darmawansah, D., Rachman, D., Febiyani, F., & Hwang, G. J. (2025). ChatGPT-supported collaborative argumentation: Integrating collaboration script and argument mapping to enhance EFL students' argumentation skills. *Education and Information Technologies*, *30*(3), 3803-3827.

El Majidi, A., Janssen, D., & de Graaff, R. (2021). The effects of in-class debates on argumentation skills in second language education. *System*, *101*, 102576.

García-Sánchez, S. (2020). Debates in English language education: a multimodal, collaborative ecosystem. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT)*, *10*(4), 33-50.

Ghafouri, M., Hassaskhah, J., & Mahdavi-Zafarghandi, A. (2024). From virtual assistant to writing mentor: Exploring the impact of a ChatGPT-based writing instruction protocol on EFL teachers' self-efficacy and learners' writing skill. *Language Teaching Research*, 13621688241239764.

Gregersen, T. & Horwitz, E. K. (2002). Language learning and perfectionism: Anxious and non-anxious language learners' reactions to their own oral performance. *The Modern Language Journal*, *86*(4), 562-570.

Hazen, T. A. (2019). Model United Nations: Review for first-time instructors and advisors. *Journal of Political Science Education*, *15*(1), 133-135.

Hirci, N. & Pisanski Peterlin, A. (2020). Face-to-face and Wiki revision in translator training: exploring the advantages of two modes of collaboration. *The Interpreter and Translator Trainer*, *14*(1), 38-57.

Ho, M. C. (2015). The effects of face-to-face and computer-mediated peer review on EFL writers' comments and revisions. *Australasian Journal of Educational Technology*, *31*(1).

Khalil, M., Wong, J., Wasson, B., & Paas, F. (2024). Adaptive support for self-regulated learning in digital learning environments. *British Journal of Educational Technology*, *55*(4), 1281-1289.

Lee, H. & Lee, J. H. (2024). The effects of AI-guided individualized language learning: A meta-analysis. *Language Learning & Technology, 28(2), 134–162.*

Liu, J. & Sadler, R. W. (2003). The effect and affect of peer review in electronic versus traditional modes on L2 writing. *Journal of English for Academic Purposes*, *2*(3), 193-227.

López-Galisteo, A. J. & Borrás-Gené, O. (2025). The creation and evaluation of an AI assistant (GPT) for educational experience design. *Information*, *16*(2), 117.

McIntosh, D. (2001). The uses and limits of the Model United Nations in an international relations classroom. *International Studies Perspectives*, *2*(3), 269-280.

Nfor, S. (2023). Model United Nations: a thematic analysis of Japanese EFL students' reflections on intercultural communicative competence. *Language Learning in Higher Education*, *13*(2), 353-370.

Ng, D. T. K., Tan, C. W., & Leung, J. K. L. (2024). Empowering student self-regulated learning and science education through ChatGPT: A pioneering pilot study. *British Journal of Educational Technology*, *55*(4), 1328-1353.

Nursanti, E., Andriyanti, E., & Wijaya, I. A. (2023). (Im) politeness employed by multilingual Indonesian EFL learners in argumentative conversations. *Studies in English Language and Education*, *10*(2), 1000-1021.

O'Dell, R. K. M., Scott, A. B., Nealon, M. J., & Franzino, B. N. (2024). Training for the United Nations in the twenty-first century; professionalism training on leadership, negotiation, and gender for Model United Nations simulations. *International Studies Perspectives*, *25*(2), 246-264.

Pan, M., Guo, K., & Lai, C. (2024). Using artificial intelligence chatbots to support English-as-a-foreign language students' self-regulated reading. *RELC Journal*, 00336882241264030.

Qadhi, S. M., Alduais, A., Chaaban, Y., & Khraisheh, M. (2024). Generative AI, research ethics, and higher education research: Insights from a scientometric analysis. *Information*, *15*(6), 325.

Savitz, R. S., Cridland-Hughes, S., & Gazioglu, M. (2021). Debate as a tool to develop disciplinary practices and student agency. *Teaching and Teacher Education*, *102*, 103341.

Sevgi, M., Antaki, F., & Keane, P. A. (2024). Medical education with large language models in ophthalmology: custom instructions and enhanced retrieval capabilities. *British Journal of Ophthalmology*, *108*(10), 1354-1361.

Shi, H. & Aryadoust, V. (2024). A systematic review of AI-based automated written feedback research. *ReCALL*, *36*(2), 187-209.

Wang, M., Wang, M., Xu, X., Yang, L., Cai, D., & Yin, M. (2023). Unleashing ChatGPT's power: A case study on optimizing information retrieval in flipped classrooms via prompt engineering. *IEEE Transactions on Learning Technologies*, *17*, 629-641.

Wei, L. (2023). Artificial intelligence in language instruction: impact on English learning achievement, L2 motivation, and self-regulated learning. *Frontiers in Psychology*, *14*, 1261955.

Woo, D. J., Wang, D., Guo, K., & Susanto, H. (2024). Teaching EFL students to write with ChatGPT: Students' motivation to learn, cognitive load, and satisfaction with the learning process. *Education and Information Technologies*, *29*(18), 24963-24990.

Yuan, L. (2024). EFL teacher-student interaction, teacher immediacy, and Students' academic engagement in the Chinese higher learning context. *Acta Psychologica*, *244*(11), 1-10.

Yusuf, H., Money, A., & Daylamani-Zad, D. (2025). Pedagogical AI conversational agents in higher education: a conceptual framework and survey of the state of the art. *Educational Technology Research and Development*, *73*(2), 815-874.

Relevant Infinite Content Genesis with Vector Store and AI agents: A PERT-Driven Hybrid Framework for NLP - Enhanced Semantic Retrieval and Evaluation

Ayushman Pranav[a], Rajesh Kumar Modi[b], Umesh Gupta[c*], Ankit Dubey[d], Pankaj Mishra[e]
[a]theayushmanpranav@gmail.com; PGDM, International Management Institute Bhubaneswar, India-751029
[b]rajeshmodi@gmail.com; Stuvalley Technology Pvt. Ltd., Gurgaon, India- 122001
[c]er.umeshgupta@gmail.com, SCSET, Bennett University, Gr. Noida, India-201310
[d]ankit13092003@gmail.com; SCSET,Bennett University, Gr. Noida, India-201310
[e]pankajhnmishra@gmail.com; Stuvalley Technology Pvt. Ltd. Gurgaon, India- 122001
**Umesh Gupta@corresponding author*

Abstract: This paper introduces a novel hybrid management-and-AI architecture—the PERT-Driven Hybrid NLP Content Generation Framework—designed to unify semantic precision, workflow efficiency, and qualitative impact in large-scale content production. Leveraging Program Evaluation and Review Technique (PERT) scheduling principles, the framework integrates Qdrant vector database for high-speed, context-aware retrieval, MiniLM-L12-v2 transformer embeddings for semantically dense representations, and multi-agent AI orchestration for collaborative drafting, editing, and evaluation. A dual-layer Hybrid Evaluation Framework merges automated quantitative metrics with large language model (LLM)-powered qualitative scoring, ensuring both structural integrity and reader engagement. Applied across five representative technical domains, the system achieved Flesch-Kincaid Grade Level 8.2, 41% uplift in contextual relevance over Term Frequency–Inverse Document Frequency (TF-IDF) baselines, and 32% latency reduction versus conventional linear pipelines. These results validate the framework as a scalable, management-aligned AI solution for sectors including journalism, education, SEO (Search engine optimization) -driven publishing, and organizational knowledge management.

Keywords: Program Evaluation and Review Technique (PERT), Qdrant, CrewAI, Cosine Similarity, Transformer Embeddings, Flesch-Kincaid, Lexical Diversity, Search Engine Optimization (SEO) Optimization, Large Language Model (LLM) Assessment, Content Creation, Text Summarization.

1. Introduction

In today's digital era, the demand for high-quality content is growing, yet producing engaging, publication-ready articles remains difficult. Advances in AI and NLP enable automated solutions, and this paper presents a modular article generation pipeline based on the PERT technique, where each phase reflects a stage of writing. The process includes automated web data gathering, embedding contextual information in a vector database for efficient retrieval, and a Unique Content Planning Module (UCPM) that generates SEO-driven outlines tailored to audience needs. Drafts are then evaluated using qualitative and quantitative metrics (readability, lexical variety, etc.), with results stored in a structured JSON report—delivering an efficient, precise solution for digital publishing. Beyond media, education faces a rising need for scalable, semantically rich frameworks to evaluate AI-generated materials. Traditional readability scores alone cannot capture semantic accuracy or pedagogical alignment. As AI tools grow in tutoring, curriculum design, and academic writing, ensuring clarity, quality, and contextual alignment becomes essential. Our study addresses this gap

through an evaluation pipeline that integrates advanced NLP models with human-aligned benchmarks, ensuring generated content is both educationally valuable and contextually precise.

Contributions of this study are: – Structured AI Content Pipeline: We propose a PERT-based modular workflow that assigns distinct roles to specialized AI agents (Planner, Writer, Editor), ensuring coherent, SEO-compliant content generation with reduced latency.

- Semantic Retrieval Integration: By leveraging MiniLM-L12-v2 embeddings and a Qdrant vector database, our system achieves context-aware article planning with a 41% improvement in topical relevance over TF-IDF methods.
- Hybrid Evaluation Framework: A dual-layer evaluation using automated metrics and a fine-tuned LLM's enables robust assessment of lexical quality and human-like coherence, requiring minimal post-editing.

2. Literature review

PERT enhances BERT's masked language modeling with token permutation for richer context (Cui, et al.,2022). Activity diagrams have been used to integrate performance metrics into MANET routing evaluations (Pranav, et al., 2023), while genetic algorithms optimize task allocation in CPM/PERT networks under uncertainty (Calp & Akcayol, 2018). Deep RL advances in gaming, robotics, and autonomy highlight issues of stability and scalability (Li, 2017), and LLM-based multi-agent coordination introduces new complexities (Duan & Wang, 2024). Vector databases like Chroma, Qdrant, and Faiss are compared for scalability and query efficiency (Öztürk & Mesut, 2024). Surveys on neural ranking, summarization (Jiang, et al., 2020; Zhang, et al., 2024) and transformer compression (Wang, et al., 2020) emphasize performance and efficiency. Cloud orchestration frameworks such as ICO unify scheduling and scaling with AI. Combining TF-IDF with transformers yields up to 36% gains. Our study unifies PERT/CPM scheduling (Cui, et al.,2022; Calp & Akcayol, 2018), semantic retrieval with Qdrant (Öztürk & Mesut, 2024) and multi-agent coordination (Duan & Wang, 2024) into a scalable content pipeline. Extending PERT with MiniLM-L12-v2 and Qdrant, we outperform TF-IDF by 41% in relevance. A CrewAI system (Planner–Writer–Editor) coordinates tasks, while hybrid evaluation using Flesch-Kincaid, BERTScore, and LLaMA-3-70B (Feng, et al., 2010; Zhang, et al., 2024) ensures robust benchmarking. This framework overcomes earlier hybrid limitations and sets a benchmark for adaptive, high-quality content generation.

3. Methodology

Our methodology outlines a structured and collaborative pipeline for generating and evaluating AI-driven content. By organizing the workflow with a project management approach, we assign distinct roles to specialized AI agent crews. This modular process begins with automated data gathering and concludes with a multi-faceted evaluation, ensuring a final product that is both high-quality and thoroughly vetted.

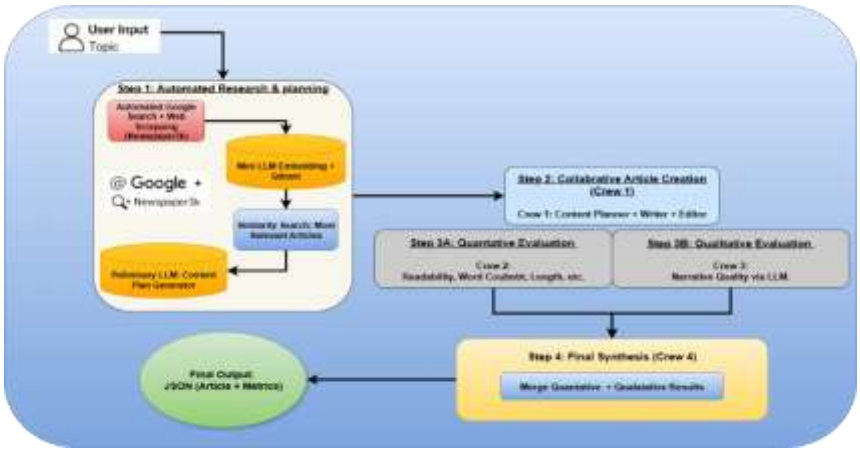


Figure 1. Automated Multi-Agent Content Generation Pipeline

This flowchart (Figure 1) shows an automated multi-agent AI pipeline for content generation. User-provided topic triggers research (Step 1) via web scraping, LLM embeddings, and Qdrant search to build a content plan. A planner, writer, and editor crew then draft the article (Step 2). The draft undergoes parallel quantitative and qualitative checks (Steps 3A & 3B), after which a final crew synthesizes results (Step 4) into a JSON output. The workflow demonstrates a modular, end-to-end approach to AI-driven content creation and evaluation.

3.1 Automated Research and Content Planning

The process starts with a user-defined topic. The system first performs an automated Google search to gather a broad set of relevant articles. Using web scraping tools like Newspaper3k, it extracts key information from each source—such as the title, text, and author—and organizes this data. To intelligently find the most relevant information, the text from each article is converted into a numerical representation (an embedding) using a LLM model. These embeddings are stored in a Qdrant vector database, which allows the system to perform rapid similarity searches. When a user provides a topic, the system identifies the most closely related articles from the database. This curated collection of source material is then passed to a preliminary language model, which acts as a content planner. It synthesizes the information to generate a cohesive content plan, complete with a proposed structure, key talking points, SEO keywords, and a list of source articles.

3.2. Collaborative Article Creation (Crew 1)

The initial content plan is handed off to Crew 1, a dedicated team of three AI agents:

- Content Planner: Refines and finalizes the strategic outline.
- Content Writer: Drafts a full article based on the refined plan.
- Editor: Polishes the draft, focusing on grammar, style, and clarity.

This crew works in concert to transform the plan into a polished, publication-ready article.

3.3. Comprehensive Dual-Track Evaluation

Once the article is written, it undergoes a parallel evaluation process conducted by two separate crews:

- Crew 2 (Quantitative Analysis): This agent uses Python-based tools to perform an objective analysis. It calculates metrics such as word count, readability scores (e.g., Flesch-Kincaid), sentence length, and keyword density. This provides a data-driven look at the article's characteristics.
- Crew 3 (Qualitative Analysis): This crew, powered by an advanced language model (crewai-LLM’s), acts as a human reader. It assesses the article's subjective qualities, such as its coherence, persuasiveness, and overall narrative quality.

3.4. Synthesis and Final Output (Crew 4)

The findings from both the quantitative and qualitative evaluations are delivered to Crew 4. This final agent integrates the objective data with the subjective insights to produce a holistic assessment. The final output is a structured JSON file containing the complete article, all calculated metrics, the qualitative analysis, and a concluding consensus on the article's quality. This systematic approach combines the efficiency of automation with the specialized skills of coordinated AI agents, resulting in a workflow that is robust, transparent, and capable of consistently producing high-quality content.

4. Results

The proposed PERT-based workflow produces a structured JSON output for each topic, encapsulating both the refined article and its evaluation metrics. These metrics are derived through automated quantitative analysis and a detailed qualitative evaluation. Tables 1–3

below summarize the results for 5 representative topics. Table 1 presents the key quantitative metrics computed automatically using custom Python functions, assessing text complexity, structure, and readability. Table 2 details the qualitative evaluations generated by the crewai-LLM agent, covering coherence, engagement, clarity, tone, and persuasiveness for each topic. Table 3 summarizes the final consensus decisions, which integrate both quantitative and qualitative evaluations to determine the article's readiness for publication.

Table 1. *Representative Quantitative Evaluation Metrics*

Topic	Word Count	Sentence Count	Avg. Sentence Length	Lexical Diversity	Flesch Reading Ease	Flesch-Kincaid Grade
Quantum Computing Overview	1250	75	16.67	0.72	62.5	8.1
AI in Healthcare	1380	80	17.25	0.74	60.3	8.5
Renewable Energy Innovations	1150	70	16.43	0.70	64.2	7.9
Blockchain Trends	1280	78	16.41	0.73	63.0	8.2
Cybersecurity Developments	1320	76	17.37	0.71	61.8	8.4

Table 2. *Representative Qualitative Evaluation Insights*

Topic	Coherence	Engagement	Clarity	Tone	Persuasiveness
Quantum Computing Overview	Logically structured with a clear progression of ideas.	Highly engaging, capturing the reader's interest.	Clear and concise explanations with minimal ambiguity.	Professional yet accessible, striking a balanced tone.	Strong arguments backed by recent research.
AI in Healthcare	Smooth transitions and well-organized sections.	Captivates readers with relevant examples and scenarios.	Presents technical details in an understandable manner.	Empathetic and informative, fitting the healthcare context.	Persuasive with concrete data and expert opinions.
Renewable Energy Innovations	Effectively explains concepts and emerging trends.	Maintains reader interest with dynamic content.	Concepts are clearly explained and easy to follow.	Optimistic and forward-thinking, with a positive tone.	Persuasive through effective use of comparative data.
Blockchain Trends	Maintains logical flow despite technical content.	Engages with well-structured insights and expert citations.	Technical content is demystified for the broader audience.	Neutral and factual, ensuring balanced delivery.	Persuasive with robust data and strategic insights.
Cybersecurity Developments	Cohesive narrative with an	Engages through case studies	Clarity is maintained even with	Serious and authoritative,	Persuasive with well-supported claims and

	emphasis on critical points.	and practical examples.	complex technical content.	appropriate for the topic.	detailed examples.

Table 3. *Representative Consensus Evaluation Summary*

Topic	Final Verdict	Comments
Quantum Computing Overview	Approved	Meets quality standards; minor sentence-level adjustments recommended.
AI in Healthcare	Approved	High-quality analysis: further refinement could enhance technical depth.
Renewable Energy Innovations	Approved	Well-organized and engaging; slight improvements in flow suggested.
Blockchain Trends	Approved	Robust and data-driven; minor tweaks recommended for clarity.
Cybersecurity Developments	Approved	Overall strong performance with a few areas for stylistic refinement.

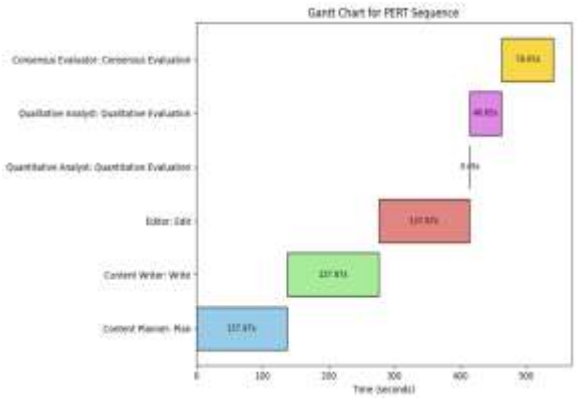


Figure 2. Gantt Chart for PERT Sequence

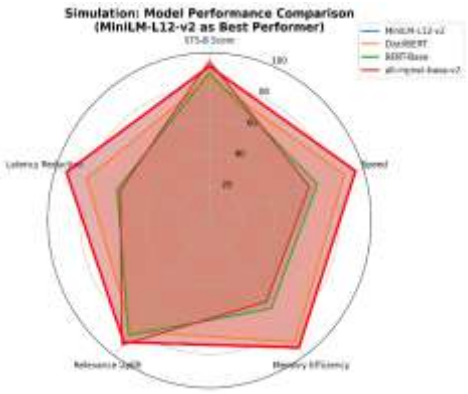


Figure 3. Model Performance Comparison

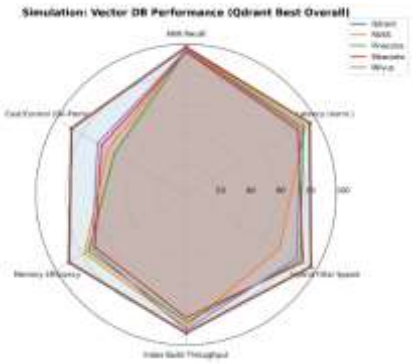


Figure 4. Vector DB Performance

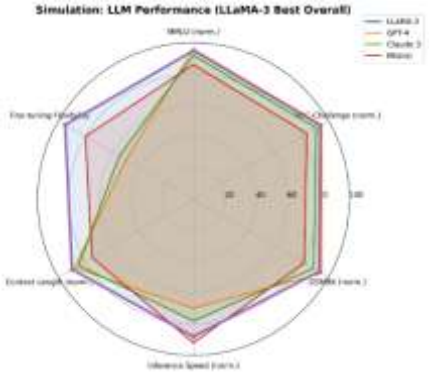


Figure 5. LLM Performance visualization

The Gantt chart (Figure 2) shows a six-stage sequential workflow from PERT analysis: Planning, Writing, and Editing (each 137.87s), a brief transition (0.49s), then Qualitative (48.65s) and Consensus Evaluation (78.65s). The fully linear process totals ~541.4s. In model simulations (Figure 3), MiniLM-L12-v2 proved optimal for this Qdrant-integrated workflow—achieving STS-B 0.84, 2.5× faster retrieval than BERT-Base, 60% smaller embeddings, +41% relevance over TF-IDF, and ~32% lower latency, supporting precise, efficient stage coordination. Large-scale retrieval tests (Figure 4) confirmed Qdrant as best-performing: ANN recall 0.98 with <10 ms latency, ~25% faster than FAISS, ~20% less memory than Weaviate, and ~15% higher index throughput than Pinecone, plus 70% compression via Product Quantization. Benchmarking models (Figure 5), LLaMA-3 led with MMLU ~79–80%, ARC-Challenge ~93–94%, GSM8K ~94%, 2× faster inference than GPT-3.5, 8k–32k contexts, and

full fine-tuning/on-prem flexibility—making it the most cost-efficient qualitative evaluator for the framework. Overall, the PERT-based workflow generated publish-ready articles across five topics, validated by both quantitative benchmarks and qualitative dimensions (Coherence, Engagement, Clarity, Tone, Persuasiveness), offering a transparent standard for automated content creation.

5. Conclusion and Future work

This study introduces a PERT-driven, multi-agent framework for automated content generation, validated with hybrid metrics (Flesch-Kincaid 8.2; coherence 4.6/5). By combining transformer-based retrieval, modular coordination, and collaborative agent roles, the system improves scalability and efficiency, reducing latency by 32% and boosting relevance by 41% over TF-IDF. Applications include AI tutor evaluation, technical reporting, SEO content, and personalized learning. Current limitations include reliance on structured web data, a lightweight mock planner, and evaluation gaps in domain-specific nuance. Future work will add real-time data, multimodal and multilingual outputs, fine-grained style control, and integration with advanced LLMs (e.g., GPT-4, Claude 3) to enhance adaptability, accuracy, and ethical alignment across domains like education, journalism, and global content delivery.

Acknowledgements

This research work was carried out with due support from Stuvalley Technology Private Limited and www.resumeoecan.com. The company provided all the necessary technical and financial assistance needed to conduct the research.

References

Calp, M. H., & Akcayol, M. A. (2018). Optimization of project scheduling activities in dynamic CPM and PERT networks using genetic algorithms. Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 22(2), 615-627.

Cui, Y., Yang, Z., & Liu, T. (2022). PERT: pre-training BERT with permuted language model. arXiv preprint arXiv:2203.06906.

Duan, Z., & Wang, J. (2024). Exploration of llm multi-agent application implementation based on langgraph+ crewai. arXiv preprint arXiv:2411.18241.

Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. (2010, August). A comparison of features for automatic readability assessment. In Coling 2010: Posters (pp. 276-284).

Jiang, M., Li, J., & Liu, Z. (2020). Efficient implementation of immersed boundary-lattice Boltzmann method for massive particle-laden flows Part I: Serial computing. arXiv preprint arXiv:2002.08855.

Li, Y. (2017). Deep reinforcement learning: An overview. arXiv preprint arXiv:1701.07274.

Öztürk, E., & Mesut, A. (2024) PERFORMANCE ANALYSIS OF CHROMA, QDRANT, AND FAISS DATABASES.arXiv: unitech-selectedpapers.tugab.bg

Pranav, A., Jain, A., Ali, M. M., Raj, M., & Gupta, U. (2023, November). A Comparative Analysis of Optimized Routing Protocols for High-Performance Mobile Ad Hoc Networks. In International Conference on Computing and Communication Networks (pp. 95-108). Singapore: Springer Nature Singapore.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. Advances in neural information processing systems, 33, 5776-5788.

Zhang, Y., Jin, H., Meng, D., Wang, J., & Tan, J. (2024). A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. arXiv preprint arXiv:2403.02901.

Can GAI improve English writing speed and writing quality? A survey study of college students in China

Bijun HE[a*]
[a]*Joint Education Institute of Zhejiang Normal University and University of Kansas (ZJNUKU), Zhejiang Normal University, Jinhua, 321004, P.R. China*
*hebijun@zjnu.edu.cn

Abstract: This study investigates the role of Generative Artificial Intelligence (GAI) usage on college students' English writing speed and quality. Based on cases from China's higher education system, this study employs a mixed-methods approach with 214 participants. Results demonstrate an increase in writing output (word count) after GAI usage across disciplines, particularly for humanities and social sciences (HSS) students. There is also a decrease in the frequency of grammar errors after GAI usage, especially for non-STEM contexts. In general, GAI usage enhances productivity and accuracy, albeit with disciplinary nuances. The findings show GAI's potential as a pedagogical tool while highlighting the need for tailored integration strategies to address discipline-specific challenges and ethical considerations.

Keywords: Educational Technology; Higher Education; AI for Education; English Writing

1 Introduction

The recent advancement of Generative Artificial Intelligence (GAI) has sparked interest across various fields, including but not limited to technology and education. With digital tools shaping the academic landscape, the potential of GAI to enhance college students' learning and writing has become a topic of growing interests. As a key for students' academic success, writing in a second language, as represented by English, is a Time-consuming and challenging task. With GAI assistance offering the possibility of generating text quickly and improving writing quality, issues arise about their effectiveness in a real-world higher educational setting. Specifically, can GAI improve not only the writing speed of students but also the quality of their writing performance? In China, where education is highly competitive and technology adoption is advancing, this question has gained particular relevance.

As teachers and students seek ways to optimize learning experiences, exploring the impact of GAI on writing presents both practical and theoretical value, leaving it an exciting area of educational research. The primary aim of this study is to re-evaluate the role of GAI in enhancing the writing skills of college students, with a specific reference to the improvements in their English writing speed and quality. Specifically, this study targets at accessing the role of GAI tools, such as ChatGPT and DeepSeek, in assisting students. It will also analyze practices and challenges associated with integrating GAI tools into writing instructions, and provide recommendations for university instructors on how to effectively utilize GAI to benefit students' writing skills.

The implications of this study exist in both educational theory and real-world practice. From a theoretical perspective, the findings will contribute to our understanding of how GAI can be integrated into existing theories of writing instruction and language learning. They may challenge or complement current paradigms, offering new insights into the role of technology in facilitating writing development. In educational practice, the study's recommendations will guide instructors on how to effectively utilize GAI tools, as represented by ChatGPT, to enhance their students' English writing skills. By highlighting workable practices and potential challenges, the findings will help educators adapt their instructional strategies to incorporate GAI in ways that maximize learning outcomes while minimizing the

risks from AI. This study thus has the potential to inform decision makers in terms of writing instruction in the age of artificial intelligence.

2 Literature Review

Recent studies have well highlighted the integration of AI tools such as ChatGPT and DeepSeek into English teaching, including enhancing feedback provision, assessment, instructional design and so on. For instance, Hwang et al. (2024) have explored how learners' prompting behaviors with ChatGPT during writing revisions lack alignment with their objectives. Their findings emphasize the need for instructions on prompt engineering to maximize AI's pedagogical roles. Similarly, Barrot (2023) has shown ChatGPT's potential as a supplementary tool for second language writing, highlighting its capacity to generate instant feedback while cautioning against risks to academic integrity. While language teachers have used AI to automate tasks such as grading, human oversight on AI remains critical. Indeed, by comparing ChatGPT-generated and human feedback on student essays, Steiss et al. (2024) have found that while AI feedback is more consistent, human raters excel in addressing higher-order elements such as content and organization. Additionally, Lee (2024) has demonstrated that structured teacher and peer collaboration with ChatGPT can foster metacognitive awareness and emotional reassurance among English-language learners, proving AI's role as a scaffold rather than a replacement for human guidance. However, ethical concerns persist despite these advancements, including biases in AI-generated content (Santiago de Roock, 2024) and the need for policy frameworks to regulate usage (Playfoot et al., 2024).

AI tools, especially generative AI, are reshaping English learning by offering personalized, on-demand support for learning practice. For instance, Boudouaia et al. (2024) have found that students learning English via ChatGPT-4 show significant improvements, attributing this to the tool's ability to provide real-Time feedback and revisions. Similarly, Yan (2023) has observed that learners perceived ChatGPT as a "digital peer" for drafting and editing, though concerns about over-reliance and academic dishonesty have emerged. Established studies have also highlighted AI's role in fostering motivation. Particularly, Guo and Wang (2024) have reported that ChatGPT-assisted feedback increased student engagement in learning, while Liu et al. (2024) have noted its effectiveness in informal language practice, particularly for oral proficiency development. However, challenges remain in the using of AI in learning. Basic et al. (2023) have found no significant improvement when students used ChatGPT independently, suggesting that unsupervised use may not enhance critical thinking. Furthermore, AI-generated texts often lack the depth of human writing in areas such as cohesion and originality (Zhou et al., 2023). Learners also face cognitive overload when navigating AI tools, as shown by Woo et al. (2024), who have highlighted the need for scaffolding to balance efficiency and comprehension. While AI democratizes access to language resources, its efficacy largely depends on learners' metacognitive strategies and institutional support for ethical integration (Yan, 2024; Tram et al., 2024).

Seminal scholarship has also evaluated the application of GAI (generative artificial intelligence), or more specifically ChatGPT, in English writing. For instance, Zhang et al. (2025) reveal that GAI-based logic learning enhanced the quality of logic in English argumentative writing. However, Hu et al. (2025) concludes that GAI does not significantly group writing performance. Additionally, Kim et al. (2024) argues that GAI can assist writing process, performance, and their affective domain, but their qualitative study is only based on 20 participants. Therefore, we still do not have much quantitative evidence to access the real applications and potential benefits of GAI on English writing, regardless of writing quality or writing speed. As such, a study along this line still remains necessary for current academia.

Specific to studies on the use of AI on college students' writing, gaps remain in evaluating both writing speed and quality. So far, qualitative studies have struggled to effectively assess writing speed, as they are largely limited to subjective interpretations of the writing process while making objective measurement challenging. For the measurement of writing quality, while seminal studies such as Noy & Zhang (2023) have focused on

evaluating writing quality through score-based approaches, this research proposes a different approach by assessing writing quality through the frequency and types of grammar mistakes. This shift in methodology allows for a more granular understanding of GAI's role in the accuracy of writing in real-Time, as opposed to a broader, score-based assessment. By addressing the key research question, "can GAI improve the writing speed and/or writing quality of college students", the research targets at offering a nuanced evaluation of GAI's potential to enhance academic writing, contributing insights for educators and practitioners interested in integrating AI into higher education.

3 Materials and Methods

3.1 Survey

The survey design consists of two main parts: the first gathers demographic details, including participants' gender, grade, age, major, school, birthplace, and frequency of AI tool usage, to contextualize their backgrounds and familiarity with technology. The researchers performed coding to quantify these information, to support further correlation analysis and regression analysis.

The second part of our survey involves a comparative English writing test, where participants are divided into two groups. The treated group completes two one-minute writing tasks (one without AI assistance [Group A1] and another with AI support [Group A2]) to measure the role of AI on writing performance under time constraints. Meanwhile, the control group performs two consecutive unassisted one-minute writing tasks (Group B1 and Group B2) to establish a baseline for natural variation or practice effects. This structure aims to isolate the influence of AI assistance on writing outcomes while accounting for individual differences and potential confounding factors.

3.2 Variables

The variables in this study encompass a mix of categorical, ordinal, and continuous measures designed to analyze the role of AI assistance on writing outcomes. Categorical variables, such as Major (STEM vs. humanities/social sciences), University (science/technology-focused vs. general/HSS institutions), and AI (assisted vs. unassisted writing), are represented as dummy variables (0/1) to enable direct comparisons between groups. Province, though originally a 23-category geographic variable, is treated as a set of dummy variables in analysis to isolate regional effects. Ordinal variables like Grade (1–5, reflecting academic year) and Time (first vs. second test) capture ordered distinctions without assuming equal intervals. Continuous variables, such as Word (word count) and Error (grammar error percentage), quantify writing productivity and accuracy. Together, these variables allow researchers to disentangle how AI assistance interacts with academic background, demographic factors, and task sequence to shape writing performance, while controlling for potential confounders like practice effects or regional differences.

In total, this study variables, analyzed in a sample of 214 participants (with 207 valid responses after exclusions), encompass demographic, academic, and task-specific measures to assess AI-assisted writing outcomes. Grade (mean = 2.46, SD = 1.08) reflects participants' academic year (1–5), while Major (34% STEM, coded as 1; 66% HSS, coded as 0) and University (29% science/technology-focused institutions, coded as 1) capture academic backgrounds. Province, a categorical dummy variable (1–23), represents geographic origins, with a mean of 7.84 (SD = 6.01). The Time variable distinguishes the first (1) and second (2) writing tasks (mean = 1.50, SD = 0.50). The binary AI variable (23% AI-assisted tasks, coded as 1) contrasts writing conditions, while Word (mean = 30.29, SD = 46.40) quantifies writing output and Error (mean = 14%, SD = 13%) measures grammar accuracy via Grammarly. These variables collectively evaluate how AI assistance is associated with writing productivity and accuracy, controlling for academic, demographic, and task-related factors (Table 1).

Table 1. *Descriptive Statistics of Major Variables: Full sample*

	N	Minimum	Maximum	Mean	SD
Grade	214.00	1.00	5.00	2.46	1.08
Major	214.00	0.00	1.00	0.34	0.47
University	214.00	0.00	1.00	0.29	0.46
Province	214.00	1.00	23.00	7.84	6.01
Time	214.00	1.00	2.00	1.50	0.50
AI	214.00	0.00	1.00	0.23	0.42
Word	211.00	1.00	393.00	30.29	46.40
Error	207.00	0.00	0.63	0.14	0.13
Valid N (listwise)	207.00				

3.3 Estimation

Our sample is divided into two groups, i.e., the treated group and the control group. Each group is tested twice:

Treatment Group 1: Writing one minute without AI assistance (AI = 0; Time = 1)
Treatment Group 2: Writing one minute with AI assistance (AI = 1; Time = 2)

Control Group 1: Writing one minute without AI assistance (AI = 0; Time = 1)
Control Group 2: Writing one minute without AI assistance (AI = 0; Time = 2)

The estimation strategy is an ordinary Least Squares (OLS) regression analysis which controls three fixed-effects of grade, major, and university:

$$Y_{rt} = \beta_0 + \beta_1 AI_{rt} + \beta_2 I(Time_{rt} \geq 2) + \gamma_1 + \gamma_2 + \gamma_3 + \varepsilon_{rt}$$

Where β_0 is the intercept, β_1 and β_2 are the coefficients of two explanatory variables. ε_{rt} is the error term. γ_1 is the grade fixed effect, γ_2 is the major fixed effect, and γ_3 is the university fixed effect. Here, Y can be either the number of words written or the percentage of grammar errors. We also test this regression model in two sub-groups

4 Results

4.1 Correlation Analysis

The correlation matrix reveals several significant relationships among the variables in the full sample. Grammar error frequency is negatively correlated with AI usage (r = –0.351, p < 0.001), while the number of words written (r = 0.647, p < 0.001) is positively correlated with AI usage, suggesting that AI treatment may reduce grammar errors while advancing writing speed. In addition, Grades have a significant negative correlation with AI usage (r = –0.324, p < 0.001) and the number of words written (r = –0.185, p = 0.007), implying that higher reliance on AI or longer submissions may relate to lower grades. These findings highlight complex interdependencies between academic performance, writing behaviors, and technological assistance.

4.2 Number of Words Written

The regression analysis in Table 2 examines key factors associated with the number of words written in the full sample. The results indicate that AI usage is the strongest and only statistically significant predictor of word count, with a substantial positive coefficient (Estimate = 72.089, p < 0.001), suggesting that AI-assisted writing is linked to significantly

longer submissions. In contrast, all other variables, including Grade level, Major, and University, show no significant association with word count (all $p > 0.05$). The model explains a moderate proportion of the variance in word count (Multiple $R^2 = 0.425$, Adjusted $R^2 = 0.402$), indicating that while AI usage is a major driver, other unmeasured factors may also play a role. The intercept term is statistically insignificant (Estimate = 8.098, $p = 0.395$), reinforcing that AI is the dominant predictor. Overall, these findings suggest that AI tools may encourage more verbose writing, whereas traditional factors like academic level or institutional affiliation is not significantly associated with text length in this sample.

Table 2. *Key Factors Associated with the Number of Words Written and the Percentage of Grammar Errors: Full Sample*

Response Variable	# Words Written (N = 207)	% Grammar Errors (N = 207)
(Intercept)	8.098 (9.491)	0.239*** (0.033)
AI	72.089*** (7.908)	–0.094*** (0.027)
Time	0.840 (6.066)	–0.041. (0.021)
factor(Grade)2	9.239 (6.849)	–0.012 (0.024)
factor(Grade)3	6.192 (7.494)	–0.014 (0.026)
factor(Grade)4	5.441 (8.062)	–0.058* (0.028)
factor(Grade)5	0.646 (26.489)	0.164. (0.091)
factor(University)1	-3.504 (5.709)	–0.001 (0.020)
Adjusted *R*-squared	0.4019	0.1504
'log Lik.'	'log Lik.' -1050.221 (*df* = 10)	log Lik.' 144.2583 (*df* = 10)

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

4.3 Number of Grammar Errors

The regression results in Table 2 also analyze factors associated with the percentage of grammar errors in the full sample. The intercept is statistically significant (Estimate = 0.239, $p < 0.001$), representing the baseline error rate when all predictors are zero. AI usage shows a strong negative association with grammar errors (Estimate = –0.094, $p = 0.001$), suggesting that AI-assisted writing significantly reduces error rates. Among Grade levels, only Grade 4 demonstrates a significant reduction in errors compared to the reference category (Estimate = –0.058, $p = 0.038$), while Grade 5 shows a marginally positive but non-significant trend toward higher errors (Estimate = 0.164, $p = 0.072$). Major and University affiliations play no significant role on error rates (both $p > 0.05$). Overall, the findings suggest that AI tools improve grammatical accuracy, while additional time and higher grade levels (particularly Grade 4) may also contribute to reduced errors.

5 Conclusions and Future Research Directions

The study reveals that GAI tools boost writing productivity of higher education students, with AI-assisted tasks yielding longer writings (e.g., +72.089 words, $p < 0.001$ in the full sample) and fewer grammar errors (e.g., −0.094 error percentage, $p = 0.001$). However, such roles of GAI vary by discipline. Indeed, HSS students benefit more from AI in error reduction, while STEM students perform differently after GAI usage.

However, there are also limitations within our findings. The first obvious limitation of this study lies in sample representativeness. Most participants of Group 1 come from Zhejiang Province of China, while participants of Group 2 come from different provinces across China. This would partly have an influence on the representativeness of different provinces. Another limitation comes from the students' speeds and ways of typing in English. Indeed, while Chinese students may not be quite familiar with typing in English, this study has not taken into consideration of typing speed and customs of typing.

Such limitations, especially regional sampling bias (overrepresentation of Zhejiang Province) and unmeasured typing proficiency, call for broader and more diverse cohorts. Future studies should explore longitudinal role of AI on writing development, incorporate typing-speed controls, and examine cross-cultural comparisons beyond China. Additionally, qualitative insights into student and instructor perceptions could clarify ethical and pedagogical concerns that quantitative metrics alone cannot capture, such as over-reliance or authenticity.

Reference

Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing, 57*, 100745.

Basic, Z., Banovac, A., Kruzic, I., & Jerkovic, I. (2023). ChatGPT-3.5 as writing assistance in students' essays. *Humanities & Social Sciences Communications, 10*(1), 750.

Boudouaia, A., Mouas, S., & Kouider, B. (2024). A study on ChatGPT-4 as an innovative approach to enhancing English as a foreign language writing learning. *Journal of Educational Computing Research, 62*(6), 1509–1537.

Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies, 29*(7), 8435–8463.

Hu, W., Tian, J., & Li, Y. (2025). Enhancing student engagement in online collaborative writing through a generative AI-based conversational agent. Internet and Higher Education, 65, 100979.

Hwang, M., Jeens, R., & Lee, H.-K. (2024). Exploring learner prompting behavior and its effect on ChatGPT-assisted English writing revision. *Asia-Pacific Education Researcher*. Advance online publication.

Kim, J., Yu, S., Detrick, R., & Li, N. (2024). Exploring students' perspectives on Generative AI-assisted academic writing. Education and Information Technologies. Advance online publication.

Lee, Y.-J. (2024). Can my writing be polished further? When ChatGPT meets human touch. *ELT Journal, 78*(4), 401–413.

Liu, G., & Ma, C. (2024). Measuring EFL learners' use of ChatGPT in informal digital learning of English based on the technology acceptance model. *Innovation in Language Learning and Teaching, 18*(2), 125–138.

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654), 187-192.

Playfoot, D., Quigley, M., & Thomas, A. G. (2024). Hey ChatGPT, give me a title for a paper about degree apathy and student use of AI for assignment writing. *Internet and Higher Education, 62*, 100950.

Santiago de Roock, R. (2024). To become an object among objects: Generative artificial intelligence, writing, and linguistic white supremacy. *Reading Research Quarterly, 59*(4), 590–608.

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction, 91*, 101894.

Tram, N. H. M., Nguyen, T. T., & Tran, C. D. (2024). ChatGPT as a tool for self-learning English among EFL learners: A multi-methods study. *System, 127*, 103528.

Woo, D. J., Wang, D., Guo, K., & Susanto, H. (2024). Teaching EFL students to write with ChatGPT: Students' motivation to learn, cognitive load, and satisfaction with the learning process. *Education and Information Technologies*. Advance online publication.

Yan, D. (2023). Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. *Education and Information Technologies, 28*(11), 13943–13967.

Yan, D. (2024). Comparing individual vs. collaborative processing of ChatGPT-generated feedback: Effects on L2 writing task improvement and learning. *Language Learning & Technology, 28*(1), 1–19.

Zhang, R., Zou, D., & Cheng, G. (2025). ChatGPT affordance for logic learning strategies and its usefulness for developing knowledge and quality of logic in English argumentative writing. *System, 128*, 103561.

Zhou, T., Cao, S., Zhou, S., Zhang, Y., & He, A. (2023). Chinese intermediate English learners outdid ChatGPT in deep cohesion: Evidence from English narrative writing. *System, 118*, 103141.

Can GPT Detect Gaming the System in Text Replays?

Cristina MAIER [a*] & Ryan S. BAKER [b]
[a]*Boston College, USA*
[b]*Adelaide University, Australia*
*cristina.maier@bc.edu

Abstract: Detecting gaming the system in an educational setting is essential because such behavior prevents educators from accurately assessing students' knowledge, which negatively impacts their learning outcomes. This study examines the potential of utilizing Large Language Models (such as GPT 4 and GPT 3.5) combined with prompt engineering to identify gaming behaviors within data from a digital system platform. Various prompting strategies are analyzed, and experimental results are compared to previous approaches.

Keywords: Gaming the Systems, ASSISTments, Large Language Model, GPT 3.5, GPT 4

1. Introduction

Gaming the system in education is a behavior in which students exploit the system in an attempt to get correct answers in a learning environment, without possessing the required knowledge to solve such problems. In (Baker, et al., 2009), gaming is defined as "attempting to succeed by exploiting properties of the system rather than by learning the material". Detecting gaming is essential in educational settings, as it limits a learning platform's ability to accurately assess students' knowledge, and pass that information on to educators, ultimately impacting the learning process. Gaming the system has also been shown to be associated with poorer learning outcomes, both in the short-term (Cocea, et al., 2009) and in the long-term (Almeda, et al., 2020; Baker, et al., 2025; San Pedro, et al., 2013)

Prior research in various educational learning environments (Baker, et al., 2008; Baker, et al., 2010; Beal, et al., 2006; Paquette, et al., 2015; Walonoski & Heffernan, 2006) has shown that gaming is correlated with poorer outcome (Beck & Rodrigo, 2014; Cocea, et al., 2009; Pardos, et al., 2014). Gaming the system has been modeled in a variety of ways. Some detectors use machine learning approaches (Baker, et al., 2008; Baker, et al., 2010; Walonoski & Heffernan, 2006), while other detectors use knowledge engineering techniques (Beal, et al., 2006; Walonoski & Heffernan, 2006). Machine learning combined with natural language processing were applied in (Zhang, et al., 2023) to identify gaming in open-ended questions.

However, except for a single case (Paquette, et al., 2015), every time a model of gaming the system is applied to a new learning environment, it has to be developed essentially from scratch. This is an expensive and time-consuming process, which limits its adoption.

One possible solution to this problem could be to use Large Language Models (LLMs) which can translate an idea to very different contexts and domains. In the last couple of years, LLMs have been used to detect complex constructs in text (Cao, et al., 2024). Part of the process of developing some of the existing machine learned and knowledge engineered detectors of gaming the system has been to first automatically transform multi-dimensional student interaction data into a textual and readable form, called "text replays" (e.g. Baker & de Carvalho, 2008; Baker, et al., 2010; Paquette, et al., 2015; Paquette, et al., 2014). Hence, it may be possible to provide this textual representation to an LLM. For example, (Zhang, et al., 2024) created textual representations of whether a student was correct on a sequence of

attempts – transforming each attempt into a sentence - and input it into an LLM. However, this approach did not perform better than a classic machine learning algorithm.

Despite this work, to the best of our knowledge, no work has yet been published on leveraging LLMs to analyze and build detectors for the more complex text replays used in many papers in our field (e.g. Baker & de Carvalho, 2008; Baker, et al., 2010; Paquette, et al., 2015; Paquette, et al., 2014). In this work, we explore whether LLMs (GPT 3.5 and GPT 4 (Achiam, et al., 2024)) combined with prompt engineering can detect gaming the system in a digital learning platform data, by providing the LLM with information from past scientific literature.

Our objectives are to evaluate whether GPT 3.5 and GPT 4 can effectively detect gaming. Additionally, we aim to determine which approach to prompting is best for this task and to identify the optimal number of labeled examples required for this task. Furthermore, we seek to explore whether including gaming patterns and asking GPT to explain its answer can influence its decision-making process in relation to detecting gaming in an educational setting.

2. Dataset

For the experiments in this study, we utilized a publicly available dataset (https://osf.io/2gh56/) on gaming the system from the ASSISTments platform. This dataset was collected through the ASSISTments math learning platform (Razzaq, et al, 2005) and had already been processed into text replays and had human data labels for those text replays. This data set involved data on middle school students' usage of ASSISTments, gathered from August 2008 to April 2010. This dataset includes details about students' answers for various types of problems and activities, such as fill-in-the-blank, single-choice, multiple-choice, and open-response questions. Additionally, the start time, end time, and correctness of each answer were recorded. Student data was organized into clips, with each clip representing a series of consecutive actions performed by a single student. Each clip was labeled as either gaming (label 1) or non-gaming (label 0). The labeling process utilized text replays (Baker, et al., 2006; Baker & de Carvalho, 2008; Paquette, et al., 2015), a method in which coders analyze log files displaying details such as the timestamp of each action, problem context, input provided, relevant skill, and the correctness of the answer.

The original dataset has 1,060 labelled clips (i.e. examples). Out of them, 996 are labelled as non-gaming (i.e. label 0), and 64 are labelled as gaming (i.e. label 1). From these 1,060 clips, we set aside five clips from each label to be used as labeled examples in prompt engineering experiments. Unless stated otherwise, all the experiments below used 1,050 clips for testing and six clips (three labeled as 1 and three labeled as 0) for few-shot prompting.

3. Prompt Engineering

Prompt Engineering has emerged as a technique for using Large Language Models (LLMs) (Sahoo, et al., 2024). It refers to designing LLM prompts (i.e. query or input to an LLM) that contain instructions given to an LLM in an attempt to get a desired response.
In *zero-shot prompting* (Radford, et al., 2019), no labelled examples are added to the prompt. *Few-shot prompting* (i.e. few-shot learning) (Brown, et al., 2020) is a technique in which we add to the prompt a few labeled examples to guide the model towards a correct response.

Ongoing research has investigated approaches for developing more effective prompt designs. In general, prompts are more effective when they include clear and unambiguous instructions, appropriate contextual information, input data, and a defined output indicator (Giray, 2023) Some work has found that asking GPT to explain its answers can enhance the quality of outputs (Bsharat, et al., 2023). *Chain-of-thought* prompting is a technique where a few chain-of-thought demonstrations are included in the prompt (Chowdhery, et al., 2023). This idea is

based on using natural language rational to solve math problems through a series of intermediate steps (Chowdhery, et al., 2023).

Fine-tuning (Jeong, 2024), another technique used with LLMs, involves the adjustment of the parameters of a pre-trained large language model (LLM) for a specific task by further training it on specialized data, resulting in more accurate performance for that task. Both prompt engineering and fine-tuning improve results for specialized tasks, although fine-tuning typically achieves higher accuracy. However, with contemporary LLMs, prompt engineering has become popular due to its many advantages over fine-tuning. It is faster to implement and more cost-efficient, as it avoids the computational costs of training. Prompts are also easier to adjust, and the process requires significantly less labeled data. A study in (Zhang, et al., 2024) examines the comparison between prompt engineering and fine-tuning across various LLMs..

Prompt engineering is an emerging field (Schulhoff, et al., 2024), with increasing usage in the domain of generative AI. Well-constructed prompts have been shown to mitigate hallucination issues in generative AI (Bubeck, et al., 2023), and have shown success in many applications such as defect detection and classification tasks (Yong, et al., 2022). A survey on prompting techniques is presented in (Schulhoff, et al., 2024), as well as in (Chen, et al., 2023).

In this work, we focused on detecting gaming by utilizing two existing pre-trained LLMs: GPT 3.5 (we used the GPT 3.5 Turbo version of GPT 3.5) and GPT 4. We employed few-shot prompting to achieve this. The prompt creation was automated, with the option to enhance the prompts by including a set of gaming patterns (Paquette, et al., 2014) and the ability to request GPT to explain its answers. Prompts began with a gaming the system definition, then ten gaming patterns (Paquette, et al., 2014) (if enabled), and a set of labeled clips.

Previous research has shown that LLMs perform better when they have clear, unambiguous prompt instructions (Chen, et al., 2023), and when examples from the prompt are separated by clear separators (Yoonjeong, et al., 2024). Popular separators include \n\n\n, ### (Yoonjeong, et al., 2024), and +++ (Schick et Schütze, 2022).

We refined our prompt design through multiple iterations to enhance performance. The final prompts were structured using === to separate three sections:

- the initial description: a definition of gaming, and gaming patterns (if enabled),
- labeled clips: a set of clips that are labelled as either gaming or non-gaming,
- output instructions for a new unlabeled clip.

Labeled clips were further separated by ###. Each clip contained a series of activities performed by a student. For each activity, we included details such as the type of problem or activity, the response entered or selected (when available), the correctness of the response, and the time elapsed between activities. This was broadly the same information as provided in human-readable text replays used in past published works (e.g. Baker, et al., 2006; Baker et de Carvalho, 2008; Baker, et al., 2010; Paquette, et al., 2015; Paquette, et al., 2014).

4. Experimental Results

We conducted multiple experiments using the ASSISTments dataset and GPT 3.5 API and GPT 4 API. Along with examining whether GPT 3.5 and GPT 4 could identify gaming, we aimed to determine if the inclusion of gaming patterns in the prompt and asking GPT to justify its response would influence the outcomes.

For the few-shot prompting, we used six labeled examples: three labeled as 0 and three labeled as 1. To determine the number of labeled examples to include in the prompt, we conducted several experiments using both GPT 3.5 and GPT 4, with and without gaming patterns, on a subset of 118 clips (59 for each label).

We created multiple prompts with varying numbers of labeled examples: no labeled examples, two labeled examples (one labeled 0 and one labeled 1), four labeled examples (two labeled 0 and two labeled 1), six labeled examples (three labeled 0 and three labeled 1), eight labeled examples (four labeled 0 and four labeled 1), and ten labeled examples (five labeled 0 and five labeled 1). Performance improved up to six labeled examples, after which no further improvement was observed. As a result, we decided to proceed with experiments using prompts containing six labeled examples (three labeled 0 and three labeled 1).

All experiments below used 1,050 clips for testing. Results were validated using Cohen's kappa (κ), Accuracy, False Positive Rate, and False Negative Rate.

4.1 Detecting Gaming with GPT 3.5

The best results were observed for the full data set when gaming patterns were provided, and no explanation was requested ($\kappa = 0.171$). In prior work (Paquette, et al., 2015), a knowledge engineered model developed on Cognitive Tutor and applied to this same ASSISTments data set achieved a much higher level of performance ($\kappa = 0.256$). Hybrid machine learning-knowledge engineering models in their work ranged in performance from substantially better than our results ($\kappa = 0.248$) to very similar ($\kappa = 0.173$) (Paquette, et al., 2015).

Within our work, lower performance was observed when gaming patterns were included in the prompt and an explanation was requested (κ = 0.112). The lowest performance ($\kappa = 0.087$) was observed when no gaming patterns were provided, and no explanation was requested. This indicates that gaming patterns were beneficial for GPT in making decisions, though not so much when an explanation was requested. Details of the experiment conducted with temperatures of 0.7 and 0.8 can be found in Table 1 below.

Table 1. Experimental Results with GPT 3.5

Gamming Patterns in Prompt	Explanation Requested	Temperature	Accuracy	False Positive Rate	False Negative Rate	Kappa
yes	yes	0.7	65.53%	35.11%	23.57%	0.112
yes	no	0.7	78.95%	19.97%	38.98%	0.171
yes	yes	0.8	66.38%	34.1%	25.42%	0.113
yes	no	0.8	77.8%	20.78%	45.7%	0.137
no	yes	0.8	57.52%	43.59%	23.72%	0.091
no	no	0.8	61.23	39.55%	25.42%	0.087

4.2 Detecting Gaming with GPT 4

To explore whether GPT 4 behaves similarly to GPT 3.5 in detecting gaming, we conducted the same experiments with GPT 4 as we had conducted with GPT 3.5. The best results were observed when gaming patterns were provided without requesting an explanation ($\kappa = 0.13$), or when an explanation was requested without providing gaming patterns ($\kappa = 0.119$).These same scenarios also resulted in the best performance for GPT 3.5, although GPT 3.5 (best $\kappa = 0.171$) outperformed GPT 4 ($best\ \kappa = 0.13$).

As with GPT 3.5, the lowest results were observed either when no patterns were provided, and no explanation was requested (κ = 0.11), or when patterns were included, and explanation requested (κ = 0.107). In these low-performance scenarios, GPT 4 outperformed GPT 3.5, which had a κ of 0.087 at its lowest performance. The results are detailed in Table 2 below.

Table 2. Experimental Results with GPT 4

Gaming Patterns in Prompt	Explanation Requested	Temperature	Accuracy	False Positive Rate	False Negative Rate	Kappa
yes	yes	0.7	68.29%	31.69%	32.2%	0.107
yes	no	0.7	70.476%	29.57%	28.81%	0.13
yes	yes	0.8	68.85%	31.28%	28.81%	0.119
yes	no	0.8	72%	27.4%	35.6%	0.12
no	yes	0.8	72.19%	27.24%	37.28%	0.119
no	no	0.8	69.4%	30.3%	33.8%	0.11

5. Conclusions and Future Work

In this study, our objectives were to evaluate whether prompt engineering with GPT 3.5 and GPT 4 could be utilized to detect gaming the system in an educational context. Additionally, we sought to investigate whether incorporating gaming patterns and requesting GPT to explain its answer impact its decision-making process. We used a previously-labeled ASSISTments dataset for our experiments, and found that GPT benefited from being provided with gaming patterns, but not as much when asked to explain its answer.

The experimental results revealed that GPT 3.5 and GPT 4 were above chance but did not perform very well at detecting gaming in this type of data, compared to previous approaches. Both GPT models were outperformed by other non-LLM models ported from another system to the ASSISTments dataset (Paquette, et al, 2015).

The reason could be that GPT's training data likely contains little that resembles text replays, making it difficult for GPT to interpret interaction data formatted as text, even though it is equally readable for humans. This suggests a potential challenge: the need to not only convert data into human-readable formats but also into formats that align with the types of data present in GPT's training set.

Future work could explore better ways to present interaction data to LLMs and identify tasks they handle well versus those they struggle with. A second idea worthy of further consideration is shifting from a standard prompting approach to a chain-of-thought prompting method. This approach would involve augmenting each labeled example in the prompt with a chain of thought that explains the reasoning behind its corresponding answer. Alternatively, it could be feasible to use an LLM better at complex reasoning (today, GPT o1-preview; GPT o1-mini).

References

Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al. (2024). GPT-4 technical report. ArXiv:2303.08774.

Almeda, M.V., & Baker, R.S. (2020). Predicting student participation in STEM careers: The role of affect and engagement during middle school. Journal of Educational Data Mining, 12(2), 33-47.

Baker, R., Corbett, A., & Koedinger, K. (2009). Educational software features that encourage and discourage “gaming the system”. AIED’09, Brighton, UK, pp. 475–482.

Baker, R.S.J.d., Corbett, A.T., Roll, I., & Koedinger, K.R. (2008). Developing a Generalizable Detector of When Students Games the System. User Modeling & User Adapted Interaction, 18, 287-314.

Baker, R.S.J.d., Corbett, A.T., & Wagner, A.Z. (2006). Human Classification of Low-Fidelity Replays of Student Actions. Data Mining Workshop at 8th Int. Conf on Intelligent Tutoring Systems, 29-36.

Baker, R., & de Carvalho, A. (2008). Labeling Student Behavior Faster and More Precisely with Text Replays. Educational Data Mining ng, Montreal, Québec, Canada, June 20-21.

Baker, R.S.J.d., Mitrovic, A., & Mathews, M. (2010). Detecting Gaming the System in Constraint- Based Tutors. Proc. 18th Conf. User Modeling, Adaptation & Personalization, 267-278.

Baker, R.S, Richey, J.E., Zhang,J., Karumbaiah, S., Andres-Bray J.M.,Nguyen H., Andres J.M.A., & McLaren B.(2025).Gaming the system mediates the relationship between gender and learning outcomes in a digital learning game. Instructional Science 203–238.

Beal, C.R., Qu, L., & Lee, H. (2006). Classifying Learner Engagement Through Integration of Multiple Data Sources. Proc. of the National Conf. on Artificial Intelligence, 151-156.

Beck, J., & Rodrigo, M.M.T.2014. Understanding Wheel Spinning in the Context of Affective Factors. Proc. of the 12th Int'l Conference on Intelligent Tutoring Systems, 162-167.

Brown, T., Mann, B., Ryder, N, Subbiah, M., Kaplan, J., & al. (2020). Language models are few-shot learners. 34th Conference on Neural Information Processing Systems, Vancouver, Canada.

Bsharat, S.M., Myrzakhan, A., & Shen, Z. *(*2023). Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4. ArXiv, abs/2312.16171.

Bubeck S., Chandrasekaran V., Eldan R., Gehrke J., Horvitz E., & al. (2023). Sparks of artificial general intelligence: early experiments with GPT-4;. ArXiv:2303.12712.

Cao C., Chen, E., Fang, Z., Cao, L., Lin, J., & Li, R. (2024). LLM-Generated Personalized Analogies to Foster AI Literacy in Adult Novices. 32nd Int. Conf on Computers in Ed.(ICCE)

Chen B., Zhang Z., Langrene N., & Zhu S. (2023). Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review. arXiv:2310.14735.

Chowdhery, A. Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A, & al. (2023). PaLM: scaling language modeling with pathways. J. Machine Learning Research. 24, 1, Article 240.

Cocea, M., Hershkovitz, A., Baker, R.S.J.d.2009. The Impact of Off-Task and Gaming Behaviors on Learning: Immediate or Aggregate? 14th Int. Conf. on Artificial Intelligence in Education,507-514,.

Giray L. (2023). Prompt Engineering with ChatGPT: A Guide for Academic Writers. Ann Biomed Eng. 51(12):2629-2633. doi: 10.1007/s10439-023-03272-4.

Jeong, C. (2024) Fine-Tuning and Utilization Methods of Domain specific LLMs. https://arxiv.org/abs/2401.02981

Paquette, L., Baker, R.S., de Carvalho, A. et al. (2015). Cross-system transfer of machine learned and knowledge engineered models of gaming the system. User Modeling, Adaptation and Personalization - 23rd International Conference, UMAP.

Paquette, L., de Carvalho, A., Baker, R., & Ocumpaugh, J. (2014). Reengineering the Feature Distillation Process: A case study in detection of Gaming the System, EDM: 284-287.

Pardos, Z.A., Baker, R.S., San Pedro, M.O.C.Z., Gowda, S.M., Gowda, S.M. (2014). Affective States and State Tests: Investigating how Affect and Engagement During the School Year Predict End of Year Learning Outcomes. J. of Learning Analytics, 1(1), 107-128.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8):9.

Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N.T., Koedinger, K., Junker, B., & al. (2005). The Assistment Project: Blending Assessment and Assisting. Proc. of the 12 Annual Conference on Artificial Intelligence in Education, 555-562.

Sahoo, P., Saha, S. , Jain, V. , Mondal, S. & Chadha, A. (2024). A Systematic Survey of Prompt Engineering in Large Language Models:Techniques and Applications. arXiv:2402.07927v

San Pedro, M. O.Z., Baker, R.S.J.D., Bowers, A.J., & Heffernan, N.T. (2013). Predicting college enrollment from student interaction with an intelligent tutoring system in middle school. Preceedings of the 6th international conference on educational data mining.

Schick, T. & Schütze, H. (2022). True Few-Shot Learning with Prompts — A Real-World Perspective. Transactions of the Association for Computational Linguistics, 10: 716 – 731.

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A. & al. (2024) "The Prompt Report: A Systematic Survey of Prompting Techniques", Art. no. arXiv:2406.06608. doi:10.48550/arXiv.2406.06608.

Walonoski, J.A. & Heffernan, N.T. (2006). Detection and Analysis of Off-Task Gaming Behavior in Intelligent Tutoring Systems. Proc. 8th Int'l Conf. Intelligent Tutoring Systems, 382-391.

Yong G., Jeon K., Gil D., Lee G. (2022) Prompt engineering for zero-shot and few-shot defect detection and classification using a visual-language pretrained model. Computer-Aided Civil and Infrastructure Engineering.;38(11):1536–1554.

Yoonjeong, P., Kim, H., Choi, C., Kim, J., and Sohn, J. (2024). Can Separators Improve Chain-of-Thought Prompting? ArXiv abs/2402.10645.

Zhang. L., Lin, J., Borchers, C., Sabatini, J., Hollander, J., Cao, M., & Hu, X. (2024). Predicting Learning Performance with Large Language Models: A Study in Adult Literacy. In Adaptive Instructional Systems: 6th International Conference, AIS 2024

Zhang, J., Pang, S., Andres, A., Baker, R., Cloude, E., Nguyen, H.A., & McLaren, B.M. (2023). Leveraging Natural Language Processing to Detect Gaming the System in Open-ended Questions in a Math Digital Learning Game.33rd Annual Mtg. of the Society in Text and Discourse.

Coded Narratives: Leveraging Women AI Scientists' Memoirs for Inspiration and Knowledge in K-12 AI Education

Shuchi GROVER
Looking Glass Ventures, USA
shuchigrover@acm.org

Abstract: This conceptual paper explores how autobiographical memoirs by prominent AI women scientists can be integrated into K-12 AI literacy curricula to inspire and educate students. Drawing on research on the motivating influences of role models and narrative stories in STEM learning, the paper analyzes *The Worlds I See* by Fei-Fei Li, a Chinese immigrant; *Unmasking AI* by Joy Buolamwini, a Black woman of Ghanaian descent; and *Girl Decoded* by Rana el Kaliouby, an Egyptian Muslim immigrant. We argue that these works from diverse women in AI offer relatable, motivating, and engaging narratives that will resonate with students from underrepresented groups. The analysis explores how such memoirs can demonstrate resilience in the face of challenges, identity affirmation, and career pathways while providing accessible insights into core AI literacy topics such as the critical role of data in AI, ethics & bias in AI, societal impacts, computer vision, and affective human-centered AI. We argue that such resources are powerful catalysts for inclusive education, serving both as motivational tools and as pedagogical resources that deepen AI understanding.

Keywords: AI Literacy, STEM role models, inclusive education, storytelling

1. Introduction

The rapid spread of artificial intelligence (AI) and machine learning (ML) across domains—from education to transportation—signals a transformative shift that necessitates AI literacy in K-12 education. As AI increasingly mediates human experiences, students must learn to understand, evaluate, and shape these technologies—not just as users, but as informed citizens and potential innovators (Touretzky et al., 2019). AI literacy, encompassing technical knowledge, ethical awareness, and societal implications, is critical for preparing youth for a world shaped by AI (Long & Magerko, 2020). Yet, current K-12 AI education often lacks depth and accessibility, especially for students from historically marginalized backgrounds who face systemic barriers, including the absence of relatable role models in STEM (NSF, 2022).

The learning sciences have long called for inclusive pedagogies that integrate STEM concepts with motivating narratives to foster comprehension and aspiration (Pena, 2014). STEM research also shows that role models enhance motivation, self-efficacy, and persistence, especially for marginalized students (Shin et al., 2016). To address this need, this conceptual paper proposes the integration of memoirs into AI literacy efforts. We analyze three recent memoirs by pioneering women in AI—Fei-Fei Li's *The Worlds I See* (2023), Joy Buolamwini's *Unmasking AI* (2023), and Rana el Kaliouby's *Girl Decoded* (2020)—to demonstrate how these narratives offer both motivation and rich AI content K-12 curricula. These deeply personal stories intersect with critical AI/ML topics, making them compelling tools for engaging students, particularly women and those from underrepresented groups. Each memoir exemplifies the blend of technical insight and human experience. Li, a Chinese immigrant, recounts her path through hardship and into leading advancements in computer vision. Buolamwini, a Black woman of Ghanaian descent, reveals how AI bias affected her personally and drove her to become a leader in algorithmic justice. El Kaliouby, an Egyptian-Muslim immigrant, explores affective computing while navigating personal and cultural

complexities. These narratives add the human element to AI development, helping students see humans like them as part of the AI story.

Additionally, storytelling as pedagogy increases engagement and helps ground abstract ideas in relatable human contexts (Dahlstrom, 2014). These memoirs, therefore, function as dual-purpose tools: they inspire through representation and provide substantive entry points into AI literacy. Each memoir contributes conceptual insights aligned with AI4K12 (n.d)'s "Big Five Ideas". Li explains the pivotal role of big data in revolutionizing ML. Buolamwini critiques algorithmic fairness and the societal impacts of biased AI, while also modeling ethical dataset creation. El Kaliouby introduces affective computing and the interplay between technology and humanity. These stories thus interweave personal, technical, historical, and ethical threads, addressing key AI literacy competencies.

Following a literature review and brief analysis of the memoirs, this paper proposes a pedagogical framework positioning them as both motivational catalysts and sources of AI conceptual content. Centering such diverse human stories supports inclusive education and broadens participation, helping students envision themselves as future AI innovators.

2. Theoretical Framing

This work is grounded in (a) the need to address AI literacy and pedagogies to support AI literacy education, (b) research on the benefit of role models to address underrepresentation in STEM fields in general and AI in particular, and (c) the power of narratives and storytelling to engage learners along the cognitive as well as intrapersonal dimensions of learning.

2.1 AI Literacy in K-12 Education

AI literacy efforts in K-12 education are growing globally to prepare students for an AI-driven world. Initiatives like "Day of AI" and "National AI Literacy Day" offer free curricula, tools, and professional development. Nonprofits like the AI Education Project, Raspberry Pi Foundation, EqualAI.org and TeachAI.org promote AI learning and provide resources for educators and communities to engage with AI responsibly. UNESCO's AI Competency Framework (2024) emphasizes a human-centered, ethics-infused approach across STEM and social studies.

Core AI literacy curricula combine technical, practical, and ethical components, captured by AI4K12 (n.d.)'s Big 5 Ideas: *Perception* (using sensors to interpret the world); *Representation and reasoning* (modeling and reasoning); *Learning* (AI learns from data); *Natural interaction* (communication through language and emotion); and *Societal impact* (AI's positive and negative effects). Grover (2024) highlights how students explore AI through accessible tools like Google's Teachable Machine and TensorFlow or code in Scratch and Python, and interdisciplinary projects link AI to real-world contexts like criminal justice and healthcare. For younger or tech-limited learners, unplugged activities introduce AI concepts. Ethical themes—bias, fairness, privacy—are central. Many curricula reference Buolamwini & Gebru's (2018) *Gender Shades* study and Buolamwini's TED Talk on AI bias. However, few go further to incorporate personal stories of underrepresented AI scientists.

2.2 Promoting diversity through role models in STEM education

Women remain underrepresented in AI, comprising under 19% of U.S. AI/CS Ph.D. graduates and only 16% of tenure-track CS faculty (Stanford HAI, 2021). African Americans and Hispanics make up just 2.4% and 3.2% of AI Ph.D. recipients. Addressing this imbalance requires continued emphasis on engaging underrepresented students.

Research shows role models with shared identities—by gender, race, or background—boost self-efficacy and aspirations for marginalized learners. Students exposed to relatable figures report higher confidence and actively seek role models who reflect their values (Gillooly, Hardt, & Smith, 2021). According to social cognitive theory, perceived similarity strengthens belief in one's potential (Gladstone & Simpian, 2021), and role models who share struggles and counter stereotypes are especially motivating (Steinke et al., 2022).

This evidence has shaped STEM curricula that bring in scientists from academia and industry, though such visits are often brief. Broader efforts include museum and online initiatives like *I Am a Scientist* (2023) which showcase diverse scientific role models.

2.3 The Power of Stories and Narrative in STEM Learning

Storytelling enhances STEM learning by anchoring technical content in emotionally resonant, human-centered contexts. As the oldest teaching form, storytelling organizes complex ideas into sequences that improve retention and critical thinking (Barchas-Lichtenstein et al., 2023). Its benefits are well-supported in STEM education (Dahlstrom, 2014).

Stories about scientists humanize STEM and make abstract concepts relatable. The "narrative effect" improves learning by boosting emotional engagement and memory (Soares et al., 2023). Personal stories also address representation gaps by challenging stereotypes and helping underrepresented students imagine themselves in STEM (Collins, 2022). Narrative pedagogies also place science within historical and cultural frames. Hobbs and Davis (2013) identify "inward-looking" narratives that connect to students' lives and "outward-looking" ones linking science to societal issues—both enhance relevance. Hearing how others overcame similar obstacles promotes a sense of belonging and encourages persistence particularly for students facing systemic barriers. (Dahlstrom, 2014).

3. Analysis of three memoirs

3.1 'The Worlds I See' by Fei-Fei Li

Background and Journey. Born in Beijing and raised in Chengdu, China, Li immigrated with her family to the U.S. at age 15. The family settled in New Jersey and faced language, cultural, and financial hardships. Despite these, Li excelled academically, earning a scholarship to Princeton to study physics. She continued her graduate work at Caltech, completing a Ph.D. in electrical engineering (computer vision), and later held CS faculty positions at Princeton and Stanford.

Inspiration and Mentors. Li's credits her father, an electrical engineer who modeled deep intellectual wonder, and her mother who pushed her to pursue a career in science. Mr. Sabella, her high school math teacher, became a pivotal mentor and family supporter, even helping finance their dry-cleaning business. Her Ph.D. advisors at Caltech, Pietro Perona and Christof Koch, shaped her scientific path.

Major Contribution to AI/ML: Data Matters. Li's signature achievement is *ImageNet*, a dataset of over 14 million labeled images in 22,000 categories. Initially met with skepticism, ImageNet proved that large-scale data could transform ML, launching the deep learning revolution. Li also co-founded *AI4ALL*, a nonprofit introducing AI to underrepresented youth.

Challenges and Persistence. Li's mother's chronic illness was a constant challenge and caused financial strain in their early U.S. years. Professionally, creating ImageNet took years and faced repeated dismissal from experts. Only after adopting Amazon Mechanical Turk could her team crowdsource the needed image labels. While she doesn't dwell on gender bias, she reflects on its impact, noting shared frustrations with fellow female AI researchers.

"North Star" and Student Advice. Li champions "human-centered AI" that enhances, not replaces, human capabilities. She warns of biased algorithms, privacy risks, and power concentration, urging AI to be ethically grounded from the start. Her guiding principle is to reimagine AI as a deeply humanistic endeavor. She also notes: "The best work is often done...in shared spaces of science...rather than within the bubbles of our own fields."

3.2 Unmasking AI by Joy Buolamwini

Background and Inspirations. Born in Canada to Ghanaian parents, Buolamwini describes herself as a "daughter of art and science." Her mother, Frema the Akan, was an artist, and her

father, Dr. John Buolamwini, a medical chemist and computational researcher. This dual influence shaped her approach to tech, blending creativity with scientific inquiry. A PBS segment featuring a young MIT professor, Cynthia Breazeal, and her robot *Kismet* sparked her early interest in AI and robotics.

Mentors and Supporters. Buolamwini was deeply inspired by her parents and ancestral lineage. At MIT, Ethan Zuckerman, her advisor at the Center for Civic Media, encouraged her to connect technical work to issues of power and inequality. Dr. Timnit Gebru, both mentor and collaborator, supported her development in computer vision and advocacy work.

Major Contributions and Achievements. Having to wear a white mask to get face detection software to recognize her face launched Buolamwini's landmark "Gender Shades" study. It revealed commercial facial recognition systems had the highest error rates on darker-skinned women, exposing racial and gender bias in tools from IBM, Microsoft, and Face++. She coined this systemic issue "the coded gaze." Her spoken-word video *AI, Ain't I A Woman?* highlighted misclassifications of iconic Black women like Michelle Obama and Serena Williams, launching her idea of "evocative audits"—artistic demonstrations of algorithmic harms. She also created the *Parliaments Benchmark Dataset*, addressing bias in training data by selecting diverse parliaments from Africa and Nordic countries and using Fitzpatrick skin types rather than racial categories. Personally labeling 1,270 images by skin tone and gender, she enabled intersectional analysis that revealed disparities hidden by aggregate metrics.

Challenges and Navigating Discrimination. When she shifted her research toward algorithmic bias, peers warned it could jeopardize her career. She faced pushback from companies like Amazon, which challenged her findings. Despite this, she gained support from researchers who publicly defended her work. As one of few Black women in AI, Buolamwini also encountered discrimination. At an EU tech panel in Brussels, she was denied entry by security. On *60 Minutes*, despite contributing to a segment, she was ultimately not featured.

Views on an AI-Driven World & Advice for Students. "AI will not solve poverty… or discrimination," she writes, noting that these are cultural, not technical, problems (p. 22). She rejects tech determinism and calls for algorithmic justice—ensuring people have "a voice and a choice" in AI systems that shape their lives (p. 305). Her message to students: "You don't need a PhD from MIT to make a difference… All you need is a curious mind and a human heart" (p. 303). She urges young technologists to "question the status quo… listen to those harmed by algorithms… and explore intellectual terrain beyond computer science" (p. 298).

3.3 Girl Decoded by Rana El Kaliouby

Background and Journey. Raised in Kuwait and Egypt, el Kaliouby grew up in an education-focused household. Her mother was among the first female programmers in the Middle East; her father taught computer science. Reading *Affective Computing* by MIT's Rosalind Picard during her undergrad at American University of Cairo sparked her interest in Emotion AI. She pursued this path through a master's at AUC, a PhD at Cambridge, and a postdoc at MIT.

Mentors and Support Networks. Her PhD advisor, Dr. Peter Robinson, and autism researcher Simon Baron-Cohen provided key research support. At MIT, Picard became both mentor and co-founder. Her mother's unwavering support was also crucial.

Major Contributions. El Kaliouby pioneered Emotion AI—teaching machines to recognize emotions. Her research led to *FaceSense* and *Affectiva*, involving training models on 9 million faces from 87 countries, with applications spanning healthcare, education, and autism support.

Challenges and Navigating a Male-Dominated Field. She faced personal challenges as an Arab-Muslim woman balancing solo parenting, cross-continental moves, and divorce while building a career. Fundraising for Affectiva was difficult in a male-dominated tech space that undervalued emotion-based technologies. Though not the focus of her memoir, she notes pitching only to white male investors and stresses the risks of homogenous AI development: "we will be unwittingly duplicating the biases that exist in society."

Vision for an AI-Driven World and Advice for Students. Her mission is "bringing emotional intelligence to computers." She champions ethical, inclusive, privacy-conscious AI.

Her story shows students that empathy, persistence, and diverse perspectives are vital to creating tech that serves everyone.

3.4 Additional AI-related Insights from the Memoirs

Across all three memoirs, recurring themes include persistence through adversity, questioning the status quo, unwavering pursuit of a “North Star”, and a clear understanding that both data and human choices shape AI at every stage. All memoirs also offer the following lessons and technical, conceptual, and humanistic insights into AI research and discovery in tangible ways:
1. AI innovation is inherently interdisciplinary,
2. It is built through collaboration across diverse teams,
3. AI augments—not just replaces—human abilities,
4. AI needs diverse data to serve everyone,
5. AI reflects societal biases—it is not neutral,
6. Personal experience can drive powerful research questions,
7. Innovation often unfolds in small, iterative steps,
8. Creative expression can convey complex technical ideas
9. AI is still evolving, and
10. Everyone has a role in shaping AI’s future.

4. A Pedagogical Framework for Integrating Memoirs for AI Literacy

Integrating memoirs of women of color AI scientists into K-12 AI literacy calls for a pedagogical approach that balances technical learning with identity affirmation. Drawing on narrative research in STEM, and inspired by Meier (1997)’s "learning in small moments," we propose a four-part framework where biographies offer accessible entry points into complex concepts.

1. **Align Narratives with AI Literacy Goals**: Use memoirs to illustrate key AI concepts such as perception, learning from data, societal impact, and reasoning (AI4K12, n.d).
2. **Foster Identity Affirmation**: Highlight diverse scientists’ contributions and stories to help students see themselves in STEM, promoting belonging.
3. **Engage Underrepresented Students**: Leverage relatable narratives to make STEM more inclusive and meaningful.
4. **Design Lessons Around Personal Stories**: Integrate technical content with memoir excerpts to deepen comprehension and motivation.

This framework supports both cognitive and emotional learning. Aligning stories with AI concepts allows students to extract conceptual insights embedded in lived experiences. For example, students might identify and discuss AI principles within memoir excerpts, grounding abstract ideas in human stories. Identity affirmation responds directly to barriers in STEM participation. Memoirs that challenge dominant assumptions about who belongs in STEM. help underrepresented students reimagine pathways into STEM. This approach moves from viewing STEM through abstract systems to “seeing it big” (Greene, 1995)—through personal narratives that illuminate both technical ideas and aspirational journeys.

.

5. Limitations, Implications, and Conclusion

This paper argues that autobiographical memoirs by pioneering women in AI serve a dual purpose: they are both motivational and rich content sources for AI literacy. By embedding technical concepts within personal narratives, these stories make AI more tangible, relatable, and inspiring—especially for underrepresented learners. While further empirical research is needed, the proposed framework offers a foundation for inclusive, human-centered AI education. Centering diverse voices is not only a matter of equity but essential to cultivating the next generation of ethical, critical, and innovative thinkers in an AI-driven world.

This conceptual paper has limitations. It draws on three memoirs which, while rich, reflect a narrow slice of the diverse experiences within AI. These narratives represent specific cultural, geographic, and disciplinary perspectives and are written in English, largely situated

within the U.S., which may limit their global resonance. Articles or interviews of women scientists in media could be used to expand the pool of stories as resources.

Although the framework is grounded in existing literature, its effectiveness must be validated through empirical studies across varied educational contexts and age groups. Meaningful integration into AI curricula will also require effort from educators and curriculum designers. Time must be allocated for reading and discussion, and teachers will need to connect story elements with technical content. Accessibility may pose additional challenges—due to reading level, language, or resource constraints—necessitating age-appropriate adaptations such as summaries or excerpts.

Despite these limitations, integrating memoirs into AI literacy holds promise for inclusive STEM education. These narratives humanize AI development, disrupt stereotypes, and foster belonging for students historically excluded from STEM. They also open doors for important ethical and societal discussions. Memoirs also offer interdisciplinary possibilities, blending computer science with literacy, history, and social studies. This aligns with learning research advocating culturally responsive, engaging teaching. To support implementation, curriculum designers would need to create supplementary materials—lesson plans, guides, and digital resources—to scaffold classroom use of these powerful stories.

References

AI4K12. (n.d.). AI4K12: Five Big Ideas in AI. Retrieved from https://ai4k12.org

Barchas-Lichtenstein, J., Sherman, M., Voiklis, J., & Clapman, L. (2023). Science through storytelling or storytelling about science? Identifying cognitive task demands and expert strategies in cross-curricular STEM education. In Frontiers in Education (Vol. 8, p. 1279861). Frontiers Media.

Collins, S. (2022). How storytelling inspires representation in STEM communities. AAAS IUSE.

Dahlstrom, M. F. (2014). Using narratives and storytelling to communicate science with nonexpert audiences. Proceedings of the national academy of sciences, 111(supplement_4), 13614-13620.

el Kaliouby, R. (2020). Girl decoded: A scientist's quest to decode our emotions. Currency.

Gillooly, S. N., Hardt, H., & Smith, A. E. (2021). Having female role models correlates with PhD students' attitudes toward their own academic success. Plos one, 16(8), e0255095.

Gladstone, J. R., & Cimpian, A. (2021). Which role models are effective for which students? A systematic review and four recommendations for maximizing the effectiveness of role models in STEM. International journal of STEM education, 8, 1-20.

Greene, M. 1995. Releasing the Imagination: Essays on Education, the Arts, and Social Change. San Francisco, CA: Jossey-Bass.

Grover, S. (2024). Teaching AI to K-12 learners: Lessons, issues, and guidance. In Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1 (pp. 422-428).

Hobbs, L., & Davis, R. (2013). Narrative pedagogies in science, mathematics and technology. Research in Science Education, 43, 1289-1305.

Li, F. (2023). The Worlds I See: Curiosity, exploration, and discovery at the dawn of AI. Flatiron Books.

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–16. Meier, D. (1997). *Learning in small moments: Life in an urban classroom*. TC Press.

National Science Foundation (NSF). (2022). STEM education for the future: A visioning report. NSF.

Peña, A. (2022). Supporting the STEM Aspirations of Youth Who Have Been Pushed Out. Science Education International, 33(4), 409-421.

Shin, J. E. L., Levy, S. R., & London, B. (2016). Effects of role model exposure on STEM and non-STEM student engagement. Journal of Applied Social Psychology, 46(7), 410–427.

Soares, S., Gonçalves, M., Jerónimo, R., & Kolinsky, R. (2023). Narrating science: Can it benefit science learning, and how? A theoretical review. *Journal of Research in Science Teaching*, *60*(9).

Stanford HAI. (2021, April 15). *AI Index Diversity Report: An unmoving needle*. Stanford University.

Steinke, J., Applegate, B., Penny, J. R., & Merlino, S. (2022). Effects of diverse STEM role model videos in promoting adolescents' identification. *Intl Journal of Sc and Mathematics Ed*, 1-22.

Touretzky, D., et al. (2019). Envisioning AI for K-12: What should every child know about AI? *Proceedings of the AAAI Conference.33*(1), 9795–9799.

UNESCO. (2024). *AI competency framework for students*. United Nations Educational, Scientific and Cultural Organization. https://unesdoc.unesco.org/ark:/48223/pf0000391105.

A Federated Learning Neural Network For Student Dropout Prediction

Anuradha Kumari SINGH[a*] & Karthikeyan S[b]
[a]*Department of Computer Science, Banaras Hindu University, India*
[b]*Department of Computer Science, Banaras Hindu University, India*
*anuradha@bhu.ac.in

Abstract: Federated learning offers a promising solution for training neural networks on distributed datasets while reducing the risk of data leakage, a critical concern in educational data mining where student records are highly sensitive. We propose FedAvg-NNFL, a privacy-preserving neural network for student dropout prediction within a federated framework. Each institution trains a local model on its private data, and a global model is created by aggregating locally trained weights, ensuring privacy and compliance with data-sharing restrictions. Using a benchmark dataset from the Polytechnic Institute of Portalegre, containing demographic, socioeconomic, academic, and macroeconomic features, we compared FedAvg-NNFL against centralized and locally trained models. FedAvg-NNFL achieved accuracy 0.9280, precision 0.9320, recall 0.8972, F1-score 0.9142, and AUC 0.92 under 10-fold cross-validation, outperforming the local model and closely matching centralized performance. These results demonstrate the potential of federated learning for developing accurate, privacy-aware predictive models in educational contexts.

Keywords: Federated Learning, Neural Network, Student Dropout, Predictive model

1. Introduction

Student dropout prediction (SDP) plays a vital role in identifying at-risk students and enabling proactive interventions that enhance retention, institutional performance, and long-term learner success. Leveraging machine learning and optimization techniques, universities can analyze diverse factors including attendance, engagement, socio-economic background, and academic performance to uncover patterns associated with dropout. However, developing a centralized SDP model remains challenging due to the difficulty of data collection and aggregating data across departments and institutions, strict privacy and compliance requirements, variability in data quality and formats, and the scalability demands posed by increasing data volumes.

In the literature review, we found very few studies Fachola et al. (2023); Farooq et al. (2024); Johany´ak, Laufer, and Kov´acs (2024); Kuang et al. (2025); van Haastrecht, Brinkhuis, and Spruit (2024) that employed a federated learning (FL) approach for training a global model for student dropout prediction. To bridge this gap, we developed a federated learning model based on neural networks (FedAvg-NNFL) for student dropout prediction in universities, effectively addressing the challenge of maintaining data privacy.

This study has the following main contributions:

- Collecting student data from the universities is creating data privacy concerns as student records contain sensitive information; to solve this problem, we proposed the use of a federated learning framework which is not explored more in educational data mining.
- Neural network-based federated learning model, FedAvg-NNFL is proposed to predict the student's dropout at university.
- We compared our FedAvg-NNFL model performance with the centralized and locally trained models, and our model outperformed.

The remainder of this article is organized as follows: Section 2 presents a concise overview of recent research on student dropout prediction and the use of federated learning. Section 3 details the proposed methodology and algorithms. Section 4 discusses the experimental results and performance evaluation of the proposed model. Finally, Section 5 highlights the key findings and contributions of the study.

2. Literature Review

Most of the studies have been done on a centralized model, and very few studies have used a decentralized model in educational data mining. The Researcher utilized a federated learning approach to identify students' unique learning patterns and individualized variations across different courses. Another research applied federated learning methods to address two learning analytics challenges: dropout prediction and unsupervised student classification Fachola et al. (2023). The researcher explored the potential of combining federated learning with strategic client selection and sampling techniques to enhance dropout prediction in healthcare applications Nikolaidis and Efraimidis (2025). Researchers introduced a new federated learning approach, DQFed, for learning analytics that consists of aggregating the local models based on their weights computed according to a quality-driven model Bernardi, Cimitile, and Usman (2024).

Another study explored the technical foundations of federated learning, its application in education systems, and its potential benefits in improving learning outcomes and fostering data-driven decision-making in education Surapaneni, Bojjagani, and Sharma (2024). Researchers introduced a new federated learning standard that safeguards dataset confidentiality and enables the prediction of student grades, classified into four categories: low, good, beverage, and drop Chu et al. (2022).

There are some recent studies that have applied federated learning techniques to student performance and dropout prediction. Kuang et al. (2025) leveraged federated frameworks to investigate depressive symptoms among students, demonstrating FL's applicability beyond academic metrics. Farooq et al. (2024) proposed a privacy-preserving federated model for predicting student grades while ensuring data confidentiality. Van Haastrecht et al. (2024) conducted a comparative study between local, centralized, and federated learning approaches across multiple educational prediction tasks, including learning outcomes, dropout rates, and question correctness. Meanwhile, Johanyák et al. (2024) introduced a collaborative federated learning initiative between two universities focused on developing a fuzzy logic-based model for predicting student results. Collectively, these studies underscore the growing interest in using federated learning to address privacy concerns while enabling effective predictive modeling in educational environments.

3. Methodology

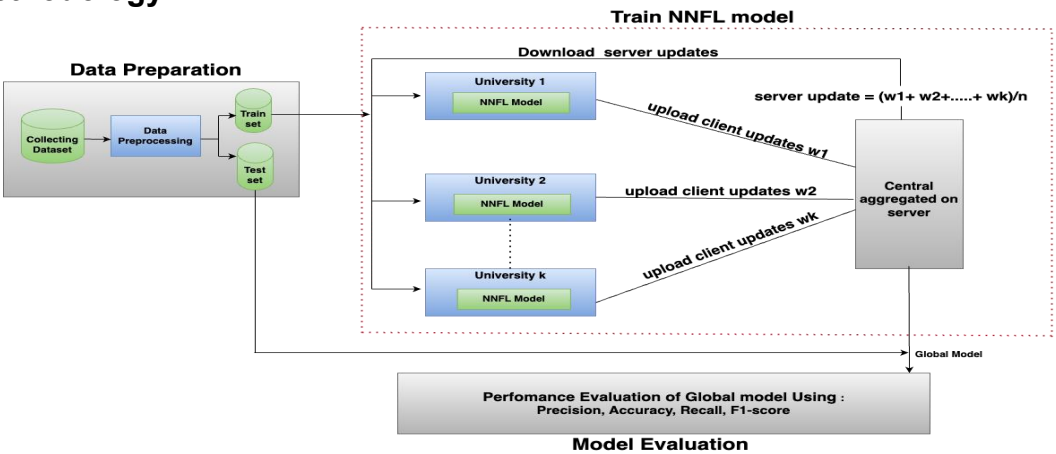


Figure 1. Overview of proposed methodology.

This section contains the discussion of the proposed methodology. Our proposed methodology follows three stages: 1. Data preparation, 2. Training of Federated model, 3. Performance evaluation. **Figure 1** depicts the workflow of our proposed methodology.

3.1 Proposed Algorithm

The Federated Averaging (FedAvg) , **algorithm 1** is a privacy-preserving training strategy that enables collaborative learning of a shared neural network model across multiple clients without centralizing their data. This approach is especially relevant in sensitive domains such as education, where sharing raw student data between institutions may violate privacy regulations.

Algorithm 1 Federated Averaging Neural Network Training (FedAvg-NNFL)

Require:
NUM_CLIENTS ← Number of clients
CLIENT_DATA ← $\{(X_i^{train}, X_i^{test}, y_i^{train}, y_i^{test})_{i=1}^{NUM_CLIENTS}$
NUM_ROUNDS ← Number of federated training rounds
EPOCHS_PER_CLIENT ← Number of local epochs per round
BATCH_SIZE ← Batch size for local training
Ensure: Final global model weights WG, mean global accuracy and loss
Initialize global model G with weights W_G
for r = 1 to NUM_ROUNDS do
    Initialize empty list client weights ← []
    for i = 1 to NUM_CLIENTS do
        Create local model M_i and set weights $W_i \leftarrow W_G$
        Train Mi on $(X_i^{train}, y_i^{train})$ using EPOCHS_PER_CLIENT and BATCH_SIZE
        Append W_i to client weights
    end for
    Aggregate local weights using element-wise mean: $W_G = \frac{1}{NUM_CLIENTS} \sum_{i=1}^{NUM_CLIENTS} W_i$.
    Update global model G with weights W_G
    for i = 1 to NUM_CLIENTS do
        Evaluate G on $(X_i^{train}, y_i^{train})$ to obtain accuracy acc_i
    end for
    Compute and print mean training accuracy: Accuracy$_r$ = $\frac{1}{NUM_CLIENTS} \sum_{i=1}^{NUM_CLIENTS} acc_i$
end for
Final Evaluation:
for i = 1 to NUM_CLIENTS do
    Evaluate G on (X_i^{test}, y_i^{test}) to get loss l_i and accuracy acc_i
end for
Compute:
    Mean Accuracy: $\bar{A} = \frac{1}{NUM_CLIENTS} \sum_{i=1}^{\text{NUM_CLIENTS}} acc_i$
    Mean Loss: $\bar{L} = \frac{1}{NUM_CLIENTS} \sum_{i=1}^{\text{NUM_CLIENTS}} l_i$
return Final weights W_G , $\bar{A}$, and $\bar{L}$

The Federated Averaging Neural Network Training (FedAvg-NNFL) algorithm begins with the initialization of a global neural network model G with random weights. This global model is iteratively refined over multiple federated training rounds based on contributions from distributed clients. During each round, every client receives a copy of the global model and initializes its local model with the same weights. The local model is then trained on the client's private dataset $(X_i^{train}, y_i^{train})$ using a fixed number of epochs and batch size. After training, only the updated local model weights W_i are sent back to the central server. The server then performs federated averaging by computing the element-wise mean of the collected weights:

$$W_G = \frac{1}{NUM_CLIENTS} \sum_{i=1}^{NUM_CLIENTS} W_i.$$

This aggregated weight set becomes the updated global model, effectively integrating knowledge from all clients while preserving their data privacy. Optionally, the global model may be evaluated on each client's training data to monitor intermediate performance, such as mean training accuracy. After completing all training rounds, the final global model is evaluated on the test datasets of all clients to measure its performance. For each client, the test accuracy and loss are recorded and averaged to compute the overall global performance:

$$\text{Mean Accuracy } \bar{A} = \frac{1}{NUM_CLIENTS} \sum_{i=1}^{\text{NUM_CLIENTS}} acc_i ;$$

$$\text{Mean Loss } \bar{L} = \frac{1}{NUM_CLIENTS} \sum_{i=1}^{NUM_CLIENTS} l_i .$$

The algorithm ultimately returns the final global model weights W_G, which are suitable for inference or further fine-tuning, along with the average accuracy and loss across all clients, reflecting the overall effectiveness of the federated learning process.

3.2 Dataset and Data Preprocessing

We used the publicly available dataset from the Polytechnic Institute of Portalegre (Realinho et al., 2022), previously employed in our earlier study. The dataset (743 KB) contains 4,424 records and 37 attributes covering demographic, socioeconomic, macroeconomic, enrollment, and first- and second-semester academic data. Originally, it had three target classes dropout, graduate, and enrolled. For this study, we excluded the enrolled category (students whose final status was unknown), leaving 3,630 records with two definitive outcomes: dropout and graduate. Preprocessing involved loading the data from Excel, removing rows with missing values, and separating features (X) from the target (y). Numerical features were standardized using StandardScaler. To address class imbalance, we applied the SMOTE–Tomek method, which oversamples the minority class and removes borderline samples. Finally, the data was split into 80% training (2,904 records) and 20% testing (858 records).

4. Results and Discussion

Experiments were conducted in Python using scikit-learn (v1.3.2) for preprocessing and TensorFlow for neural network implementation, training, and testing. All runs were performed on Google Colab with a T4 GPU. The proposed FedAvg-NNFL model combines neural networks with federated learning for student dropout prediction. The neural network comprises three dense layers: 16 ReLU-activated neurons (input: 36 features), 8 ReLU-activated neurons, and a final sigmoid-activated neuron for binary classification.

Federated learning enables collaboration among multiple parties to develop a model without directly sharing their data. This approach involves training the model on a decentralized network, where the data stays on local client devices or servers at various locations. In this study, we introduced a FedAvg-NNFL model. NUM_ROUNDS, EPOCHS_PER_CLIENTS, and BATCH_SIZE are the key hyperparameters affecting the local training process of each client. We trained our FedAvg-NNFL model with 50 clients, NUM_ROUNDS=10, EPOCHS_PER_CLIENTS=20, and BATCH_SIZE=64.

Figure 2 depicts the comparison between the local model and federated model FedAvg-NNFL training over different clients. The performance of local and federated training across epochs reveals distinct trends. In local training (represented by the blue line with circles), the accuracy starts relatively high at epoch 5 (0.923) but then drops to its lowest point by epoch 10 (0.911). It slightly improves at epoch 20 (0.912), reaches a peak again at epoch 30 (0.924), and then dips slightly by epoch 50 (0.916). This fluctuation indicates instability, likely caused by the model's sensitivity to local data distributions, suggesting possible overfitting or limited generalization.

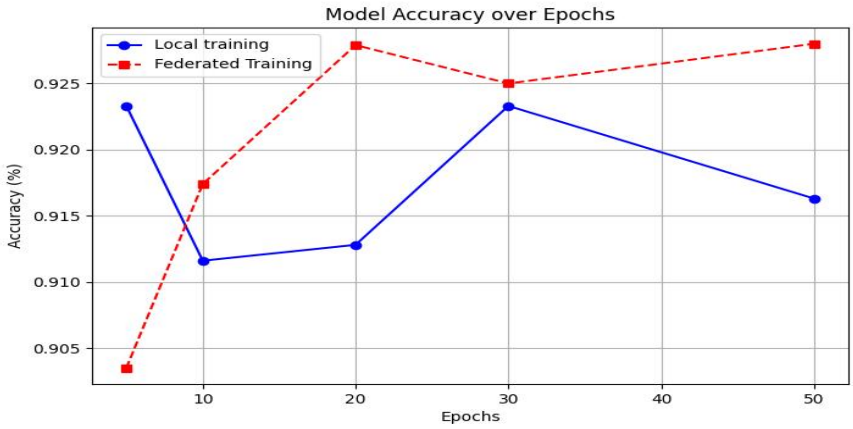


Figure 2. Comparison between Local model and federated model(FedAvg-NNFL) training over different clients at 10-fold cross-validation.

In contrast, federated training (depicted by the red line with squares) starts at a lower accuracy 0.904 at epoch 5 but improves steadily, reaching 0.917 at epoch 10 and peaking at 0.927 by epoch 20. Although there's a minor dip at epoch 30 (0.925), it regains peak accuracy 0.927 by epoch 50. This consistent improvement and stabilization demonstrate the robustness and generalization capabilities of federated learning, likely due to the aggregation of knowledge from diverse client data. Notably, federated training consistently outperforms local training from epoch 10 onward.

Table 1. *Comparison between centralized, local and federated model (FedAvg-NNFL) at 10-fold cross-validation*

Training Mode	Accuracy	Precision	Recall	F1-score	AUC
Centralized Training	0.9628	0.9083	0.9036	0.9026	0.96
Local Training	0.9175	0.9384	0.8909	0.9140	0.92
FedAvg-NNFL (federated)	0.9280	0.9320	0.8972	0.9142	0.92

Table 1 compares centralized, local, and federated (FedAvg-NNFL) training. Centralized training achieved the highest accuracy (0.9628) and AUC (0.96) under 10-fold cross-validation, as expected from full data access. The local model attained the highest precision (0.9384) but lower accuracy (0.9175) and recall (0.8909) due to client-specific training. FedAvg-NNFL offered a balanced performance, with accuracy 0.9280, precision 0.9320, recall 0.8972, F1-score 0.9142, and AUC 0.92, outperforming the local model in most metrics while preserving privacy and closely approximating centralized results.

Limitations and Future Work

While this study demonstrates the feasibility of federated neural networks for student dropout prediction, several limitations remain. First, our implementation uses a standard FedAvg algorithm without customization for handling non-IID data or client-specific behavior. Future work will explore client personalization, asynchronous updates, and adaptive aggregation to enhance model robustness. Second the use of signal dataset. We plan to incorporate Dirichlet-based non-IID partitioning to better assess generalizability. Additionally, fairness analysis across demographic subgroups and performance under class imbalance conditions will be introduced to ensure equitable model behavior. These enhancements will strengthen both the technical contribution and the practical utility of our approach in privacy-sensitive educational settings.

5. Conclusion

Ensuring student data privacy is a major challenge in educational data mining, as university records often contain highly sensitive information. To address this, we propose FedAvg-

NNFL, a privacy-preserving federated learning framework for predicting student dropout without sharing raw data. Each institution trains a local neural network on its private dataset, and a global model is constructed by aggregating locally trained weights, ensuring compliance with data-sharing restrictions. We evaluated FedAvg-NNFL on the benchmark dataset from the Polytechnic Institute of Portalegre, which includes macroeconomic, socioeconomic, demographic, admission, and academic performance attributes. Compared with a centralized neural network and a locally trained model, FedAvg-NNFL achieved balanced results: accuracy 0.9280, precision 0.9320, recall 0.8972, the highest F1-score of 0.9142, and an AUC of 0.92 under 10-fold cross-validation. The findings highlight federated learning's potential to deliver accurate, privacy-aware predictive models, enabling institutions to identify at-risk students and design timely interventions to improve retention and success.

Data Availability Statement

The data presented in this study are publicly available at https://doi.org/10.5281/zenodo.5777339.

Acknowledgements

We would like to thank all the people who prepared and revised previous versions of this document.

References

Bernardi, M. L., Cimitile, M., & Usman, M. (2024). An Entropy-based Federated Learning approach for the student's dropout prediction. In *Book of Abstracts* (p. 48).

Chu, Y. W., Hosseinalipour, S., Tenorio, E., Cruz, L., Douglas, K., Lan, A., & Brinton, C. (2022, October). Mitigating biases in student performance prediction via attention-based personalized federated learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (pp. 3033-3042).

Fachola, C., Tornaría, A., Bermolen, P., Capdehourat, G., Etcheverry, L., & Fariello, M. I. (2023). Federated learning for data analytics in education. *Data*, *8*(2), 43.

Farooq, U., Naseem, S., Mahmood, T., Li, J., Rehman, A., Saba, T., & Mustafa, L. (2024). Transforming educational insights: strategic integration of federated learning for enhanced prediction of student learning outcomes. *The Journal of Supercomputing*, *80*(11), 16334-16367.

Johanyák, Z. C., Laufer, E., & Kovács, L. (2024, October). Fuzzy Logic-Based Preliminary Risk Assessment for a Supervised Federated Learning Solution in Predicting Student Results. In *2024 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)* (pp. 1-8). IEEE.

Kuang, Y., Liao, X., Jiang, Z., Gu, Y., Liu, B., Tan, C., ... & Li, K. (2025). Federated learning-based prediction of depression among adolescents across multiple districts in China. *Journal of Affective Disorders*, *369*, 625-632.

Nikolaidis, C. C., & Efraimidis, P. S. (2025). Advancing elderly social care dropout prediction with federated learning: client selection and imbalanced data management. *Cluster Computing*, *28*(2), 114.

Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting student dropout and academic success. *Data*, *7*(11), 146.

Surapaneni, P., Bojjagani, S., & Sharma, N. K. (2024, November). Federated Learning-based Big Data Analytics For The Education System. In *2024 International Conference on Intelligent Computing and Emerging Communication Technologies (ICEC)* (pp. 1-6). IEEE.

van Haastrecht, M., Brinkhuis, M., & Spruit, M. (2024, July). Federated learning analytics: Investigating the privacy-performance trade-off in machine learning for educational analytics. In *International Conference on Artificial Intelligence in Education* (pp. 62-74). Cham: Springer Nature Switzerland.

Automatic Distractor Generation in Multiple-Choice Questions Using Large Language Models with Expert-Informed Distractor Strategies

Yusei NAGAI[a*], Masaki UTO[a*]
[a]*The University of Electro-Communications, Japan*
*{nagai, uto}@ai.lab.uec.ac.jp

Abstract: In recent years, automatic generation of reading-comprehension questions with artificial intelligence has attracted considerable attention. In particular, producing high-quality distractors remains a critical challenge when generating multiple-choice questions (MCQs). Recent studies have increasingly employed large language models (LLMs) to generate distractors for MCQs. However, prior research has relied solely on the implicit, black-box knowledge of LLMs and has seldom exploited human expertise in distractor design. Therefore, in this study, we propose an LLM-based distractor-generation method that explicitly incorporates expert-informed distractor strategies, which represent typical heuristics used by human experts when crafting distractors. Experiments demonstrate that our method produces distractors of higher quality than those generated by previous approaches.

Keywords: Large language models, question generation, distractor generation, deep learning, natural language processing

1. Introduction

An effective way to develop reading comprehension skills is to provide learners with diverse reading passages accompanied by comprehension questions tailored to each passage. However, manually creating large numbers of such questions for various reading passages is time-consuming and labor-intensive. Recent studies have therefore explored the use of artificial intelligence to automatically generate reading comprehension questions (Chan et al., 2022; Tomikawa et al., 2024). Automatic question-generation methods can be used to create a variety of question types, among which the generation of multiple-choice questions (MCQs) is particularly prominent (Dutulescu et al., 2024; Gao et al., 2019; Maity et al., 2024; Shuai et al., 2023; Yu et al., 2024).

A crucial challenge in automatic MCQ generation is the generation of high-quality distractors. There are several criteria that high-quality distractors must generally satisfy. For example, (1) the distractors should not be easily identified and eliminated as incorrect options, and (2) the distractors should not be semantically equivalent or overly similar to the correct answer (Dutulescu et al., 2024). Therefore, creating distractors that satisfy these criteria is crucial for generating effective MCQs.

Numerous studies have been conducted with the aim of improving distractor quality in MCQs (Dutulescu et al., 2024; Gao et al., 2019; Maity et al., 2024; Shuai et al., 2023; Yu et al., 2024). For example, one approach generates candidate distractors using large language models (LLMs) and subsequently filters or ranks these candidates to remove inappropriate options (Dutulescu et al., 2024). Another approach involves multi-step prompting, in which distractors and questions are generated sequentially through structured prompts (Maity et al., 2024). However, these existing methods rely solely on the implicit, black-box knowledge embedded within LLMs without explicitly utilizing expert knowledge employed by human experts when creating distractors. Consequently, distractors produced

by these automatic methods may differ from those that human experts typically create, potentially reducing their appropriateness or effectiveness within specific contexts.

Therefore, in this study, we propose a novel distractor-generation method that explicitly incorporates human expert knowledge. Our approach first systematically organizes expert-informed distractor strategies, which reflect the common intentions and heuristics that human experts use when creating distractors. Next, our method uses a fine-tuned neural model to select the most suitable strategies for a given reading passage, question, and correct answer. Finally, an LLM guided by the selected strategies generates contextually appropriate distractors. Experiments show that our method yields higher-quality distractors than previous approaches.

2. Related Works

Dutulescu et al. (2024) proposed a method for automatically generating MCQs using LLMs that combines knowledge bases such as WordNet and DBpedia with the T5 model (Text-to-Text Transfer Transformer). Their approach generates candidate distractors, filters out inappropriate ones, and ranks the remaining candidates to select effective distractors. Yu et al. (2024) introduced a retrieval-augmented generation framework that incorporates external knowledge sources. Their method leverages the semantic relationships between words retrieved from knowledge bases to enhance the relevance and quality of generated distractors. Furthermore, Maity et al. (2024) proposed a multi-stage prompting method based on the chain-of-thought paradigm. In their approach, an LLM is prompted in four sequential stages: (1) paraphrasing the passage, (2) extracting keywords, (3) generating a question based on those keywords, and (4) generating distractors.

Although these methods have demonstrated improvements in distractor generation, they rely primarily on the implicit, black-box knowledge in LLMs. As noted in the Introduction, they do not explicitly incorporate expert-informed distractor strategies, which limits their ability to consistently produce distractors that align with those created by human experts.

3. Proposed Method

To overcome this limitation, the present study proposes an LLM-based distractor-generation method that explicitly leverages expert-informed distractor strategies. The proposed method comprises four steps: 1) classification of expert-informed distractor strategies, 2) construction of a strategy-labeled question dataset, 3) training of a strategy selection model, and 4) distractor generation. The first three steps constitute the preparation phase, while the final step corresponds to the actual distractor generation. Each of these steps is described in detail in the following subsections.

3.1 Classification of Expert-Informed Distractor Strategies

The initial preparation step involves systematically identifying and categorizing expert-informed distractor strategies by thoroughly reviewing prior research (Freedle and Kostin, 1991; Goodrich, 1977; King et al., 2004; Terao, 2019) that provided explicit guidelines or methodologies for distractor creation in MCQ design. Specifically, we carefully extracted, grouped, and synthesized these guidelines into distinct and coherent strategy categories. The resulting comprehensive set of strategies is presented in Table 1.

3.2 Construction of a Strategy-Labeled Question Dataset

The second preparation step involves annotating each distractor in an existing reading comprehension MCQ dataset with a corresponding strategy label based on the classification

Table 1. Expert-informed distractor strategies

	Strategy	Description
1	Use of Opposite Facts	Generating distractors that convey a meaning directly opposite to the correct answer.
2	Use of Irrelevant Facts	Generating distractors based on information unrelated to the topic or context of the reading passage and the correct answer.
3	Incorrect Combination of Facts	Generating distractors by incorrectly combining separate facts from different parts of the reading passage to create plausible yet incorrect options.

of expert-informed distractor strategies. The annotation process is conducted using Llama3, an open-source LLM, as described below.

Let an MCQ dataset be denoted as $\mathcal{D} = \{(c_n, q_n, a_n, \boldsymbol{d}_n) \mid n \in \mathcal{N}\}$, where c_n, q_n, a_n, and $\boldsymbol{d}_n$ represent a reading passage, a question, a correct answer, and a set of K distractors, respectively, for the n-th MCQ. The distractor set is defined as $\boldsymbol{d}_n = \{d_{nk} \mid k \in \mathcal{K}\}$ with $\mathcal{K} = \{1, \ldots, K\}$, and $\mathcal{N} = \{1, \ldots, N\}$ denotes the index set for MCQs with a size of N. To handle each distractor individually, we convert $\mathcal{D}$ into $\mathcal{D}' = \{(c_n, q_n, a_n, d_{nk}) \mid n \in \mathcal{N},\ k \in \mathcal{K}\}$. For each record in $\mathcal{D}'$, Llama 3 assigns the appropriate expert-informed distractor strategy, yielding the labeled dataset $\mathcal{D}_\mathrm{t} = \{(c_n, q_n, a_n, d_{nk}, y_{nk}) \mid n \in \mathcal{N},\ k \in \mathcal{K}\}$, where y_{nk} denotes the label of the expert-informed distractor strategy from Table 1 assigned to d_{nk}.

3.3 Construction of a Strategy Selection Model

The third preparation step is to construct a strategy-selection model that determines which strategies should be employed for a specific triple consisting of reading passage, question, and correct answer. We implement it as a classifier based on BERT (Bidirectional Encoder Representations from Transformers) The classifier receives the concatenated text of the reading passage, question, and correct answer as input and then outputs the label of the appropriate distractor-generation strategy. The classifier is trained on the dataset $\mathcal{D}_\mathrm{t}$.

3.4 Distractor Generation Leveraging Expert-Informed Strategies

After the above preparation steps, the proposed method generates distractors for an arbitrary reading passage c, question q, and correct answer a, using an LLM. Specifically, the strategy-selection model first predicts the most suitable distractor-generation strategy for the input triplet (c, q, a). Then, given the selected strategy y and the same triplet, a prompt is constructed using the template shown in Table 2, and this prompt is fed into Llama3 to generate distractors.

The prompt incorporates few-shot examples pertinent to the selected strategy, retrieved from $\mathcal{D}_\mathrm{t}$ by measuring their similarity to the input triplet $\{c, q, a\}$. The procedure for selecting similar examples is as follows:

1. The given triple (c, q, a) is input into SimCSE-BERT (Simple Contrastive Learning of Sentence Embeddings BERT) to obtain an embedding vector. Embedding vectors for all samples in $\mathcal{D}_\mathrm{t}$ are also precomputed in the same manner.
2. The cosine similarity between the embedding vector of the input triplet and those of the samples in $\mathcal{D}_\mathrm{t}$ labeled with the selected strategy y is computed. The three most similar samples $\{(c_i^{(\mathrm{e})}, q_i^{(\mathrm{e})}, a_i^{(\mathrm{e})}, d_i^{(\mathrm{e})}) \mid i \in \{1, 2, 3\}\}$ are chosen as few-shot examples.

4. Experiments

This section describes the experiments conducted to evaluate the performance of the proposed method. Our experiments used the RACE dataset (Lai et al., 2017), a well-known dataset for reading-comprehension MCQs. Each question comprises a passage, a question,

Table 2. Prompt template for generating distractors

You are an expert in creating multiple-choice questions for reading comprehension. You are provided with a set consisting of reading passage, a question, and a correct answer. The question and correct answer are both related to the content of the reading passage. Your task is to create a distractor (an incorrect option) based on the reading passage, question, and correct answer according to the following strategy. {*The description of expert-informed distractor strategy* y} During generation, please follow the steps below: 1. Understand the above strategy thoroughly. 2. Carefully read and understand the provided reading passage, question, and correct answer. 3. Generate a distractor following the strategy, considering the given reading passage, question, and correct answer. 4. Output ONLY the distractor you generated. The input passage, question, and correct answer are given below: ### Reading passage: c, Question: q, Answer: a
The following are the 3-shot examples (for $i \in \{1, 2, 3\}$): ### Reading passage: $c_i^{(e)}$, Question: $q_i^{(e)}$, Answer: $a_i^{(e)}$, Distractor: $d_i^{(e)}$

the correct answer, and three distractors. The experimental procedure was as follows:

1. We split the RACE training dataset into a training set $\mathcal{D}_{\text{train}}$ and a test set $\mathcal{D}_{\text{test}}$ in a 9:1 ratio, then constructed $\mathcal{D}_{\text{t}}$ from $\mathcal{D}_{\text{train}}$ following the procedure in Section 3.2.
2. We trained the strategy selection model following the procedure in Section 3.3.
3. For each of 100 randomly selected instances from $\mathcal{D}_{\text{test}}$, the proposed method generated one distractor according to the procedure in Section 3.4.
4. The generated distractors were evaluated manually on the following two criteria.
   - **Incorrectness:** This criterion indicates whether the option is actually incorrect. We assigned a score of 1 when the option was indeed incorrect; otherwise, the score was 0.
   - **Plausibility:** This criterion indicates whether the option cannot be dismissed without reading the passage. We assigned a score of 1 when recognizing the option as incorrect required proper comprehension of the passage; otherwise, the score was 0.
5. For comparison, we conducted the same experiments on the following two methods:
   - **Baseline:** A method that generates distractors without providing either expert-informed distractor strategies or few-shot examples
   - **Proposed Variant:** A variant of the proposed method in which distractors are generated using randomly selected few-shot examples, while the optimal strategy is still selected by the strategy selection model.

Table 3 presents the mean and standard deviation (in parentheses) of each score across the 100 generated distractors. The table shows that the proposed method obtained the highest values for both evaluation criteria. To determine whether these differences were statistically significant, we compared the methods using the Kruskal–Wallis test. The test revealed a significant difference in plausibility at the 1% significance level. Subsequent Dunn post-hoc tests with Holm correction confirmed significant plausibility differences (1%) between the proposed method and both the baseline and the proposed variant. These results indicate that the proposed method with appropriate few-shot examples produces higher-quality distractors that serve as effective and misleading options.

Table 3. Experimental results

Criteria	Proposed	Baseline	Proposed Variant
Incorrectness	**0.97** (0.17)	0.93 (0.25)	0.95 (0.21)
Plausibility	**0.56** (0.49)	0.20 (0.40)	0.32 (0.46)

5. Analysis

For a more detailed analysis, we assessed the generated distractors using item response theory (IRT), a family of probabilistic models widely used in educational and psychometric measurement to analyze the relationship between an examinee's latent ability and their item responses. IRT enables the estimation of item parameters, such as difficulty and discrimination, as well as examinee abilities on a common scale, offering more precise measurement than classical test theory. Based on this IRT framework, along with a virtual-examinee approach (Tomikawa et al., 2024; Uto et al., 2024), we conducted the following evaluation to assess the generated distractors:

1. We first built 59 virtual examinees as question-answering (QA) systems, following Tomikawa et al. (2024). Specifically, they were constructed using various pretrained neural models trained on the RACE validation dataset with different sample sizes, so that their accuracies spanned a wide range. We then collected their correct and incorrect responses on MCQs in $\mathcal{D}_{\text{train}}$, and estimated their abilities with the Rasch model, a widely used model in IRT.
2. For each record in $\mathcal{D}_{\text{test}}$, we generated three distractors: one using the baseline method, one using the proposed method, and one using the proposed variant method. Treating each set of three distractors together with the corresponding reading passage, question text, and correct answer as a single question, we collected responses from the 59 virtual examinees for each.
3. We converted the responses to distractor-level data, treating every distractor as an individual item and each response as an indicator of whether the examinee selected it. Using these data, we fitted the two-parameter logistic model, another widely-used IRT model, by fixing examinee abilities to their Rasch estimates, and obtained the discrimination parameter for each distractor.

Figure 1 plots the average absolute discrimination against the correct ratio threshold: the x-axis gives the maximum correct ratio (e.g., $x = 0.5$ covers questions with a correct-answer rate < 0.5), and the y-axis shows the mean absolute discrimination for the distractors of the corresponding questions. Note that we reported the absolute discrimination values, whereas all discrimination estimates were originally negative values because higher-ability examinees were less likely to choose distractors. The three curves correspond to the distractors produced by the baseline, the proposed method, and the proposed variant, respectively.

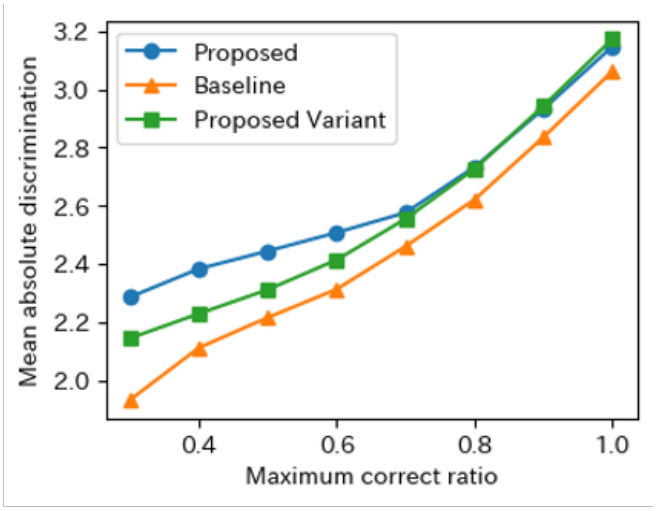


Figure 1. Discriminatory power of the generated distractors

The results show that distractors generated by the proposed method achieve the highest absolute discrimination, especially on difficult questions (i.e., those with low correct-answer rates). For easier questions, where the correct-answer rate is high, the gap between methods

narrows because distractors play a smaller role. Overall, these findings confirm that the proposed method produces distractors that better differentiate examinees' abilities, enabling automatically generated MCQs to diagnose reading-comprehension proficiency more accurately and to yield finer-grained ability estimates.

6. Conclusion

In this study, we proposed a method for generating high-quality distractors for multiple-choice reading-comprehension questions. Human evaluation and IRT analysis demonstrated that our approach yields distractors with higher quality and greater discriminatory power. Future work will proceed along three directions. First, because our experiments were limited in scale and detail, we will evaluate the method on more diverse datasets in greater detail. Second, the current taxonomy of distractor strategies remains coarse; therefore, we plan to refine it into finer subtypes and enable it to discover additional strategies automatically. Third, although the human evaluation currently relies on a binary scoring system, this may lack the granularity needed to capture nuanced differences in distractor quality. Future work will adopt a more fine-grained scoring rubric to help uncover deeper insights.

References

Chan, Y.-H., Chung, H.-L., & Fan, Y.-C. (2022). Keyword provision question generation for facilitating educational reading comprehension preparation. Proceedings of the 15th International Conference on Natural Language Generation, 196–202.

Dutulescu, A., Ruseti, S., Iorga, D., Dascalu, M., & McNamara, D. S. (2024). Beyond the obvious multi-choice options: Introducing a toolkit for distractor generation enhanced with NLI filtering. Proceedings of the Artificial Intelligence in Education, 242–250.

Freedle, R., & Kostin, I. (1991). The prediction of SAT reading comprehension item difficulty for expository prose passages. ETS Research Report Series, 1–52.

Gao, Y., Bing, L., Li, P., King, I., & Lyu, M. R. (2019). Generating distractors for reading comprehension questions from real examinations. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 6423–6430.

Goodrich, H. C. (1977). Distractor efficiency in foreign language testing. TESOL Quarterly, 69–78.

King, K. V., Gardner, D. A., Zucker, S., & Jorgensen, M. A. (2004). The distractor rationale taxonomy: Enhancing multiple-choice items in reading and mathematics.

Lai, G., Xie, Q., Liu, H., Yang, Y., & Hovy, E. (2017). RACE: Large-scale reading comprehension dataset from examinations. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 785–794.

Maity, S., Deroy, A., & Sarkar, S. (2024). A novel multi-stage prompting approach for language agnostic MCQ generation using GPT. Proceedings of the 46th European Conference on Information Retrieval, 268–277.

Shuai, P., Li, L., Liu, S., & Shen, J. (2023). QDG: A unified model for automatic question-distractor pairs generation. Applied Intelligence, 53, 8275–8285.

Terao, T. (2019). Eigo bunsho dokkai koumoku ni okeru sakuranshi no sentakuritsu: Jukensha no tenkeiteki na goto ni chakumoku shite (distractor selection rates in English reading comprehension items: Focusing on examinees' typical errors) [Doctoral dissertation, Nagoya University] [(In Japanese)].

Tomikawa, Y., Suzuki, A., & Uto, M. (2024). Adaptive question–answer generation with difficulty control using item response theory and pretrained transformer models. IEEE Transactions on Learning Technologies, 17, 2186–2198.

Uto, M., Suzuki, A., & Tomikawa, Y. (2024). Question difficulty prediction based on virtual test-takers and item response theory. Workshop on Automated Evaluation of Learning and Assessment Content.

Yu, H. C., Shih, Y. A., Law, K. M., Hsieh, K., Cheng, Y. C., Ho, H. C., Lin, Z. A., Hsu, W.-C., & Fan, Y.-C. (2024). Enhancing distractor generation for multiple-choice questions with retrieval augmented pretraining and knowledge graph integration. Proceedings of the Association for Computational Linguistics, 11019–11029.

The Construction of Evaluation Indicators for Intelligent Educational Software and the Measurement of Its Developmental Level

Bo ZHANG[a*], Yuanyuan LI[b], Qianchao XING[c], Yuhan CHEN[d] & Xinying WANG[e]
[a]*XJTLU Entrepreneur College (Taicang), Xi'an Jiaotong-Liverpool University Bo ZHANG, China*
[b]*Shanghai Normal University Yuanyuan LI, China*
[c]*Shanghai Normal University Qianchao XING, China*
[d]*Minhang Yongde School Attached to East China Normal University Yuhan CHEN, China*
[e]*Shanghai Huangpu District Labor and TechnicalEducation Centre, China*
*Bo ZHANG@corresponing.author

Abstract: The evaluation revealed notable disparities across educational scenarios, with strengths in technical quality and user experience but persistent shortcomings in functionality completeness, business alignment, and attractiveness. In 2024, intelligent educational software in China is transitioning from the early phase to more advanced and optimized stages, yet it remains distant from full maturity. By establishing a systematic evaluation framework, this study offers an empirical basis for guiding the high-quality development of intelligent educational software and promoting the broader digital transformation of education.

Keywords: Intelligent educational software, maturity model, evaluation indicator, level measurement

1. Introduction

With the rapid development of generative AI, big data, cloud computing, and the metaverse, educational software has increasingly acquired intelligent features, becoming a critical medium for integrating AI into education. This study defines such tools as intelligent educational software (IES)—applications that leverage technologies such as IoT, big data, AI, and 5G to improve learning performance, optimize teaching, and support diverse educational scenarios (Wang, Li & Yan, 2024). As a key element of "soft" infrastructure, IES plays a central role in advancing educational digital transformation. However, existing assessments of such software remain largely descriptive, lacking systematic and quantitative evaluation frameworks. Current studies typically evaluate educational software along educational and technical dimensions—for example, through usability, functionality, and student engagement—but face persistent challenges: reliance on subjective judgment, poorly structured indicator systems, limited applicability beyond teaching scenarios, insufficient attention to intelligent features, and a lack of empirical validation (Zhang, Goodman & Gu, 2022). To address these gaps, this study reviewed domestic and international evaluation frameworks and identified four key indicators tailored to IES: Functional Integrality, Requirement Matching, Functional Realization, and Charisma Attribute.

Based on these, the Evaluation Indicators System for IES (2024) was developed through collaboration with educators, software developers, and frontline teachers. Using this system, 503 IES products were evaluated and their developmental stages based on the Maturity Model of Digital Transformation for Education (Zhu, Zhang & Dai, 2024). This research aims to advance the transformation and upgrading of IES, provide empirical evidence for optimizing educational practices, and offer insights for implementing China's educational digital transformation strategies.

2. The Construction of Evaluation Indicators

2.1 Design of Indicators

The central role of IES in China's digital transformation lies in embedding AI technologies into diverse educational scenarios. Existing products provide support for personalized learning, smart classrooms, intelligent campus management, and data-driven assessments, thereby empowering three main stakeholders—students, teachers, and administrators (Zhang, et al., 2024). Educational scenarios have gradually shifted from peripheral management contexts to core teaching and learning environments, creating a dynamic system characterized by personalization, intelligence, and ubiquity. In educational research (Yang, et al.,2022), scenarios are typically categorized into learning, teaching, management, and evaluation. Building on this classification, the Evaluation Indicators System for IES (2024) identifies four first-level dimensions: student learning, teacher instruction, educational administration, and teaching evaluation. Under these dimensions, 16 specific educational scenarios are further defined as second-level indicators (Figure 1).

2.2 Design of Indicators

Drawing on Garrett's user experience model, the evaluation framework incorporates four primary indicators—Functional Integrality (FI), Functional Realization (FR), Requirement Matching (RM), and Charisma Attribute (CA)—to assess the effectiveness of Intelligent Education Systems (IES) *(Garrett, 2007)*. FI examines whether IES provides the core functions required across diverse educational scenarios, with Core Functions (CF) serving as a secondary indicator to capture scenario-specific needs. FR evaluates both functional effectiveness and sustained user adoption, operationalized through the Technology Quality Indicator (TQI), which assesses functionality, reliability, usability, maintainability, efficiency, and portability, and the Experience Quality Indicator (EQI), which emphasizes prioritization, consistency, and resilience in user experience. RM measures the degree to which IES functionalities align with scenario-specific requirements, with Educational Scenario Requirements (ESR) ensuring that software functions are systematically developed to match practical educational needs. Finally, CA addresses innovation potential by considering Advanced Function (AF), which evaluates the novelty of functionalities, and Practice Innovation (PI), which examines whether these functionalities generate new teaching or learning practices. Overall, the framework integrates four first-level dimensions (student learning, teacher instruction, educational administration, and teaching evaluation), 16 second-level scenarios, and the four primary indicators (FI, FR, RM, and CA), each supported by corresponding secondary indicators (Figure 2).

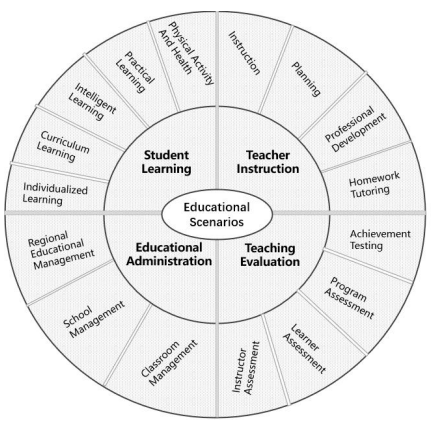


Figure 1. Evaluation Indicators for IES (a)

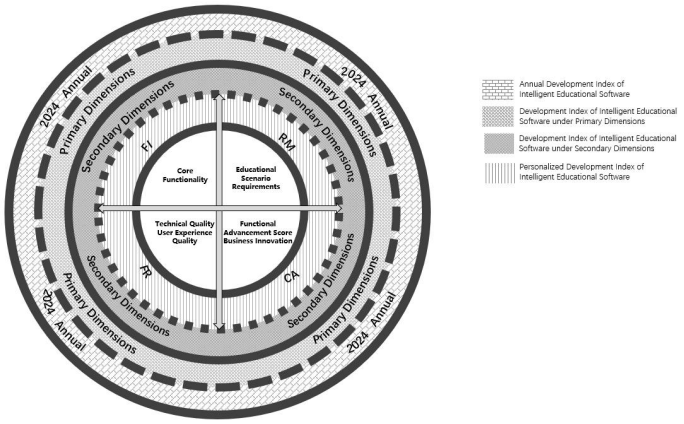


Figure 2. Evaluation Indicators for IES (b)

2.3 Design of Indicator Weights

The Delphi method was employed to determine indicator weights, involving multiple rounds of consultation with frontline educators, software developers, and education experts (Table 1). As this research represents an initial baseline investigation, the Charisma Attribute (CA) was assigned relatively lower weight, while priority was placed on indicators reflecting functional effectiveness and scenario alignment.

Table 1. *Weight Design of Evaluation Indicators for IES*

Primary Dimension	*Secondary Dimension*	*Primary Indicator*	*Secondary Indicator*
Equally Distributed Weights	Equally Distributed Weights	FI, 0.35	CF, 1
		FR, 0.2	TQI, 0.5
			EQI, 0.5
		RM, 0.4	ESR, 1
			AF, 0.5
		CA, 0.05	PI, 0.5

2.4 Quantitative Assessment of Software Scores

The scores of the four primary indicators were first computed and subsequently converted into a percentage scale. Specifically, n denotes the number of core functionalities, m the number of educational scenario requirements, and o the number of charisma attributes. The value ranges for the secondary indicators were defined as follows: CF = 0 or 1; DQI, EQI, ESR = 0 −4; AF, PI = 0−2. Quantitative scoring for each secondary indicator was based on either a three-point or five-point Likert scale.

Step 1. Calculation of primary indicator scores.

$$FI = \frac{\sum_{i=1}^{n} CF_i}{n} \qquad FR = \frac{\sum_{i=1}^{6} TQI_i + \sum_{i=1}^{3} EQI_i}{36} \qquad RM = \frac{\sum_{i=1}^{m} ESR_i}{4m} \qquad CA = \sum_{i=1}^{o} \frac{AF + PI}{2}$$

Step 2. Dimension-level aggregation.

The scores of the four primary indicators under each dimension were calculated using a hierarchical iterative approach.

Step 3. Calculation of the Development Level of IES (DLIES).The final development level scores for each dimension were derived using the following formula:

$$DLIES_i = FI_i * 35\% + FR_i * 20\% + RM_i * 40\% + CA_i * 5\%$$

where i = a, b, c; a represents the year, b denotes the primary dimension, and c corresponds to the secondary dimension.

2.5 Delimitation of Development Stages

Building on expert ratings of development levels, this study further delineates the maturity stages of IES within China's broader educational digital transformation. Drawing on a maturity model of educational digital transformation (Zhu, Zhang & Dai, 2024), a five-level maturity framework was constructed, with score ranges of 0–50, 50–60, 60–70, 70–80, and 80–100, corresponding respectively to the stages of Cognitive, Initial, Intermediate, Optimized, and Mature.

At the Cognitive stage, software lacks core functionalities, exhibits very limited use of intelligent technologies, and shows weak requirement alignment, low realization quality, and no evident innovative attributes. The Initial stage is characterized by the presence of basic functionalities, though integration of intelligent technologies remains limited and software performance across indicators is generally poor. In the Intermediate stage, software demonstrates moderate functional completeness, introduces basic intelligent technologies, and achieves average realization quality and requirement matching, with initial signs of innovative attributes. The Optimized stage reflects substantial functional completeness, widespread application of intelligent technologies, high realization quality, and strong

requirement alignment, with innovation becoming increasingly visible. Finally, at the Mature stage, software exhibits comprehensive core functionalities, effective integration of advanced intelligent technologies, and outstanding performance across all dimensions, with innovation widely and consistently achieved.

3. Investigation into the developmental level of IES

IES should not be viewed merely as a product of intelligent technologies, but rather as a developmental stage in the evolution of educational software toward intelligence. To evaluate the current status, this study identified software registered on the Ministry of Education's Educational Software (APP) Filing Platform (http://app.eduyun.cn) as primary evaluation targets, given its regulatory oversight. Additional widely used and market-recognized applications were also included. Selection criteria required that software: (1) allow registration and trial use; (2) apply to primary education; and (3) integrate intelligent functional features. After screening 3,080 applications, 503 were retained for analysis. These applications were evaluated using the Evaluation Indicators System for IES (2024) and the proposed maturity framework, enabling assessment of both development levels and maturity stages.

3.1 Analysis of Developmental Levels under Secondary-Dimension Scenarios

Evaluation scores across secondary-dimension scenarios ranged from 54.07 to 72.26 (Table 4). Educational administration within school management achieved the highest performance (Optimized stage), while curriculum learning in student learning recorded the lowest (Initial stage), reflecting weak classroom integration. Most scenarios clustered at the Intermediate stage, including homework tutoring, assessment, online instruction, and teacher professional development, showing that intelligent software is beginning to support core teaching and evaluation practices. By contrast, applications concentrated in student learning remain largely at the Initial stage, underscoring substantial gaps but also significant potential for future innovation.

Table 4. *Developmental Levels of IES Across Different Dimensional Educational Scenarios*

Educational Scenario under Secondary Dimensions		***FI***	***FR***	***RM***	***CA***	***Evaluation Score for Development Level***	***Developmental Stage***
Educational Administration	School Management	77.62	75.74	71.89	23.74	72.26	Optimized
	Classroom Management	73.44	77.44	70.11	22.95	70.38	Optimized
Teacher Instruction	Homework Tutoring	65.32	90.73	67.90	14.43	68.89	Intermediate
	Instruction	62.11	89.54	64.05	10.88	65.81	Intermediate
	Professional Development	58.98	88.78	61.06	12.28	63.44	Intermediate
Teaching Evaluation	Planning	56.64	88.06	58.62	10.00	61.38	Intermediate
	Learner Assessment	69.48	83.07	68.13	0.00	68.18	Intermediate
	Achievement Assessment	63.97	86.34	59.56	9.76	63.97	Intermediate
Student Learning	Intelligent Learning	57.35	85.97	48.90	14.90	57.57	Intermediate
	Individualized Learning	54.61	85.01	46.59	12.95	55.40	Initial
	Practical Learning	53.95	84.81	45.90	12.43	54.83	Initial
	Physical Activity and Health	53.72	84.77	45.65	12.38	54.63	Initial
	Curriculum Learning	52.79	84.51	45.10	12.93	54.07	Initial

Educational Scenario under Primary Dimensions	*FI*	*FR*	*RM*	*CA*	*Evaluation Score for Development Level*	*Developmental Stage*
Educational Administration	75.53	76.59	71.00	23.34	71.32	Optimized
Teacher Instruction	60.76	89.28	62.91	11.89	64.88	Intermediate
Teaching Evaluation	66.73	84.71	63.84	4.88	66.08	Intermediate
Student Learning	54.48	85.01	46.43	13.12	55.30	Initial
2024	64.38	83.90	61.05	13.31	64.39	Intermediate

3.2 Analysis of Development Levels of IES in Primary-Level Educational Scenarios

As shown in Table 4, Intelligent Education Systems (IES) varied notably across scenarios. Educational administration achieved the highest score (71.32, Optimized), followed by teaching evaluation (66.08) and teacher instruction (64.88), both at the Intermediate stage. Student learning, however, scored lowest (55.30), remaining at the Initial stage. This reveals a clear disparity: administration and evaluation are relatively advanced, instruction is mid-level, and student learning lags behind.

Administration software recorded high FI and RM scores and the strongest CA performance, but its FR score was the lowest. This suggests bold but uneven experimentation, reflecting both progress and challenges given the complexity of administrative functions. Teacher instruction tools scored moderately in FI and RM but high in FR, indicating efficiency gains in areas such as homework tutoring, though innovation remains limited. Teaching evaluation software showed similar patterns, with strong FR yet modest CA, pointing to partial digital transformation but insufficient integration with evaluative needs.

Student learning software performed relatively well in FR but weakly in FI, RM, and CA. Given the complexity of learning processes, effective design requires both adaptive resources and interactive tools, alongside cautious integration of intelligent technologies. Despite current limitations, this domain holds the greatest potential for future development.

3.3 Annual Analysis of Development Levels of IES

Evaluation of 503 software products indicates that in 2024, the overall development level of IES reached a score of 64.39, placing it within the “Intermediate” stage. FI and RM scored 64.38 and 61.05 respectively, while FR and CA scored 83.90 and 13.31. These findings highlight several trends. First, software design has begun to address the core functional demands of educational practice, though greater precision is needed in identifying user requirements and improving iterative development. Second, while the technical and experiential quality of functionalities generally satisfies educational stakeholders, substantial scope remains for enhancing FI and RM to better align with actual practice. Third, innovation in CA is particularly limited, signaling the need for greater emphasis on charisma attributes to ensure wider acceptance and sustained engagement. Overall, IES has transitioned beyond the initial stage, yet remains far from maturity.

4. Conclusions and Recommendations

This study applied a 2024 evaluation framework comprising four indicators—FI, FR, RM, and CA—to assess 503 IES products across diverse educational scenarios. Three key findings emerged. First, significant disparities exist: educational administration software reached the Optimized stage, teaching evaluation and teacher instruction achieved the Intermediate stage, while student learning software remains at the Initial stage. Second, although FR and user experience scores were relatively strong, FI and RM barely passed, reflecting limited functional coverage and weak stakeholder alignment. Third, overall IES development has

advanced beyond early phases but has not yet reached maturity. Based on these findings, three recommendations are offered.

4.1 Promote High-Quality and Balanced Development of IES

IES supports multiple domains—student learning, teacher instruction, evaluation, and administration. It facilitates personalized learning, instructional efficiency, evidence-based evaluation, and administrative effectiveness. However, development is uneven: administration has advanced rapidly, evaluation and instruction show moderate progress, while student learning lags significantly. Achieving high-quality digital transformation requires prioritizing student learning software, while sustaining innovation in instructional and evaluative tools.

4.2 Strengthen Core Functionalities and User Alignment

Although technical reliability is adequate, insufficient FI and RM scores indicate misalignment between software capabilities and user needs. Governments should establish structured feedback channels between educators and developers to ensure context-relevant functionalities. Researchers should ground design in educational theory and scenario-specific requirements, while developers should enhance usability, simplify operations, and integrate coherent functions. The consistently low CA scores further highlight the need for scenario-adaptive innovation and more intelligent, pedagogy-driven applications.

4.3 Foster Multi-Stakeholder Collaborative Innovation

Advancing IES from intermediate or optimized stages to maturity requires coordinated efforts from government, industry, academia, and schools. Governments should refine regulatory frameworks and provide adoption incentives. Developers should prioritize core functions and iterative improvement in close collaboration with educators. Researchers should generate theoretical insights and practice-oriented feedback, while schools should foreground authentic classroom needs. Building dynamic, multi-stakeholder feedback loops will accelerate innovation, deepen educational transformation, and ensure the sustainable evolution of IES.

Acknowledgements

This work was supported by the XJTLU Research Development Fund Project (Grant No. RDF-24-02-021)

References

Garrett, J., “The Elements of User Experience: User-Centered Web Design, ” Beijing: Mechanical Industry Press, 2007, pp. 34–153.

Wang, C., Li, Y., & Yan, K. (2024). Key Scenarios and Promotion Strategies for Intelligent Education Software Promoting Future Education Development: Excerpt from the 2023 Annual Report on Artificial Intelligence Promoting the Development of Education (Ⅶ). *Chinese Journal of ICT in Education, 30*(07), 73-82.

Zhang, B., Goodman, L., & Gu, X. (2022). Novel 3D contextual interactive games on a gamified virtual environment support cultural learning through collaboration among intercultural students. SAGE Open, 12(2), 21582440221096141.

Zhang, B., Chen, Y., Wang, X., and Li, Y., “Construction of an Evaluation Index System for Intelligent Educational Software Oriented to Student Learning: Excerpt from the 2023 Annual Report on Artificial Intelligence Promoting the Development of Education (Ⅴ),” Chinese Journal of ICT in Education, vol. 30, no. 07, pp. 54-63, 2024.

Zhu, Z., Zhang, B., & Dai, L. (2024). The change and the unchanged in education empowered by digital intelligence technology. *Chinese Journal of ICT in Education*, 30(3), 3-14.

Zhang, B., Goodman, L., & Gu, X. (2023). Telecollaboration tool preferences for online intercultural learning in higher education: Perspectives of Chinese international students. Sage Open, 13(2), 21582440231180087.

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

A Study of Precise Instructional Intervention Strategies Based on Learner Portraits

Jingyi Zhang[a*], Xinyu Sun[a], Nan Dang[a], Meng Li[a], Minsheng Fan[b], Huilong Li[a] & Weitong Guo[a]
[a]*Institute of Educational Technology*
Northwest Normal University
Lanzhou, China
[b]*Key Laboratory of Education Digitalization of Gansu Province*
Lanzhou, China
*Minsheng Fan@corresponing.author

Abstract: Personalized education is a key trend in educational innovation, but traditional teaching often fails to meet individual student needs. The development of big data and artificial intelligence has enabled learner profiling, which can support precise teaching. This study explores precise teaching interventions based on learner profiles to improve teaching quality and learning effectiveness. It provides new ideas for personalized teaching and is significant for enhancing educational outcomes.

Keywords: Precision teaching, Instructional interventions, Intervention strategies, Learner profiles

1. Introduction

With the advancement of information technology, education has transformed profoundly, and society's expectations for education have evolved. The information age and national development demand a new talent cultivation model. China's "Core Literacies for Student Development" has outlined six major literacies that students should possess, guiding talent cultivation. The innovative power of digital technology is deeply integrating information technology with learning and teaching, giving rise to a new data- and technology-driven model. In 2023, the U.S. Department of Education's Office of Educational Technology released its policy report, "Artificial Intelligence and the Future of Teaching and Learning" (2023), emphasizing AI's role in education and its potential to address student learning variability. In 2025, the Outline of the Plan for the Construction of a Stronger Education Nation (2024–2035) proposed using AI to drive educational change and establish a big data- and AI-supported evaluation and decision-making system. In the Intelligent Era, education now focuses on personalized learner growth, with data analysis technology becoming crucial for implementing precise teaching. By deeply analyzing group and individual student data to build learner portraits, we can reveal inherent group learning rules and accurately grasp each student's personalized needs.

This study uses intelligent learning technology to create precise teaching interventions based on learner profiles in an English course. By collecting and analyzing student data (basic info, progress, and behavior) in a smart learning environment, we build detailed group and individual learner portraits. These portraits help identify learning issues, support precise teaching strategies, and guide targeted interventions to improve teaching quality and meet individual student needs.

2. Definition of relevant concepts

2. 1 Learner Profile

Portraits were initially used to describe appearance, and later evolved into user portraits in the business field, proposed by Cooper(Cooper A,2004), which analyze user behavior through big data for product positioning and personalized recommendation. With the penetration of information technology in the field of education, learner profiling came into being, which is an expansion of user profiling in the field of education, and analyzes learner characteristics through visualization. In the era of big data, learner profiling technology mines learners' learning needs, motivations and preferences, and shows learning behaviors in a visual way, helping learners master their own learning conditions, providing a basis for personalized learning recommendations, and providing teachers with personalized teaching guidance.

Currently, academia lacks a unified definition of "learner portrait." From a teaching service perspective, some researchers view it as evaluating or characterizing learners. For example, Cui Jiafeng sees it as an intelligent learning evaluation system, Zhao Linglang et al. consider it a method for analyzing learners' big data using machine learning, and Dinh et al. define it as depicting the behavior of virtual learning communities. In essence, a learner portrait describes the learner's current state and summarizes their characteristics. In this study, the learner portrait is based on the learning analytics framework, aiming to provide personalized learning support by collecting, analyzing, and presenting learner behavior patterns through labels.

2.2 Precision teaching

In the 1960s, American education pioneer Ogden Lindsley first introduced "precision teaching," which uses standard variable speed charts to make educational decisions based on observed changes in student behavior frequency (Binder C, 1998). In China, Prof. Zhu Zhiting's team emphasizes combining precision teaching with information technology to enhance efficiency through human-machine collaboration. Others like Hao Jianjiang and Guo Jiong focus on precision teaching supported by intelligent technology, highlighting the importance of precise design and implementation of teaching objectives, evaluation, and activities.

In summary, precision teaching leverages intelligent technology, big data, and information technology to set precise teaching goals, conduct intelligent evaluations, and design personalized activities. Despite different definitions, the common goal is to improve teaching efficiency and quality through technology while emphasizing the role of teachers and teaching innovation. This approach integrates educational concepts and practices, offering a new direction for modern education.

3. Constructing and Analyzing Learner Profiles

Smart classrooms integrate advanced technologies to enhance teaching and learning, enabling instant communication and real-time data storage and retrieval. These classrooms collect data via digital tools like interactive whiteboards, tablets, and learning management systems. The collected data is used to create comprehensive learner portraits, which help diagnose learning issues and design targeted interventions. This supports individual learning needs and promotes a more effective, interactive educational experience.

The learner portrait modeling system proposed in this study is one of the constituent elements in the field of educational data mining. At the practical level, the construction and application of learner portraits need to follow a systematic process framework as support. According to the principles of pedagogy and data science methodology, this study designs a complete set of construction process system, whose technical route contains three core links of feature extraction, model validation and dynamic optimization, and the specific implementation steps are detailed in Figure1.

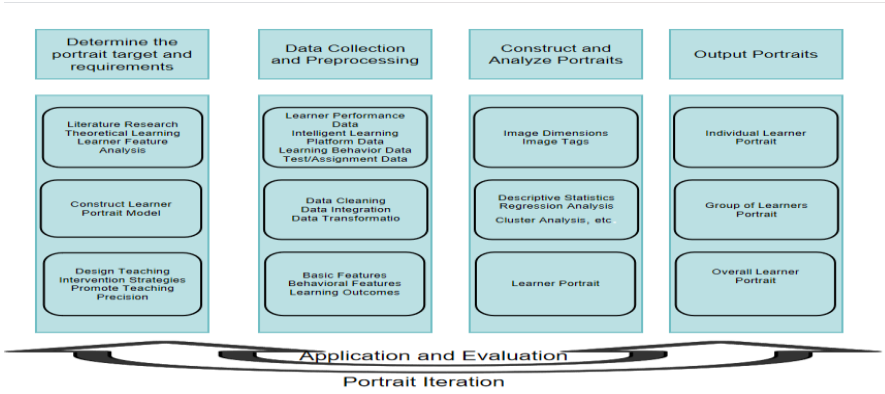


Figure. 1 Learner Profile Construction Process

In this study, we applied K-means clustering to categorize learners across four dimensions: engagement, interaction, learning, and knowledge acquisition. Three distinct learner types emerged:

1.Excellent Learners: These students excel in learning and knowledge acquisition but show limited participation and interaction. To enhance their engagement, strategies should focus on integrating them into classroom interactions and collaborative learning.

2. Diligent Learners: Despite positive behaviors, these students often struggle with competence and participation, leading to lower accuracy in assignments and exercises. Providing them with more challenging tasks and targeted learning strategies can improve their efficiency and quality.

3. Marginal Learners: Low performance and engagement; they need personalized support to boost motivation and improve participation and competence.

These findings are consistent with broader research showing that K-means clustering effectively identifies distinct performance levels and informs targeted interventions.

4.1 Types of interventions

Instructional interventions, categorized by the Response to Intervention (RTI) framework, include three tiers: universal, targeted, and personalized. Universal interventions address common issues for all students through adjusted teaching strategies and optimized resources, focusing on basic knowledge reinforcement and general learning difficulties. Targeted interventions focus on specific groups, addressing common problems through collaborative learning and group discussions to promote communication and problem-solving. Personalized interventions provide tailored support for individual students to meet unique learning needs and solve specific problems. Additionally, interventions can target cognitive skills, enhancing thinking and problem-solving abilities; behavioral habits, improving active learning attitudes; and competencies, boosting overall learning and communication skills.

4.2 Elements of intervention

During the practice of precision teaching, the classroom becomes the center point of attention for both teachers and students, and in doing so, it aims to promote the holistic development of learners. In order to effectively match instructional elements with students' individual characteristics, this study developed appropriate instructional interventions based on the findings. Incorporating elements of instructional design and classroom structure, instruction is designed around student characteristics with precise objectives, tasks and activities to motivate learning. Resources and tools need to accurately match students' needs, and evaluation methods should integrate process and summation, and clarify multiple evaluation subjects. This study clearly defines the core elements of precise instructional interventions, as shown in Table 1.

Table 1. *Elements of Precision Instructional Interventions*

Elements of intervention	concrete content
goal	Identify students' learning goals and assist them in developing individualized independent learning progress plans.
mandates	Based on the results of learner profiling, differentiated learning tasks and requirements are assigned to groups of learners with similarities and groups of learners with differences.
maneuver	Teachers focus on teaching, group cooperative inquiry, individual independent training, individualized tutoring and guidance, learning tips and information, supervision and assessment feedback.
resource (such as manpower or tourism)	Produce resources for students' varied learning preferences and push relevant resources based on learning.
estimation	Students are assessed based on the accuracy of their answers, presentation and communication of their reports, and their behavior and conduct.

In precision teaching, intervention elements are interrelated and mutually reinforcing. Learning objectives must be personalized, considering individual learner characteristics, serving as both the starting and end point of learning. These objectives significantly influence task difficulty and guide resource and evaluation system construction. Learning tasks should closely align with activities to promote learner participation and ability enhancement. Precise matching of resources and tools is crucial. Each intervention element requires tailored strategies to ensure learning continuity and integrity, optimizing teaching and learning outcomes.

4.3 Framework for intervention

Accurate teaching is based on in-depth analysis of learner characteristics, and learner profile is the key element, the construction of which covers the collection and analysis of multi-dimensional data, including basic information, behavior, knowledge mastery and ability level. These data are presented through the visualization technology to visualize the learning situation, identify problems, and provide a basis for teaching. This data-driven accurate teaching strategy can effectively improve teaching effectiveness and learner development.

Zhu Zhiting and Peng Hongchao constructed an information technology-enabled precision teaching model in a smart classroom environment, centered on the RTI instructional intervention model and implementing interventions in three progressive levels. The model screens and imposes interventions step by step in the teaching process to ensure accurate teaching (as shown in Figure 2.).

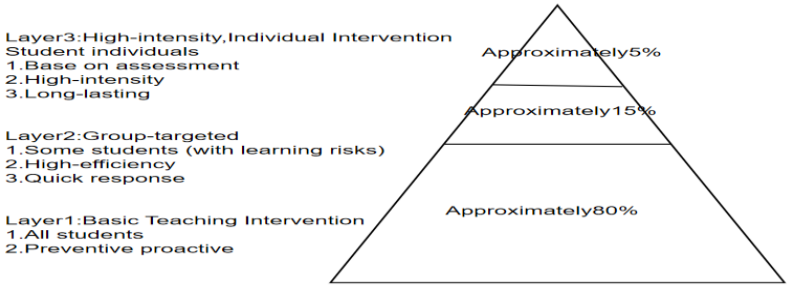


Figure 2. Intervention-Response Model

The ADDIE model is a classic instructional design model widely used in the field of teaching and learning, which, as a systematic instructional design framework, comprehensively integrates the major elements of instructional design, covering both the learning subject (teachers and students) and the learning object (instructional content, instructional methodology, instructional resources, and assessment of learning, etc.)(Liu Caiyan,2008). The ADDIE model is a student-centered holistic instructional design, which is in line with many theories of teaching and learning and achieves good The ADDIE model is a student-centered holistic instructional design that is compatible with many instructional theories and achieves good teaching results(Qi Huixuan,2016).

5. Basic Strategies for Precision Instructional Interventions

The Precision Teaching Model deeply integrates information technology with teaching practices. Initially, it constructs learner profiles to precisely grasp each student's learning situation. Subsequently, based on in-depth analysis of these profiles and precise problem diagnosis, cohort intervention strategies are designed and implemented. These strategies leverage artificial intelligence and technology to meet students' differentiated learning needs, reducing teachers' workload in large-scale classrooms. Precision teaching not only promotes students' overall development but also enhances teaching and learning efficiency and quality, advancing the modernization and innovation of education

5. 1 Cohort Intervention Strategies

5. 1. 1 Intervention strategies for exceptional learners

Based on the diagnostic analysis, outstanding learners tend to focus on results rather than the learning process, have limited self-concept, and lack self-awareness for in-depth and self-directed learning. To address these characteristics, we need to start with goal setting and carefully plan task and activity strategies to enhance commitment to learning behaviors. We also need to develop appropriate resource strategies based on learners' personality styles and cognitive development levels to stimulate their motivation and create an atmosphere of self-directed inquiry and collaborative learning. Ultimately, the learning process and outcomes are evaluated based on "excellence"

5. 1.2 Intervention Strategies for Diligent Learners

Diligent learners tend to exhibit a tendency to emphasize quantity over quality; they focus on the quantity of tasks accomplished during the learning process but may neglect the quality of their learning, and as a result, learning outcomes tend to remain at the level of literacy, and lack in-depth independent thinking, effective teaching and leadership. With this in mind, it is important that we focus on diligent learners and improve the quality of their behaviors.

5. 1.3 Intervention strategies for marginalized learners

As a result of the problem diagnosis, marginal learners were found to suffer from a lack of self-management, as evidenced by a lack of behavioral engagement, frequency of interaction and concentration, as well as a lack of solid mastery of basic knowledge. Given this situation, these students need to rely on teacher support and individual attention. The main thing they need to do is to complete learning tasks, submit assignments and tests in a timely manner in order to gradually improve their learning outcomes.

In this study, learning strategies were developed for different types of learners in terms of five dimensions: objectives, tasks, activities, resources, and evaluation (as shown in Table 2.).

Table 2. *Cohort intervention strategies*

Type of learner	Excellence in Learning	Diligent learner	Marginal learners
Target strategy	Enhance engagement and self-directed learning; develop self-motivation.	Improve behavior quality; increase interactions and task quality.	Boost engagement and competencies through personalized strategies.
Mission strategy	Progressive tasks; Deepen cognitive level.	Moderate difficulty tasks; teacher-monitored practice.	Basic learning activities; personalized instruction.
Campaign Strategy	Independent exploration; challenging tasks; peer feedback.	Collaborative learning; reinforce basic knowledge.	Scaffolded activities; effective questioning; group collaboration.
Resource strategy	Advanced exercises and resources; challenging test questions.	Video materials; practice exercises; supplementary guides.	Microteaching; easier test questions.
Evaluation strategy	Reflective and performance assessments based on "excellence".	Reflective assessment; focus on quality of outcomes.	Reflective assessment; targeted strategy guidance.

6. Conclusions of the study

From the systematic documentation, it is obtained that student data, as a direct source of information for teaching and instruction, is crucial for the implementation of precision learning. Precision teaching interventions based on learner profiles can rely on technology to gain insight into the characteristics and learning needs of each student and implement highly targeted personalized learning strategies. In view of this, this study reviewed relevant literature and combined with fieldwork data to scientifically categorize student profile parameters and propose a learner profile and its construction process. Based on the results of the profiling, an accurate teaching intervention framework was constructed and a multidimensional intervention strategy was formulated, which provides a practical solution for the current implementation of personalized learning.

Acknowledgements

This work was supported by the [Humanities and Social Sciences Youth Foundation of Ministry of Education] under Grant [24YJC880091]; [National Natural Science Foundation of China] under Grant [62467007]. This work was supported by Key Laboratory of Education Digitalization of Gansu Province.

References

Core Literacy Research Group. Core literacy for student development in China[J]. Chinese Journal of Education,2016(10):1-3

Office of Educational Technology, United States Department of Education. Artificial Intelligence and the Future of Teaching and Learning [EB/OL]. (2023-05-10) . http: www.e .gov Technology in ex.html//d//d

Cooper A. The Inmates Are Running the Asylum: Why High Tech Products Drive Us Crazy and How to Restore the Sanity[M]. 2004: 288.

Binder C. Precision Teaching: Measuring and attaining exemplary academic achievement. Youth Policy Journal, 10 (7), 12-15[EB/OL]. (1988)

Liu Caiyan. Theoretical research on instructional design of online courses [D]. Inner Mongolia Normal University,2008.

Qi Huixuan. On the Implications of ADDIE Model for Flipped Classroom Teaching Design[J]. China Adult Education,2016,(17):107-10

Generative AI pedagogical agent as learners' inquiry assistant: Designing and implementing Situational Teaching

Lin LI[a], Yan LIU[a*]
[a]*School of Education, Shanghai Jiao Tong University, China*
*flyliuyan0707@sjtu.edu.cn

Abstract: The rapid adoption of GenAI is deeply reshaping education, enriching the educational scenarios and providing the new approach for innovative teaching designing and personalized learning. However, in traditional science education, the learning scenarios are disconnected from the real inquiry situation, which makes it difficult to enhance students' interest and scientific inquiry skills. This study designed the GenAI pedagogical agent and constructed a GenAI-empowered situational teaching model to improve high school students' scientific inquiry skills. The study was conducted under two separate experiments through a quasi-experimental approach. Results revealed that the model was effective in promoting students' scientific inquiry skills and showed significant progress in all steps of scientific inquiry. In addition, the results showed that the teaching model can significantly stimulate students' interest in learning. Surprisingly, it was found that GenAI may produce content illusion and students' over-reliance on GenAI. We discussed the implications of the findings and recommended it to educators seeking to prepare students for the challenges Higher-order Thinking Skills.

Keywords: GenAI, scientific inquiry skills, pedagogical agent, situational teaching

1. Introduction

Scientific literacy drives societal progress and human advancement. The PISA 2025 framework prioritizes students' abilities to identify research questions, design investigations, and critically analyze data – aligning with global science education standards (White, 2023; Pumfrey, 1991; NRC, 1996; Milford et al., 2010). While students should explore real-world phenomena like scientists, abstract textbook content often diminishes engagement and understanding (Herrington & Oliver, 2000). To address this, situational teaching helps students grasp concepts through meaningful scenarios (Papalazarou et al., 2024). Generative AI (GenAI) now offers new tools: its interactive, adaptive nature enables personalized learning environments, real-time support, and tailored guidance to deepen scientific inquiry (Casheeka et al., 2024). However, existing research focuses primarily on GenAI's basic functions, lacking classroom-ready models or empirical studies on using GenAI-enhanced contextual teaching to develop these crucial skills. This study therefore aims to build and test a GenAI-powered contextual teaching model specifically for cultivating scientific inquiry skills among Chinese high school students.

2. Model Construction and Activity Design

The GenAI-empowered situational teaching model proposed in this study for the cultivation of high school students' scientific inquiry skills is guided by the theory of situational cognition and learning, and builds an authentic and immersive scientific inquiry context. Based on this, the study designs a Generative AI pedagogical agent that can support this situational teaching and learning, aiming to enhance the students' immersive experience and interest in learning in the process of scientific inquiry, and to improve their Scientific Inquiry Skills.

2.1 Theoretical Foundation

2.1.1 Flow theory

Flow theory posits that when learners become fully immersed in a situation, they become highly focused and enter an immersive state (Csikszentmihalyi & Graef, 1975). Based on this, the theory emphasizes that teachers should establish learning conditions conducive to student immersion during instruction. Therefore, when designing the instructional model, this study fully considered students' psychological needs, thus addressing the fundamental question of why learning contexts should be created in teaching.

2.1.2 Situational cognition and learning theory

Situational cognition and learning theory emphasizes that learners should construct knowledge through authentic activities within specific contexts, rather than through passive learning via abstract symbol systems (Wilson & Myers, 2000). Based on this principle, when designing the instructional model, this study carefully considers how to integrate situation creation with teaching content, thus addressing the essential question of what types of situations should be established in teaching.

2.1.3 A constructivist-based theory of technology use

Jonassen's constructivist framework proposes that information technology supports learning through five key functions: multimodal resource libraries, information tools, contextual support tools, reflective thinking scaffolds, and social mediation (Jonassen, 2003). This study fully integrates the characteristics of GenAI technology to design a situation-based teaching model for developing scientific inquiry skills, thereby addressing the critical question of how to effectively create teaching situation.

2.2 Model construction: GenAI-empowered situational teaching model for the cultivation of scientific inquiry skills（GSM for SIC）

GenAI-empowered situational teaching does not simply aim at knowledge transfer, but more importantly guides students to be able to independently explore and dis-cover problems in real-life situations, independently investigate problems using ideas and methods of scientific inquiry, and transfer and apply what they have learned from their investigations. Therefore, this study designed a GenAI-empowered situational teaching pointing to the development of scientific inquiry skills, as shown in Figure. 1. By summarizing the ideas of situational cognition and learning theory, the model proposes that situation creation should follow the principles of authenticity, interactivity, generativity, Constructiveness and Integrity. Building on Jonassen's five principles and technological framework, this study establishes GenAI integration guidelines to create authentic, active, collaborative (human-AI), constructive, and intentional learning environments. We define roles, functions, and dialogue features for GenAI learning companions to support scientific inquiry: Teachers first create authentic situations with investigable problems, where GenAI engages students to identify scientific questions. During inquiry, students—with teacher scaffolding—hypothesize, experiment, and analyze data while GenAI provides multimodal resources, personalized tutoring, and Socratic dialogue to foster knowledge construction. Finally, teachers consolidate learning and promote transfer, while GenAI evaluates the inquiry process to drive reflection. Throughout, students remain active agents constructing knowledge through scaffolded exploration to develop scientific inquiry skills.

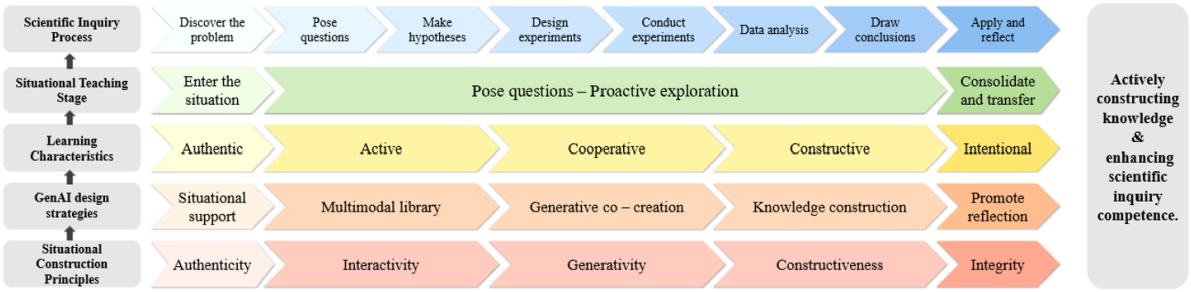


Figure 1. GenAI-empowered situational teaching model for the cultivation of scientific inquiry skills (GSM for SIC)

2.3 Activity Design: Medical Innovation in Epidemic Prevention and Control

To validate the effectiveness of the aforementioned situated instructional model, the study designed a 6-class-hour interdisciplinary scientific inquiry activity centered on the authentic, life-relevant theme of "Medical Innovations in Epidemic Prevention and Control."

3. Teaching Practices and Outcomes

3.1 Participants

The experimental subjects of this study were two classes of 74 students, 29 male and 45 female, in the first grade of a middle school in S city, China. The students had some prior but not in-depth knowledge about immunization-related issues and had been exposed to simple scientific inquiry experiments in their previous studies. Through random assignment, 37 students were assigned to the experimental group, i.e., they were taught in GSM for SIC, which points to the development of scientific inquiry competencies. 37 students were assigned to the control group, i.e., taught in the traditional teaching mode.

3.2 Research tools

3.2.1 Questionnaire

In this study, the “Scientific Inquiry Skills Questionnaire” was administered to students in the experimental group and the control group before and after the experiment. The questionnaire was adapted from Liu's (2018) “Scientific Inquiry Skills Questionnaire” and referred to the description of scientific inquiry skills in the Chinese curriculum standards (MOE, 2020) and the requirements of the PISA 2025 Scientific Literacy Framework for students' scientific inquiry skills (White, 2023). The questionnaire was divided into a total of seven dimensions of posing questions, making hypotheses, designing experiments, conducting experiments, analyzing data, drawing conclusions, and expressing and communicating with a total of 21 questions on a five-point Likert scale (1 = Strongly Disagree; 5 = Strongly Agree) with reverse questions. The scale had good reliability (Cronbach 's α of 0.95) in this study.

In this study, a learning interest questionnaire on the topic was administered to the students in the experimental group before and after the experiment. This questionnaire was adapted from PISA 2015 (OECD, 2017) regarding the description of interest in science learning based on the content of the lectures. The questionnaire consisted of 4 questions on a five-point Likert scale (1=strongly disagree; 5=strongly agree) with reverse questions. The scale had good reliability in this study (Cronbach 's α of 0.84).

3.2.2 GenAI pedagogical agent design

In this study, based on General Language Model 4, we designed and developed an intelligent learning companion, and set up roles and personalized prompts for this intelligent learning companion. Among them, the personalized prompts are developed from four aspects: roles, goals, thinking paths and personalized characteristics. In the thinking path section, add the paragraph “Use questions to guide students to deconstruct and deeply analyze the core question rather than give direct answers”, in the thinking path section of the personalized prompts, emphasize “point out the next step in the thinking path in the form of a question”, and in the personalization section, emphasize “heuristic guidance”. In the personalized prompts' thinking path section, it emphasizes “pointing out the next thinking path in the form of a question”, and in the personalized features section, it emphasizes “heuristic guidance” and promotes students' active thinking through heuristic dialogue.

3.3 Data analysis

This study used SPSS 26 software to statistically analyze the obtained data. After passing the normal distribution test, independent samples t-test was utilized to compare the differences between the pre-test scores and post-test scores of scientific inquiry skills of the experimental group and the control group. The paired-samples t-test was used to compare the differences between the pre-test and post-test scores of scientific inquiry skills and learning interest of the experimental group. The above analyses were synthesized to verify the effectiveness of the GSM for SIC.

3.4 Findings

3.4.1 Differences in scientific inquiry skills between experimental and control groups

In this study, independent samples t-tests were conducted on the pre-test scores and post-test scores of the subject students respectively. The results of the test showed that in the pre-test, there was no significant difference (t=1.857, p=0.067) between the scores of the experimental group (M=3.61, SD=0.49) and the control group's scientific inquiry skills (M=3.36, SD=0.68). In the posttest, the experimental group's scientific inquiry skills (M=4.19, SD=0.59) was significantly higher (t=2.305, p=0.012) than the control group's (M=3.89, SD=0.54) scores. It indicates that GSM for SIC can effectively enhance students' scientific inquiry skills. (see Figure 2)

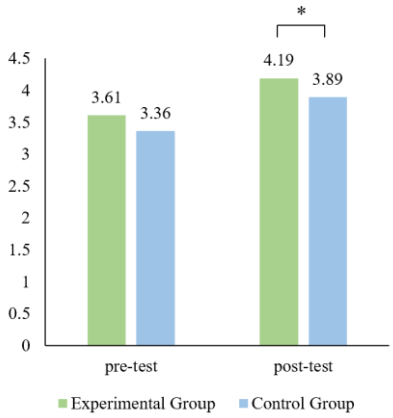


Figure 2. Differences in scientific inquiry skills between the experimental and control groups in the pre-test and post-test respectively

3.4.2 Improvement of scientific inquiry skills in the experimental group

In this study, a paired-samples t-test was conducted to examine the pre and post-test levels of scientific inquiry skills of students in the experimental group in each dimension. The results of the test show (Figure 3) that the posttest scores of the students' scientific inquiry skills in

each dimension are higher than the pretest scores, indicating that GSM for SIC can effectively enhance the students' skills in all processes of scientific inquiry.

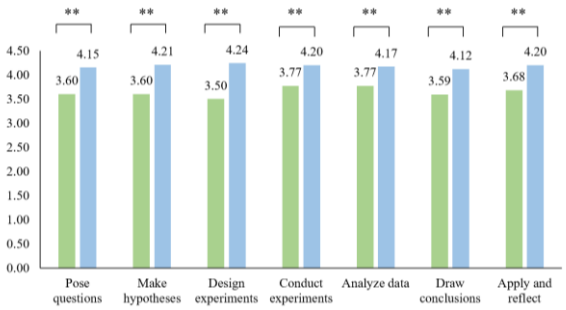


Figure 3. Differences between pretest and posttest of students' scientific inquiry ability in each dimension in the experimental group

3.4.3 Improvement of learning interest in the experimental group

In this study, a paired sample t-test was conducted on the pre-test and post-test scores of students' learning interest in the experimental group. The test results showed (Figure 4) that students' learning interest was significantly higher after the experiment (M=4.27, SD=0.67) than before the experiment (M=3.77, SD=0.74), indicating that GSM for SIC can significantly enhance students' learning interest.

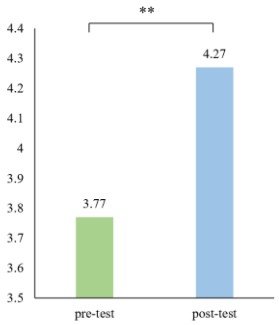


Figure 4. Differences between pre-test and post-test of students' learning interest in the experimental group

4. Conclusion and Prospect

This study makes full use of the features of GenAI, integrates it with teaching in depth, and designs and develops a GenAI pedagogical agent. Based on this, a GenAI-empowered situational teaching model for the cultivation of scientific inquiry skills was constructed, and its effectiveness was simply verified. The results show that the model can significantly improve students' scientific inquiry skills and stimulate students' learning interest.

In order to further realize the potential of GenAI in pedagogy, this study will explore the following aspects in depth:

First, in this study, GenAI pedagogical agent is integrated into the teaching de-sign as a carrier of situation, which can further stimulate students' interest in learning and keep them immersed in the inquiry situation through role-playing and other means. In this process, students maintain relative trust in GenAI. However, due to the existence of a “black box” in the generated content, AI hallucinations may occur from time to time. AI hallucinations generally manifest in two ways. First, providing incorrect answers or fabricating facts. To address this, the study implemented safeguards through knowledge base configuration. Second, failing to

understand student queries and giving irrelevant responses. This may increase cognitive load, causing students to disengage from immersive learning contexts and disrupting inquiry processes. Mitigating this requires further training of the GenAI-based learning companion. Subsequently, this research will explore methods to minimize the second type of hallucination, refine the design process, and provide frontline teachers with reliable, effective strategies for integrating GenAI into situated teaching.

Second, this study verified the impact of GSM for SIC on students' scientific inquiry skills and learning interest. Although certain restrictions were imposed on the output content of GenAI in this study so that it could not tell the answers directly to students. However, in the course of the study, we still find that a few students are able to make GenAI output the answers directly through the subtle design of cue words. The convenience of GenAI is prone to induce learners to fall into the trap of cognitive outsourcing, which makes students rely too much on GenAI and leads to the weakening of cognitive level and higher-order thinking. Therefore, when integrating GenAI into instructional design, its impact on students' higher-order thinking needs to be further considered.

Acknowledgements

This research is supported by the 2024 Young Scientists Fund of the National Natural Science Foundation of China " Multidimensional Dynamic Assessment of Complex Task-Driven Scientific Inquiry Ability and Individualized Intervention Research" (Grant No. 62407028)

References

Casheekar, A., Lahiri, A., Rath, K., Prabhakar, K. S., & Srinivasan, K. (2024). A contemporary review on chatbots, AI-powered virtual conversational agents, ChatGPT: Applications, open challenges and future research directions. *Computer Science Review, 52*, 100632.

Csikszentmihalyi, M., & Graef, R. (1975). Socialization into sleep: Exploratory findings. *Merrill-Palmer Quarterly of Behavior and Development, 21*(1), 3-18.

Herrington, J., & Oliver, R. (2000). An instructional design framework for authentic learning environments. *Educational technology research and development, 48*(3), 23-48.

Liu Dongfang. (2018). An Investigation of Middle School Students' Performance and Current Status of Scientific Inquiry Ability. *Journal of Northeast Normal University (Philosophy and Social Sciences Edition)*, (05), 178-184. doi:10.16164/j.cnki.22-1062/c.2018.05.028.

Milford, T. M., Jagger, S., Yore, L. D., & Anderson, J. O. (2010). National influences on science education reform in Canada. *Canadian Journal of Science, Mathematics and Technology Education, 10*, 370-381.

Ministry of Education. (2020). *Senior High School Biology Curriculum Standards (2017 Edition, 2020 Revision)*. People's Education Press.

National Research Council. (1996). *National Science Education Standards*. Washington, DC: National Academy Press.

OECD.(2017).*PISA 2015 Assessment and Analytical Framework*.OECD Publishing.https://doi.org/10.1787/9789264281820-en

Papalazarou, N., Lefkos, I., & Fachantidis, N. (2024). The effect of physical and virtual inquiry-based experiments on students' attitudes and learning. *Journal of Science Education and Technology, 33*(3), 349-364.

Pumfrey, S. (1991). History of science in the National Science Curriculum: a critical review of resources and their aims. *The British Journal for the History of Science, 24*(1), 61-78.

Jonassen, D. H. (2003). Learning to solve problems with technology: A constructivist perspective. *Merrill Prentice Hall.*

White, P. J., Ardoin, N. M., Eames, C., & Monroe, M. C. (2023). Agency in the Anthropocene: Supporting document to the PISA 2025 Science Framework. *OECD Education Working Papers,* (297), 0_1-43.

Wilson, B. G., & Myers, K. M. (2000). Situated cognition in theoretical and practical context. *Theoretical foundations of learning environments*, 57-88.

Ambiguity Resolution with Human Feedback for Code Writing Tasks

Aditey NANDAN[a] & Viraj KUMAR[b*]
[A]*Mathematics and Computing, Indian Institute of Science, India*
[b]*Kotak-IISc AI-ML Centre, Indian Institute of Science, India*
*viraj@iisc.ac.in

Abstract: Specifications for code writing tasks are usually expressed in natural language and may be ambiguous. Programmers must therefore develop the ability to recognize ambiguities in task specifications and resolve them by asking clarifying questions. We present and evaluate a prototype system, based on a novel technique (ARHF: Ambiguity Resolution with Human Feedback), that (1) suggests specific inputs on which a given task specification may be ambiguous, (2) seeks limited human feedback about the code's desired behavior on those inputs, and (3) uses this feedback to generate code that resolves these ambiguities. We evaluate the efficacy of our prototype, and we discuss the implications of such assistive systems on Computer Science education.

Keywords: Computer Science education, Ambiguous Specifications, Code LLMs.

1. Introduction

Real-world specifications for code writing tasks are often expressed in natural language, and such specifications can be ambiguous (Jinwala & Shah, 2015). A failure to resolve such ambiguities early in the software development lifecycle can lead to costly errors (Fernández et al., 2017). Therefore, existing computing curricula such as CS2023 (ACM/IEEE-CS/AAAI Joint Task Force, 2023) emphasize the importance of requirements elicitation (pg. 242) and regard the ability of students to assess the "lack of ambiguity" in technical documents as a core learning outcome (pg. 280). Since students have easy access to powerful Generative AI tools, and since these tools are increasingly being used in professional software development contexts, Becker et al. (2023) have highlighted an urgent need to review educational practices in computing. In this spirit, we ask the following question: Can assistive systems help programmers identify and resolve ambiguities in task specifications? If so, we are interested in understanding the impact of such systems on computing education. In this paper, we limit our attention to code writing in the context of introductory programming courses (CS1).

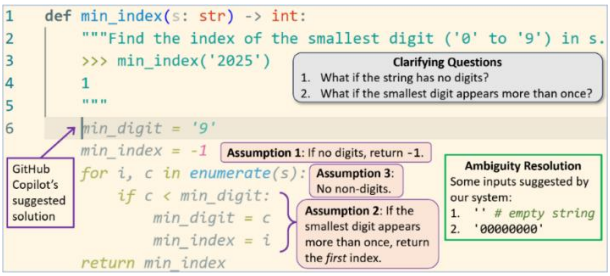


Figure 1. A solution with assumptions generated by GitHub Copilot (Line 6 onwards) for an ambiguous code-writing task (problem P1 in this study, Lines 2 to 5). A good programmer would ask clarifying questions instead of making assumptions to resolve ambiguities.

Figure 1 shows a typical CS1 code-writing task with a specification that is ambiguous because it is incomplete: the purpose statement (or docstring) on Line 2 and the example (or doctest) on Lines 3 and 4 fail to specify the code's expected behavior on certain inputs. As Schneider (1978) notes, such ambiguous specifications can be interpreted in multiple

"reasonable" ways that are functionally inequivalent. Before writing any code, a good programmer should ask clarifying questions to distinguish between these inequivalent interpretations. In contrast, modern code-generation tools such as GitHub Copilot often write code by making assumptions. For instance, the code in Figure 1 makes two seemingly "reasonable" assumptions: the function should return `-1` if the input string `s` has no digits (Assumption 1), and if the smallest digit occurs multiple times in `s`, it should return the index of the *first* occurrence (Assumption 2). However, we think it is "unreasonable" for the code to assume that all letters in `s` are digits (Assumption 3). We explore whether assistive tools can help programmers recognize that a given task supports multiple reasonable interpretations, and can help distinguish between these interpretations by suggesting clarifying questions. If so, we hope to identify skills that students will need to develop if such tools become widely available. We propose a technique, Ambiguity Resolution with Human Feedback (ARHF), whose novelty stems from the insight that code generation models can be prompted to (1) use an ambiguous task specification to generate *specific* inputs on which the code's desired behavior *may* be ambiguous, and (2) use human feedback that clarifies the code's desired behavior on these specific inputs to produce code that handles *general* inputs correctly. We have developed a prototype system that uses ARHF, and we have evaluated it on six CS1-level code-writing tasks in Python. Our research questions are:

- **RQ1**: How effectively can our ARHF system generate inputs that distinguish between all reasonable interpretations of an ambiguous task specification?
- **RQ2**: How effectively can our ARHF system use human feedback to generate accurate code given an ambiguous task specification?

2. Related Work

As powerful Generative AI tools are being integrated into professional workflows, computing education researchers are rethinking pedagogy, assessments, and learning outcomes (Prather et al., 2023, Raman & Kumar, 2022). For instance, code-writing tasks are typically simple and well-specified, and large language models trained to generate code (henceforth: Code LLMs) can solve such tasks reliably (Prather et al., 2023). Thus, recent research has explored Probeable Problems (Pawagi & Kumar, 2024; Denny et al., 2025): code-writing tasks where the specification can be interpreted in multiple ways because it lacks certain details. To solve such problems, programmers must issue 'probes' to clarify the target interpretation before writing code. In contrast, our approach has numerous similarities to interactive test-driven code generation (Lahiri et al., 2022). Although the task specification *S* may be ambiguous (i.e., it may be consistent with multiple "reasonable" interpretations), both approaches assume that the programmer has a target implementation *X* in mind. Both use a Code LLM to generate a candidate implementation *C* from specification *S* and a set of test inputs *T* that may reveal differences between multiple interpretations of *S*. Both execute implementation *C* on each input in *T*, and both rely on feedback from the programmer to determine whether *C* matches the target implementation *X*. The key difference is the nature of the human feedback: whereas Lahiri et al. (2022) seek "lightweight" Boolean feedback (i.e., whether or not implementation *C* has the desired behavior on each input in *T*), we expect programmers to accurately specify the behavior of *X* whenever this differs from the behavior of *C* on an input in *T*. Hebbar et al. (2025) call this skill reactive task comprehension.

3. Methods

3.1 ARHF System Architecture

Our ARHF system is implemented as a web based, locally run application. The frontend is built using HTML, CSS and JS, while the backend is implemented using Python (Flask). The input to ARHF is a (potentially ambiguous) specification *S*. For this study, we assume that *S*

is a Python function *signature* (the function name as well as the name and type-hint for each argument), a *docstring* explaining the function's purpose, and at least one *doctest* (an example showing the desired output as per the target interpretation on some test input). ARHF has 3 components: Initial Code Generation (which generates a candidate implementation *C* from *S*), Test Input Generation (which generates test inputs *T* from *S* and *C*), and Code Correction (which attempts to fix implementation *C* using human feedback on *C*'s output for each input in *T*). All components use Code LLMs. Our prototype system uses a common Code LLM: Qwen2.5-Coder-32B-Instruct (Hui et al., 2024) with temperature = 0.7, top_p = 0.8, and top_k = 20, via the Hugging Face Inference API (https://huggingface.co/Qwen/Qwen2.5-32B-Instruct/blob/main/generation_config.json). The detailed prompts for each component are at: https://github.com/Nandan-Aditey/ARHF-Resources/tree/main/ARHF-Prompts

In addition to using a Code LLM, the Test Input Generation component generates test inputs targeting edge cases using CrossHair (Schanely, 2017). Each input in this combined set *T* is then executed on the implementation *C*, and our system displays each input-output pair as a doctest. As an example, Figure 2 shows two such doctests for two interpretations of Problem P1 (`min_index`). The implementation *C* is equivalent to interpretation I_1: the function returns `-1` when string `s` contains no digit. If I_1 is the desired interpretation, the programmer can "Accept" both doctests (as shown in Figure 2 left). If there are no failing doctests, our system proposes *C* as the desired implementation. On the other hand, suppose the desired interpretation is I_2: the function returns `len(s)` when string `s` contains no digit. In this case, the programmer can "Reject" the doctests and specify the correct outputs: `0` for the empty input string, and `5` for the string `'abcde'` (Figure 2, right). Given these failing doctests and implementation *C*, the Code Correction component must inductively deduce that the correct return expression when string `s` has no digits is `len(s)`. If the revised implementation again fails some doctests, it is possible to rerun the Code Correction component. However, we do not perform reruns in this study. Instead, our system presents the "best" of the two generated implementations (the function that passes the maximum number of doctests, and the most recently generated function in case of a tie).



Figure 2. ARHF's interface to obtain human feedback. The two doctests differentiate between interpretations I_1 and I_2 for Problem P1 (`min_index`).

3.2 Quantitative Analysis

We evaluate our system using two types of problems. First, we created three problems with multiple interpretations (P1, P2, P3 in Table 1). Treating each of these interpretations as the target, we attempt to use ARHF to generate the desired implementation (25 times for each interpretation). In addition, we selected three Probeable Problems proposed by Denny et al. (2025), each with 16 "reasonable" interpretations (P7, P8, P9 in Table 1). We lightly modified problem P8 to return a value rather than to print an answer. For these problems, we fixed the target interpretation as the one chosen by Denny et al. (2025), and we attempted to use ARHF to generate the corresponding implementation (100 times each).

Table 1. *The six problems in this study, each with multiple "reasonable" interpretations. The implementations corresponding to each of these interpretations are available at:* https://github.com/Nandan-Aditey/ARHF-Resources/tree/main/ARHF-Problems

Problem	Docstring	Initial Doctest	Interp.
P1	Find the index of the smallest digit ('0' to '9') in *s*.	>>> min_index('2025') 1	4
P2	Find the minimum frequency. Return None if *data* is empty.	>>> min_freq([1, 2, 1]) 2	3
P3	Count the number of digits in *n*.	>>> num_digits(123) 3	2
P7	Count the number of integers in data between a and b.	>>> count_between([1, 2, 3], 0, 5) 3	16
P8	Search data for the smallest even value.	>>> smallest_even([50, 25, 2, 30, 45]) [2]	16
P9	Find the first vowel in s.	>>> first_vowel('apple') 'a'	16

For each problem and each target interpretation *X*, we evaluate the set of test inputs *T* generated by the Test Input Generation component of our system. Specifically, for each non-target interpretation *Y*, we check whether there is at least one input in *T* on which *Y* and *X* disagree. If so, we say that the set of inputs *T* distinguishes interpretation *X* from *Y*. To answer RQ1, we define Input Ambiguity Resolution (IAR) as the fraction of non-target interpretations that are distinguished by the inputs in *T*. Further, we check whether the implementation *C* presented by ARHF is functionally equivalent to *X*. We check for equivalence by manual inspection, and we use the Pass@1 metric (Chen et al., 2021) to estimate the probability that *C* is functionally equivalent to *X*. Finally, we define Code Ambiguity Resolution (CAR) as the fraction of non-target interpretations that are functionally inequivalent to *C*. Since equivalence with *X* implies inequivalence with every non-target interpretation, the Pass@1 metric is always upper bounded by CAR. We use both metrics to answer RQ2.

4. Results

The following table summarizes the performance of ARHF on the 6 problems in this study.

Table 2. *IAR, CAR, and Pass@1 of our ARHF system for the six problems in this study*

Prob.	IAR	CAR	Pass@1	Remarks
P1	(3×25)/(3×25) = 1.0	(4×75)/(4×3×25) = 1.0	(4×25)/(4×25) = 1.0	Each interp. 25 times
P2	(2×25)/(2×25) = 1.0	(50+48+47)/(3×2×25) = 0.97	(25+23+22)/(3×25) = 0.93	
P3	(1×25)/(1×25) = 1.0	(25+0)/(2×1×25) = 0.5	(25+0)/(2×25) = 0.5	
P7	1495/(15×100) = 0.99	1489/(15×100) = 0.99	95/100 = 0.95	Single interp. 100 times
P8	1473/(15×100) = 0.98	1323/(15x100) = 0.88	61/100 = 0.61	
P9	1412/(15×100) = 0.94	1370/(15x100) = 0.91	0/100 = 0.0	

4.1 RQ1: How effectively do ARHF-generated test inputs resolve ambiguities?

To answer RQ1, we evaluate the IAR of our system for the six problems in this study. Our results (Table 2) demonstrate that ARHF achieves the highest possible IAR (1.0) for problems

with only a few reasonable interpretations (P1, P2, and P3) and it maintains a high IAR (at least 0.94) even when problems admit greater ambiguity (P7, P8, and P9). Our system's performance is poorest on problem P9. To understand why, we analyze the test inputs and code generated by ARHF in relation to the 16 "reasonable" interpretations for each problem. When the target interpretation is I_7, ARHF correctly distinguishes the non-target interpretation I_8 from I_7 in only 12% of invocations. Interpretation I_7 is: "first vowel in vowel order, ignoring differences in upper-case and lower-case vowels". The test inputs generated by ARHF are often unable to distinguish this from interpretation I_8: "first vowel in index order, ignoring differences in upper-case and lower-case vowels". Since the task specification lacks so many crucial details, it is perhaps unsurprising that ARHF cannot generate appropriate test inputs. Nevertheless, the high IAR score indicates that the inputs generated can help eliminate many reasonable interpretations from further consideration. Thus, ARHF-generated test inputs appear to be reasonably effective at resolving ambiguities.

4.2 RQ2: How effectively does ARHF resolve ambiguities in the generated code?

To answer RQ2, we evaluate both the Pass@1 and CAR rates of ARHF for each of the six ambiguous code-writing problems. ARHF generally achieves high Pass@1 and CAR rates, except for problems P3 and P9.

For P3, the Pass@1 rate is 0 when the target interpretation is I_2: "count the unique number of digits". Recall that our system achieves an IAR of 1.0 for P3 regardless of which interpretation was chosen as the target (Table 2). Thus, when the target implementation is I_2, the doctests created from human feedback on test inputs generated by our system are inconsistent with interpretation I_1. Nevertheless, the Code LLM in our system generates an implementation corresponding to interpretation I_1 of P3 on all 25 invocations. This suggests an inability of current Code LLMs to generate code that strictly adheres to provided doctests. This limitation of Code LLMs has been noted previously by Heo et al. (2024).

The particularly poor performance on problem P9 is unsurprising since the test inputs are insufficiently rich to compensate for the many missing details in the task specification (see Section 4.1). For P9, 95% of ARHF invocations reveal an implementation corresponding to interpretation I_8 instead of the target interpretation I_7. While the Pass@1 rate is zero, the high CAR rate (0.91) shows that ARHF successfully rejected most of the unwanted interpretations. While this is promising, we also note that when the implementation presented by ARHF is *inequivalent* to the target implementation, it corresponds to an interpretation that is slightly different from the one desired. In other words, when the implementation generated by ARHF is incorrect, it is likely to be subtly buggy!

5. Limitations, Future Work, and Conclusion

We note two key limitations of this study and discuss how they can be addressed in future work. First, our study is limited to just six code-writing tasks (specifically, writing a single function), in a single programming language (Python). Future work can explore a richer variety of code-writing tasks. Second, although our results demonstrate that ARHF can be effective at resolving ambiguities, our system requires the programmer to accurately specify the desired outputs on inputs suggested by our system. It will therefore be useful to evaluate whether students, especially novice programmers, can make effective use of a system such as ours. We are presently evaluating ARHF in a CS1 course setting. We reiterate that our system is a prototype that can be improved in several ways, such as by rerunning the Code Generation module when doctests fail. It is also possible that breakthroughs in Code LLMs will improve their ability to adhere to specified doctests. Thus, the results reported in this paper should be viewed as a conservative lower bound on the performance of such systems.

In conclusion, this paper has presented a prototype system that uses off-the-shelf components to effectively resolve ambiguities in code-writing task specifications using a novel

strategy: Ambiguity Resolution with Human Feedback (ARHF). We have shown that although ARHF is imperfect when key details are missing in the specification, it is nevertheless helpful in rejecting unwanted interpretations of that specification. Since such systems can reduce the risk of costly errors in software development (Fernández et al., 2017), we anticipate that systems more capable than our prototype will become readily available. For educators, such systems could be helpful in two ways. First, code-writing tasks in summative assessments often need to be clear, and systems such as ours could help detect potentially ambiguous wording. Second, learning outcomes and assessments associated with identifying ambiguities (Pawagi and Kumar, 2024; Denny et al., 2025) may require rethinking if students have access to powerful assistive ARHF-like systems.

Acknowledgements

The authors respectively acknowledge support from the Intel AI PC Experiment Development Centre (FA/OTHR-24-0031) and the Kotak-IISc AI-ML Centre towards this research.

References

ACM/IEEE-CS/AAAI Joint Task Force. (2023). Computer Science Curricula 2023 (CS2023): The Final Report. ACM Press, IEEE Computer Society Press and AAAI Press.

Becker, B. A., Denny, P., Finnie-Ansley, J., Luxton-Reilly, A., Prather, J., & Santos, E. A. (2023, March). Programming is hard-or at least it used to be: Educational opportunities and challenges of ai code generation. In *Proc. of the 54th ACM Technical Symposium on CS Education V. 1* (pp. 500-506).

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., ... & Zaremba, W. (2021). Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374.

Denny, P., Kumar, V., MacNeil, S., Prather, J., & Leinonen, J. (2025). Probing the Unknown: Exploring Student Interactions with Probeable Problems at Scale in Introductory Programming. In *Proceedings of the 30th ACM Conference on Innovation and Technology in Computer Science Education* (pp. 618-624). DOI: 10.1145/3724363.3729093.

Fernández, D. M., Wagner, S., Kalinowski, M., Felderer, M., Mafra, P., Vetrò, A., ... & Wieringa, R. (2017). Naming the pain in requirements engineering: Contemporary problems, causes, and effects in practice. *Empirical software engineering*, 22, 2298-2338.

Hebbar, S. V., Harini, S., & Kumar, V. (2025). Refuting LLM-generated Code with Reactive Task Comprehension. In *Proceedings of the 30th ACM Conference on Innovation and Technology in Computer Science Education* (pp. 30-36). DOI: 10.1145/3724363.3729100

Heo, J., Heinze-Deml, C., Elachqar, O., Chan, K. H. R., Ren, S., Nallasamy, U., Miller, A., & Narain, J. (2025). Do LLMs "know" internally when they follow instructions? In Proceedings of the International Conference on Learning Representations (ICLR 2025). *arXiv:2410.14516* [cs.AI].

Hui, B., Yang, J., Cui, Z., Yang, J., Liu, D., Zhang, L., ... & Lin, J. (2024). Qwen2. 5-coder technical report. *arXiv:2409.12186*.

Jinwala, D. C., & Shah, U. S. (2015). Resolving Ambiguities in *Natural Language Software Requirements*. ACM SIGSOFT Software Engineering Notes, 40, 1-7.

Lahiri, S. K., Fakhoury, S., Naik, A., Sakkas, G., Chakraborty, S., Musuvathi, M., ... & Gao, J. (2022). Interactive code generation via test-driven user-intent formalization. *arXiv:2208.05950*.

Pawagi, M., & Kumar, V. (2024). Probeable Problems for Beginner-level Programming-with-AI Contests. In *Proceedings of the 2024 ACM Conference on International Computing Education Research-Volume 1* (pp. 166-176).

Prather, J., Denny, P., Leinonen, J., Becker, B. A., Albluwi, I., Craig, M., ... & Savelka, J. (2023). The robots are here: Navigating the Generative AI Revolution in Computing Education. In *Proceedings of the 2023 Working Group reports on Innovation and Technology in CS Education* (pp. 108-159).

Raman, A., & Kumar, V. (2022, November). Programming pedagogy and assessment in the era of AI/ML: A position paper. In *Proceedings of the 15th ACM India Compute Conference* (pp. 29-34).

Schanely, P. (2017). CrossHair - An analysis tool for Python that blurs the line between testing and type systems. https://github.com/pschanely/CrossHair.

Schneider, G. M. (1978, February). The introductory programming course in computer science: ten principles. In *Papers of the SIGCSE/CSA technical symposium on CS education* (pp. 107-114).

Write or not to Write. Attitudes of Students and Academics towards Written Dissertations in the Light of Artificial Intelligence Development

Ewa DUDA [a*], Ewelina MŁYNARCZYK-KARABIN [b] & Julia WIĘCKIEWICZ-MODRZEWSKA [b]
[a]*Institute of Education, Maria Grzegorzewska University, Poland*
[b]*Institute of Special Education, Maria Grzegorzewska University, Poland*
*eduda@aps.edu.pl

Abstract: The written format of bachelor's and master's dissertations in the social sciences has long adhered to established academic conventions, emphasising clarity, methodological transparency, and theoretical coherence. However, the rapid integration of artificial intelligence into academic research and writing is reshaping these traditions, introducing both transformative opportunities and unprecedented challenges. In this paper, we explore the academic community's attitudes towards the traditional dissertation format, addressing a topic that is underrepresented in the literature. Drawing on qualitative research with 332 students and 81 teachers at a university specialising in pedagogy, we examine arguments for preserving or changing the traditional dissertation format in this field, where AI is less popular than in other scientific disciplines. We also present alternatives to the current form of diploma thesis proposed by those involved in the educational process who will be responsible for educating the next generation. Our findings contribute to the broader debate about academic adaptation in the age of artificial intelligence. They highlight the need to equip current and future teachers with technological proficiency and the ethical determination required to navigate complex, human-centred research fields, where the development of artificial intelligence presents growth opportunities and ethical challenges. Given the scarcity of research on the current thesis form in pedagogy, an article addressing this novelty topic is an important voice in the academic discourse.

Keywords: Academic writing, Artificial Intelligence, AI in education, Bachelor's thesis, Ethical use of AI, Higher education, Master's thesis, Students' competences

1. Introduction

In the current era of rapid advancement in the field of artificial intelligence (AI), particularly in the domain of generative tools exemplified by large language models such as ChatGPT and Claude, traditional methods of assessing students' knowledge and competencies have become the subject of significant debate (Dempere et al., 2023; Grassini, 2023). One such method is the written dissertation, which has long been a mandatory requirement for completing higher education at both the undergraduate and master's levels in many countries around the world. Serving as both a demonstration of scholarly rigor and a bridge to professional or doctoral studies, the thesis demands the synthesis of theoretical knowledge and methodological competence, as well as critical analysis, and often, empirical research. It must also conform to established methodological and editorial standards (Bereźnicki, 2017).

An undergraduate or master's thesis in the social sciences plays a key role in academic training by cultivating independent research skills, critical thinking, and scholarly communication. The path to completion is fraught with conceptual, methodological, logistical and psychological obstacles. The consequences are tangible: the emerging problem of plagiarism (Pınar et al., 2023), insufficient academic writing competence, the gaps in mentorship (Ge et al., 2024), delayed graduation, uneven quality of theses and lost

opportunities for meaningful student engagement in their chosen fields. Some countries have decided to change the format of final academic work to offer students a more diverse choice. In Germany and Austria, the classic written thesis model still prevails, although project-based or practice-oriented works are increasingly accepted. In the United Kingdom, some universities offer alternatives such as portfolios, practical projects, and even multimedia presentations (EHEA, 2020).

These differences raise the question of whether, in a world where advanced AI tools, are widely accessible, the traditional written dissertation remains the optimal method for verifying a student's academic competence. This article examines the evolving format of social science dissertations in the age of AI. It explores attitudes towards the traditional dissertation format among students and academics, addressing a topic that is underrepresented in the literature. Drawing on qualitative research, it explores arguments for preserving or changing the traditional dissertation form, and the alternatives proposed by those involved in the educational process at different levels. The article also highlights the differences in access to AI tools that may exacerbate the existing inequalities between students at institutions with different levels of resources.

2. Dissertations in the Light of Artificial Intelligence Development

In recent years, tools based on artificial intelligence have assumed an increasingly significant role across various domains of life, including higher education. Among the most advanced applications of natural language processing (NLP) technologies are large language models (LLMs), which underpin intelligent chatbots capable of generating text, responding to queries, and supporting both teaching and learning processes (Yigci et al.,2024). Within the realm of academic research and thesis writing, LLMs carry substantial implications, as they can be employed at nearly every stage of the research process, from the formulation of research objectives and the development of literature reviews (Schwenke et al., 2023), to the analysis and interpretation of both qualitative and quantitative data (Ishaq et al., 2023).

The results of the research conducted by Digital Care in 2023 indicate that AI technology is popular among students, 68% of respondents declared using chatbots while studying, and every fifth student planned to use AI in writing their diploma thesis (Musiał, 2024). The research results indicate an urgent need to regulate these issues at universities. Students indicated that, in addition to using AI for language translation or creating presentations, they are willing to use chatbots to write term papers (34% of respondents). Additionally, the vast majority of respondents declared that they positively assess the use of artificial intelligence in the education process. This poses a specific challenge for scientists and universities, which is to update the methods of verifying students' knowledge. Many research centres are therefore returning to the formula of oral exams, because at the moment this form of checking students' knowledge seems to be the most objective and excludes the use of AI tools.

On the other hand, some of universities are implementing procedures banning the use of AI. The use of chatbots has been banned by public schools in New York, the University of Edinburgh and York University. To counterbalance this, permission to use chatbots in science was issued by the prestigious British University of Cambridge or Warwick University (Musiał, 2024). Therefore, one of the ideas that is increasingly being raised in the academic community is to change the formula of tests and exams in order to value oral exams and exams conducted in a traditional form, because they exclude any possibility of using AI tools. Researchers increasingly emphasise that AI technology has made traditional exams more important than ever and chatbots have undermined the arguments for greater use of test papers in the final assessment (Musiał, 2024).

3. Method

3.1 Procedure and Participants

At the university we affiliate is enrolled 3832 students and employed 282 academics, including 181 who officially declare pedagogy as their research discipline. Among them, we distributed

a survey questionnaire in March 2025. Respondents were informed that the survey was voluntary and anonymous, and that they could withdraw at any time. The questionnaire was developed and distributed in Polish using Qualtrics software. The question asked to participants about whether to retain the written form of bachelor's and master's theses offered only two possible answers (yes or no), and was intended to encourage respondents to give an unambiguous answer. Depending on the response given, respondents were invited to present arguments for leaving the current form or to suggest an alternative, preferred form.

There were 546 participants, including 445 students and 101 university teachers, who took part in the study. As a first step, we discarded data sets that did not have open-ended questions answered. Finally, we included data collected from 413 participants in the analysis, including 332 students (80.4%) and 81 university teachers (19.6%). The sample included 356 women (96.2%, M_{age} = 25.8, *SD* = 9.6, range 18-73), 48 men (11.6%, M_{age} = 34.7, *SD* = 16.0, range 19-77) and 9 (2.2%, M_{age} = 25.1, *SD* = 10.4, range 18-52) of those who preferred not to provide gender information. The predominance of women is due to the typical student structure in education.

Regarding the field of study, the sample consisted of 143 students of pedagogy with a teaching qualification (43.1% of students), 113 students of pedagogy without a teaching qualification (34.0%), 73 students of psychology (22.0%) and 3 other students (arts education, sociology). Regarding the level of study, there were 89 students (26.8%) of bachelor's level (1st cycle), 225 students (67.8%) of master's level (long cycle) and 18 students (5.4%) of master's level (2nd cycle). Regarding the study mode, there were 224 full-time students (67.5%, studying during week days) and 108 part-time students (32.5%, studying during weekends).

For university teachers, the sample included 54 (66.7%) representing the discipline of pedagogy (29.8% of the targeted group) and 27 participants (33.3%) representing other disciplines (psychology - 6 participants, sociology - 10 participants, philosophy - 3 participants, other social sciences - 8 participants). In terms of academic degrees, the sample consisted of 20 participants with a Master's degree (24.7%), 42 participants with a Doctorate (51.9%), 16 with a Postdoctoral degree (19.8%) and 3 participants with a Professor's degree (3.7%). In terms of professional experience, the study sample included 19.8% of academics with less than 5 years of experience in higher education, 22.2% with 5-10 years of experience, 32.1% with 11-20 years of experience and 25.9% with more than 20 years of experience in HE.

3.2 Data Analysis Strategy

Quantitative data analysis was conducted using IBM SPSS Statistics 29.0.2 software, employing statistics and frequency calculations. For qualitative analysis, the methodology outlined by Strauss and Corbin (1998) was applied, which similarly emphasises frequency-based thematic evaluation. The process began with a comprehensive review of the collected data, after which the lead researcher generated preliminary codes based on recurring themes emerging from participant responses. These initial codes were then collaboratively reviewed by the team to establish consensus on final code assignments. Discrepancies were resolved through iterative discussions until unanimous agreement was achieved. This approach ensured methodological rigor while accommodating the complexity and multiplicity of perspectives inherent in the dataset.

4. Results

The majority of participants believe that the written form of the thesis should be retained, with similar distributions of results in both groups (χ^2 = 0.16, *p* = .900). Among the student group, 222 participants (66.9%) agree with this opinion, while 33.1% believe that the thesis form should be changed. Among teachers, 54 participants (66.7%) would retained current thesis form. In terms of academic degrees, the results were distributed uniformly across the groups, (χ^2 (3) = 1.86, *p* = .601), with all three professors recommending the retention of the written form of the thesis. A parallel distribution of results was observed across the groups based on declared academic disciplines (pedagogy or other social sciences; χ^2 = 0.22, *p* = .638) and professional experience (χ^2 (3) = 0.96, p = .811).

According to the mode of study, the results were equally distributed across the groups (χ^2 = 0.54, p = .461). The proportion of full-time students who believed that the form of the thesis should be retained was 69.9%, while the corresponding figure for part-time students was 64.8%. Among the group of students who already have experience of writing a thesis, who have completed a bachelor's degree, 52.9% of participants believe that the thesis form should be retained, while 47.1% believe that it should be changed, but the difference is not statistically significant (p = .196).

In the next stage of the analysis, the qualitative arguments presented by the participants to justify their positions were analysed. The table 1 shows assigned codes with examples given by students (normal text) and academics (italics). The same thematic codes could be assigned in relation to both groups, but the frequency of occurrence of these codes varies.

Table 1. *The frequency of codes representing arguments for not changing the form of bachelor's and master's theses as perceived by students* N_1=222 *and academics* N_2=54.

Code: Lack of a better alternative	Students (S): 22.6%	*Academics (A): 9.2%*
- Students cannot imagine any other form of thesis work. - There is no other ideal option that would guarantee the independence of the student's work. - The format of the thesis should be retained, but that more emphasis should be placed on talking about the thesis during the defence.		
Code: Current form is a good solution		S: 21.2%; *A: 3.7%*
- This form is transparent. - It has worked well for many years. - It gives students the opportunity to demonstrate their intelligence and creativity, to show their area of interest related to their field of study. - The use of AI should not affect the quality and independence of the thesis. - In addition to writing the thesis, it is necessary to conduct research, and the tasks involved in this process cannot be replaced.		
Code: Verification by an anti-plagiarism system		S: 15.3%; *A: 3.7%*
- Trust in an anti-plagiarism system. - In the creative and generative aspect, an AI is not able to generate anything better than a human		
Code: Tradition		S: 13.1%; *A: 7.5%*
- It has its own atmosphere. - It would be problematic and even unfair to change the form of thesis.		
Code: Confirmation of the work put into thesis' development		S: 9.5%; *A: 11.1%*
- Even it is written with the help of an AI, it still mobilises the student's own work on a particular topic. - V*erification of student's ability to analyse the literature, selection of content, writing skills and, consequently, scientific potential, and, in the context of research methodology, the ability to select a research paradigm for the topic undertaken, which is the essence of the student's own work.*		
Code: Development of important skills		S: 5.9%; *A: 46.2%*
- It is a good way to ensure that people do not completely lose their writing and thinking skills. - Students are already noticing problems with the formulation of more complex sentences or excessive reasoning. - Despite the existence of chatbots, it is very difficult to write a challenging dissertation. *- Whether students use AI or not, they learn to arrange theoretical and methodological content according to the requirements for scientific research, exercise their mind and self-control, and develop editorial competence.* *- The process of preparing a thesis is an important way of presenting one's thinking, planning, as well as understanding the material read, selecting information.* *- With the diligent work of the dissertation supervisor represented by frequent meetings, conversations, in-depth discussions on the knowledge and critical reflection of sources, it is still a very good opportunity to shape the valuable research skills of the diploma student, including analytical and synthetic ones.*		
Code: A form of memorabilia		S: 5.0%; *A: 0%*
- "I would like to write my thesis like my predecessors and hold it in my hands, put it on a shelf in the family home like a trophy"		
Code: Material evidence		S: 5.0%; *A: 7.5%*
- An important document that should be printed on paper for greater security. *- A document that gives a sense of authorship and responsibility for the outcome.*		
Code: Other		S: 2.4%; *A: 11.1%*
- Changing the form of bachelor's and master's theses should be the last point in the revolution of the whole system. - The final form of studies should not be changed when there are no official standardised and regulated rules for the use of AI by students, there are no limits.		

Of the subgroup of participants who identified the need to change the current format of diploma theses, nearly one quarter (24.6%) reported finding it difficult to propose an alternative assessment method. The next thematic clusters concerns alternatives for written format of thesis. Students (22.7%) indicated that an exam or test would be the most appropriate form of assessment. Another students (22.7%) prefer oral presentations, demonstrating their knowledge and ability to defend own arguments. Another suggested form of thesis are projects prepared either individually or collaboratively (12.7%). They provide a multidimensional picture of competencies, including planning, execution and presentation of results. Students (10.9%) preferred also a practical examination involving teaching simulations and questions from an examiner mimicking typical or unforeseen school scenarios, in order to verify pedagogical flexibility. Finally, a small group (6.4%) considered written work to be an acceptable form of assessment, provided that new requirements or conditions were introduced, such as 'written work produced during class under the supervision of a tutor'.

On the other hand, majority of teachers (40.7%) regard examination-based proposals to be the most versatile and authentic form of assessment. They would replace or supplement theses with exams, favouring oral formats and a combination of written and oral verification. University teachers (37.0%) also value projects tailored to a specific discipline and real-world contexts. Teachers emphasise projects that mirror professional and research practices, engage students collaboratively and result in tangible outputs and presentations. Some of teachers (14.7%) have not formulated specific ideas, or they doubt the effectiveness of current processes without new verification tools. A minority (3.8%) believe that the conventional dissertation is still valid, provided safeguards such as plagiarism detection or an oral defence ensure integrity of authorship. One teacher proposes a case-study format to simulate workplace decision-making and assess the application of theory to concrete problems.

5. Discussion and Conclusion

The results of the conducted research indicate that the majority of respondents believe that the thesis should remain in written form. A key argument is the preservation of a long-standing tradition. Furthermore, it is difficult to ensure that another form would guarantee the independence of students' work in the current circumstances. The current thesis preparation process allows students to select their area of interest and present it in a creative way, while also enabling them to practise their writing skills, acquire new knowledge, and develop editorial skills. Teachers also emphasised that, when preparing their theses, students acquire new skills and learn to organise theoretical content and conduct scientific research in accordance with current methodological concepts. Additionally, as mention other studies (Ge et al., 2024) this form of work empowers students and fosters a sense of responsibility for their work.

The traditional method of preparing theses enables teachers to evaluate students' ability to analyse literature and select content, and to assess their academic potential. For students, the traditional thesis-writing process also has a symbolic dimension, marking the culmination of an important stage in their professional development. As in the research by Schwenke et al. (2023), both students and teachers believe that using ChatGPT does not negatively impact the key educational and assessment functions of the traditional thesis-writing process. This tool does not diminish valuable academic elements such as the independent processing of literature, the formulation of arguments, or the conduct of analyses. Instead, artificial intelligence can increase productivity by enabling students to complete the various stages of their thesis more quickly and efficiently.

The master's and bachelor's theses are intended to demonstrate students' ability to conduct and report on research. If a thesis is written with AI support but represents good quality and cites reliable sources, it still demonstrates this ability. The only difference is that it takes advantage of the opportunities offered by modern technological developments. The ability to write is a civilisational achievement that should not be abandon just because modern technologies have been invented. They should be incorporated into in the process of writing papers, set rules for their use, and they should be used to best express students’ thoughts and goals. Our study shows that university teachers still emphasise that preparing a written thesis is an important and formative part of the study process. It is a time for developing a

methodical approach to the subject matter and considering the ethics of action. Therefore, despite concerns about the unauthorised use of AI by students, the ethical context and values inherent in creating one's own work and text should be preserved.

When talking about social science dissertations, AI can primarily be used as a tool to improve the writers' language skills, as many students struggle to write properly. Improving writing style is one of the most challenging aspects of supporting graduate students. Providing an indication of the extent to which AI can be used in a paper, alongside consistent mentorship from the supervisor, should give students the opportunity to learn how to conduct research and write a scientific paper. This is a useful life skill. AI can and should be helpful in thesis writing, too. However, the thesis should clearly indicate the extent to which a chatbot was used. As with anti-plagiarism software, there should be an upper limit for admitting work.

The findings of this study are subject to certain limitations. The research was conducted at a single public pedagogical university in Warsaw, meaning the results are inherently shaped by the institutional context of the participants and may not be generalisable to students in other settings. Replicating the study across pedagogical institutions in different cities or within private universities could yield divergent trends. To strengthen the validity of the conclusions or identify patterns in the use of AI tools in higher education, expanding the research to include students from diverse academic institutions is recommended.

In conclusion, the dissertation in the AI era must remain a rite of passage that cultivates not only technical competence but also the intellectual courage to question, adapt, and innovate. By reimagining, rather than resisting, AI's role, the social sciences can preserve their humanistic ethos while equipping future scholars to tackle complex societal challenges with both technological fluency and ethical resolve. The path forward lies not in dogmatic adherence to tradition but in a dynamic synthesis of innovation and integrity, ensuring that the dissertation remains a testament to the enduring value of critical, compassionate inquiry in an increasingly automated world.

References

Bereźnicki, F. (2017). *Praca licencjacka i magisterska z pedagogiki, psychologii i socjologii. Poradnik dla studentów*. Impuls.

Dempere, J., Modugu, K., Hesham, A., & Ramasamy, L. K. (2023). The impact of ChatGPT on higher education. *Frontiers in Education*, *8*, 1206936. https://doi.org/10.3389/feduc.2023.1206936

EHEA (2020). *National Qualifications Frameworks across EHEA countries*. European Higher Education Area.

Ge, Z. G., Zhao, Y., & Liu, Y. (2024). Exploring quality issues in academic master theses of education majors and corresponding countermeasures. *Journal of Pedagogical Research*, *8*(4), 30-47. https://doi.org/10.33902/JPR.202428171

Grassini, S. (2023). Shaping the Future of Education: Exploring the Potential and Consequences of AI and ChatGPT in Educational Settings. *Education Sciences, 13*(7), 692. https://doi.org/10.3390/educsci13070692

Ishaq, K., Jhatial, S., & Parveen, F. (2023). Enhancing Statistical Understanding: A Brief Discussion on the Role of ChatGPT in Business Management. *Journal for Business Education and Management, 3*(2), 59-76. https://doi.org/10.56596/jbem.v3i2.50

Musiał, A. (2024). Sztuczna inteligencja w służbie nauki. Gdzie są granice wykorzystania AI w pracy naukowca? *Progress. Journal of Young Researchers, 14*, 44-57. https://doi.org/10.26881/prog.2024.14.03

Pınar, Y., Gür, D., Pınar, N. K., Demir, K., Iltar, E. K., Songören, S. A., & Özenici, S. (2023). A comparative study of postgraduate theses in pedagogy and preschool education in Austria and Turkey. *Frontiers in Psychology*, *13*, 1051923. https://doi.org/10.3389/fpsyg.2022.1051923

Schwenke, N., Söbke, H., & Kraft, E. (2023). Potentials and Challenges of Chatbot-Supported Thesis Writing: An Autoethnography. *Trends in Higher Education, 2*(4), 611-635. https://doi.org/10.3390/higheredu2040037

Strauss, A. L., & Corbin, J. M. (1998). Basics of qualitative research : Techniques and procedures for developing grounded theory (2nd ed.). Los Angeles, CA: Sage Publications.

Yigci, D., Eryilmaz, M., Yetisen, A. K., Tasoglu, S., & Ozcan, A. (2025). Large language model-based chatbots in higher education. *Advanced Intelligent Systems*, *7*(3), 2400429. https://doi.org/10.1002/aisy.202400429

Small but Significant: On the Promise of Small Language Models for Accessible AIED

Yumou WEI[a], Paulo CARVALHO[a] & John STAMPER[a]
[a]*Human-Computer Interaction Institute, Carnegie Mellon University, USA*
{yumouw, pcarvalh, jstamper}@andrew.cmu.edu

Abstract: GPT has become nearly synonymous with large language models (LLMs), an increasingly popular term in AIED proceedings. A simple keyword-based search reveals that 61% of the 76 long and short papers presented at AIED 2024 describe novel solutions using LLMs to address some of the long-standing challenges in education, and 43% specifically mention GPT. Although LLMs pioneered by GPT create exciting opportunities to strengthen the impact of AI on education, we argue that the field's predominant focus on GPT and other resource-intensive LLMs (with more than 10B parameters) risks neglecting the potential impact that small language models (SLMs) can make in providing resource-constrained institutions with equitable and affordable access to high-quality AI tools. Supported by positive results on knowledge component (KC) discovery, a critical challenge in AIED, we demonstrate that SLMs such as Phi-2 can produce an effective solution without elaborate prompting strategies. Hence, we call for more attention to developing SLM-based AIED approaches.

Keywords: Small Language Models, Accessible AIED, KC Discovery

1. Introduction

It is an exciting time for AIED. Technological breakthroughs in large language models (LLMs) (Brown et al., 2020) have provided unprecedented opportunities for AIED researchers and practitioners to solve some of the long-standing challenges in the field (Kasneci et al., 2023). The excitement is aptly exemplified by the community's fast adoption of LLMs in AIED research—of the 76 long and short papers accepted for AIED 2024, 61% (47 papers) describe innovative solutions using LLMs, as revealed by a simple keyword-based search in the proceedings (Olney et al., 2024). Among the ever-expanding constellation of available LLMs, the GPT family, including ChatGPT (OpenAI, 2022) and GPT-4 (OpenAI, 2023), appears to be the community's favorite: 33 of the 47 papers (70%) adopting LLMs also mention GPT. Although LLMs pioneered by GPT herald exciting possibilities to reinforce AI's positive influence on education, we argue that the community's predominant focus on GPT and other similar resource-intensive gigantic language models (with more than ten billion parameters) risks neglecting the critical impact that small language models (SLMs) can make in creating equitable and accessible education central to the mission of AIED.

The definition of SLMs is constantly changing as new technologies emerge to shape the landscape of language models. The BERT model (Devlin et al., 2019) in its largest configuration, for example, has 340 million parameters—an overwhelming amount in 2018 but only a fraction by today's standard. In relation to the current state of the art, we consider a language model small if it has fewer than ten billion parameters and requires modest hardware resources, such as a consumer-grade GPU. Canonical examples of SLMs include Llama-2 7B (Touvron et al., 2023), Mistral 7B (Jiang et al., 2023), and Phi-2 (Javaheripi et al., 2023). Phi-2, a lightweight but capable model that has only 2.7B parameters, might be a particularly good fit for the AIED community and the range of problems we are trying to address. Trained on high-quality "textbook-like" data (Gunasekar et al., 2023), Phi-2 subsumes deep knowledge about various academic disciplines and aligns better with educational contexts, which require precision and reliability, than other SLMs trained on mixed-quality data sourced from the

Internet. Its smaller size also enables local deployment on consumer-grade hardware, desirable for most educational settings where computational resources are limited.

Educational institutions operate under distinct constraints that make their AI implementation needs different from those of commercial environments. Budget limitations, technical infrastructure, privacy requirements, and equity considerations all influence technology adoption in educational settings (Reich & Ito, 2017). GPT-scale LLMs typically require substantial computational resources for local deployment or incessant API costs for cloud access, not affordable to all teachers or students (Kasneci et al., 2023). SLMs, however, only require a fraction of the resources entailed by LLMs and can be deployed on modest hardware at a much lower cost—Phi-2's 2.7 billion parameters only require about 5.4 GB of memory for storage with a 16-bit representation of floating-point numbers[1], which can fit comfortably to a consumer-grade GPU.

One argument that justifies the higher costs of GPT-scale LLMs is their superior performance in various tasks. However, we argue that the more affordable and accessible SLMs can also deliver impressive results if we manage to exploit their potential adequately. In Section 3, we present a case study of knowledge component (KC) discovery (Koedinger et al., 2012), a critical challenge in AIED, and describe our unique solution using Phi-2 (Wei et al., 2025). Our approach makes creative use of Phi-2 as a probability machine to measure question similarity and applies a clustering algorithm to identify questions belonging to the same KC; results on two datasets show that instructors can better predict student performance using the KCs generated by our approach than using those produced by experts or the more powerful GPT-4o. These positive findings from the case study reinforce our position that **small language models such as Phi-2 can provide effective solutions to critical AIED problems and hold great promise as a catalyst for inclusive, personal, and ethical education equitably accessible to teachers and students**.

2. Background

2.1 The Rise of Large Language Models in Education

The field of education has tremendously benefited from the advances in natural language processing (NLP) in recent decades, which have evolved from rule-based approaches to statistical methods and eventually to neural-network models (Litman, 2016). Early educational applications used relatively simple NLP techniques for tasks such as automated essay scoring (Shermis & Burstein, 2013); more recent work, however, uses advanced language models to tackle increasingly complex challenges in education.

Introduced in 2017, the Transformer architecture (Vaswani et al., 2017) enables researchers to build more sophisticated language models with enhanced language understanding and generation capabilities. Together with more efficient hardware and better available corpora, this architectural innovation spurred the development of models with progressively larger parameter counts—some prominent milestones include GPT-3 (Brown et al., 2020) (175B parameters), PaLM (Chowdhery et al., 2022) (540B parameters), and GPT-4 (OpenAI, 2023) (estimated 1.76T parameters). These gigantic language models have demonstrated remarkable capabilities across various educational applications, including but not limited to hint creation (Pardos & Bhandari, 2023), question generation (Sarsa et al., 2022), and KC discovery (Moore et al., 2024).

Concomitant to the development of more capable models is the emphasize of scaling—increasing model size, training data, and computational resources—as the primary mechanism for improving model performance (Kaplan et al., 2020). This scaling law suggests that many unexpected capabilities can emerge as model size increases, with larger models generally outperforming smaller ones across diverse tasks (J. Wei et al., 2022). While the successful application of the scaling law has nearly depleted the available benchmarks to

[1] 2.7B float numbers require 16×2.7B=43.2B bits, which translate to 43.2B/8B = 5.4 gigabytes if 8 bits make 1 byte.

measure the progress of LLMs, urging the development of the "Humanity's Last Exam"[2], it has also raised the computational and financial requirements that prevent resource-constrained educational institutions from equitably using LLMs, and necessitated stricter, more private access to the source code and training data that could have helped researchers build more effective AIED tools. Moreover, the community's widespread predilection for large and even larger models can exacerbate the danger of overlooking the impact that SLMs can make in providing effective and accessible AIED solutions.

2.2 The Potential of Small Language Models in Education

In contrast to the scaling efforts, researchers have also developed smaller and more efficient models that challenge the dominance of scaling as the only way to attain good performance. More recently, models like Phi-2 (2.7B parameters) have demonstrated that careful data curation and innovative training methodologies can produce surprisingly capable models at significantly smaller scales (Javaheripi et al., 2023).

Developed by Microsoft Research, Phi-2 is an epitome of efficient language models. This SLM is built on the standard Transformer decoder-only architecture and is trained with the conventional next-token prediction objective. What makes it special, however, is not architectural innovations but the unique training methodology used. Unlike many larger models trained on vast but heterogeneous corpora sourced from the Internet, Phi-2 was trained predominantly on what the researchers call "textbook-quality data" (Gunasekar et al., 2023)—carefully curated content with an emphasis on educational materials, synthetic texts designed for reasoning capabilities, and filtered web content with high educational value.

This unique training methodology, which ranks data quality higher than quantity, results in an efficient SLM that is particularly useful for educational applications. In competitive benchmarks that evaluate reasoning skills in math (GSM8k (Cobbe et al., 2021)) and coding (HumanEval (Chen et al., 2021), MBPP (Austin et al., 2021)), Phi-2 substantially outperformed Mistral 7B (Jiang et al., 2023) and Llama-2 13B (Touvron et al., 2023), which are 1.6× and 3.8× larger than Phi-2. Compared to the 25× larger Llama-2 70B (Touvron et al., 2023), Phi-2 achieved significantly better performance in coding and demonstrated comparable reasoning skills in math (Javaheripi et al., 2023). In the MMLU benchmark (Hendrycks et al., 2021), which assesses language model knowledge in 57 academic subjects, Phi-2 outperformed Llama-2 13B (54.8) and achieved a score (56.7) comparable to that achieved by Mistral 7B (60.1).

From a computational efficiency perspective, Phi-2 also offers distinct advantages for educational applications. Requiring approximately 5.4 GB of memory for storage (with additional memory for inference), Phi-2 can be deployed on consumer-grade hardware with modest requirements (the conventional 16-GB GPU), enabling local inference without cloud infrastructure dependencies. This flexibility in deployment helps reduce the first digital divide (Attewell, 2001) that prevents resource-constrained schools from using the latest AI tools, and protects student privacy (Prinsloo & Slade, 2017) by not requiring student data to be shared with a third party.

Phi-2's solid results on academic benchmarks and modest requirements on computer hardware make it a competitive alternative to gigantic language models that entail substantial computational resources and provoke critical privacy concerns. Its extensive pre-training on high-quality textbook-like data makes Phi-2 particularly tuned to educational applications. In what follows, we describe a concrete case study in which we creatively used Phi-2 to design a KC discovery algorithm that outperformed instructional experts and its GPT counterpart.

3. Case Study: Knowledge Component Discovery

Representing specific concepts or skills that students acquire through learning to perform a task or solve a problem, knowledge components (KCs) are essential elements in the KLI framework (Koedinger et al., 2012) that help instructors assess student learning. Traditionally, instructional experts are elicited to participate in Cognitive Task Analysis (CTA) (Clark et al.,

[2] https://agi.safe.ai/

2008) to identify the KCs associated with each assessment item, but CTA incurs considerable time and labor cost (Stamper et al., 2011) even when applied to moderately sized question banks. The accelerating adoption of AI in education aggravates the burden on instructional designers, who are overwhelmed by the growing amount of AI-generated questions that each needs to be analyzed by hand.

To address this challenge, a recent approach (Moore et al., 2024) uses GPT-4 (OpenAI, 2023) to extract KCs from multiple-choice questions (MCQs). The authors devised elaborate prompting strategies to ask GPT-4 to simulate instructional experts or textbook authors. Although in an evaluation study, the majority of the three participants preferred GPT-generated KCs to those designed by experts for more than 60% of the evaluated questions, this approach produced KC labels with slightly different wording for questions that instructors think should belong to the same KC (Moore et al., 2024). In our replication of their study using more advanced GPT-4o, the most intelligent non-reasoning LLM offered by OpenAI, we obtained 614 unique KC labels for 630 MCQs from the same e-learning dataset[3] used by Moore et al., 2024. The large number of KC labels comparable to the number of questions suggests that some labels can be merged. In fact, we discovered that GPT-4o had produced unnecessarily refined labels (e.g., "Analyze CTA", "Analyze CTA in E-learning", and "Analyze CTA methodologies") that could have been merged.

In our recent work proposing a new KC discovery method called KCluster (Wei et al., 2025), we demonstrate that exploiting the native potential of a language model as a "probability machine" rather than the more conventional text generation capabilities can lead to a strong KC discovery algorithm even with SLMs such as Phi-2. The core idea is that language models can induce a novel measure of question similarity, which a clustering algorithm can use to identify groups of similar questions that are likely to share the same KC. Specifically, for an arbitrary pair of questions $\boldsymbol{q_s}$ and $\boldsymbol{q_t}$ from a large collection of questions, we evaluate the change in log-probability of $\boldsymbol{q_s}$ with and without the presence of $\boldsymbol{q_t}$:

$$\Delta\,(\boldsymbol{q_s}, \boldsymbol{q_t}) \stackrel{\text{def}}{=} \mathbf{log\,Pr}(\boldsymbol{q_s} \mid \boldsymbol{q_t}) \;-\; \mathbf{log\,Pr}(\boldsymbol{q_s})$$

and define a novel measure of question similarity called "congruity":

$$\text{Congruity}\,(\boldsymbol{q_s}, \boldsymbol{q_t}) \stackrel{\text{def}}{=} \frac{\mathbf{1}}{\mathbf{2}}\,[\Delta\,(\boldsymbol{q_s}, \boldsymbol{q_t}) + \Delta\,(\boldsymbol{q_t}, \boldsymbol{q_s})]$$

The formula for Δ is mathematically equivalent to the pointwise mutual information (PMI) (Church & Hanks, 1989) between two questions (instead of words). The idea was nevertheless inspired by word collocations. We postulate that if one question increases the likelihood of another question appearing, the two questions are congruent and likely to relate to the same KC; on the other hand, if one question barely changes or even decreases the probability that another question occurs, the two questions are incongruent and likely to belong to different KCs. SLMs such as Phi-2, which were *specifically* trained to calculate next-token probabilities, are exceptional at evaluating conditional log-probabilities of the form $\mathbf{log\,Pr}(\boldsymbol{q_s} \mid \boldsymbol{q_t})$. An algorithm that uses Phi-2 to calculate various required probabilities is described in the original paper (Wei et al., 2025).

We evaluated our approach against instructional experts and our replication of the previous study (Moore et al., 2024) using more advanced GPT-4o, on two datasets collected in a graduate e-learning course taught by two different instructors in 2022 and 2023[4]. A common practice to compare different KC discovery approaches is to fit an Additive Factors Model (AFM) (Cen et al., 2007) with the KCs generated by each method to student response data; a better KC discovery approach should allow an instructor to predict student responses with a lower root mean square error (RMSE). On the 2022 dataset, KCluster generated 114 KCs (comparable to the 101 KCs in the best expert KC model) and achieved an RMSE of 0.4220, outperforming both experts (0.4235) and GPT-4o (0.4395); likewise, on the 2023 dataset, KCluster generated 92 KCs (comparable to the 75 KCs in the best expert KC model) and scored 0.4066, leading both experts (0.4075) and GPT-4o (0.4101). We did not merge similar KC labels generated by KCluster unless they were identical. Notably, GPT-4o, a highly capable LLM, performed the worst on the two distinct datasets. This strengthens our claim

[3] https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=5426

[4] https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=5843

that **SLMs can also deliver superior results if their potential is adequately exploited**.

4. Conclusion

Through this forward-looking paper, we did not argue that the AIED community should eschew LLMs in favor of their more efficient counterparts, nor did we suggest that SLMs are capable of everything LLMs can do. Similar to many AIED researchers, we share the excitement about the complementary development of both lines of NLP research and their potential application to education. However, to empower teachers and students for an equitable future, the promise of SLMs in providing accessible AIED solutions is not to be neglected. As shown in the case study and more in our recent paper (Wei et al., 2025), an innovative exploitation of SLM's potential can deliver superior results than the standard use of LLMs based on intensive prompt engineering. **We urge the AIED community to reconsider, next time when making a convenient API call to an LLM, whether it endangers the accessibility to the target audience, who may actually benefit from an SLM**.

References

Attewell, P. (2001). Comment: The first and second digital divides. *Sociology of Education*, 74(3), 252–259. Retrieved March 12, 2025, from http://www.jstor.org/stable/2673277

Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C., Terry, M., Le, Q., & Sutton, C. (2021). Program synthesis with large language models.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

Cen, H., Koedinger, K. R., & Junker, B. (2007). Is over practice necessary? improving learning efficiency with the cognitive tutor through educational data mining. *Proceedings of the 2007 Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work*, 511–518.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., ... Zaremba, W. (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*. https://arxiv.org/abs/2107.03374

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., ... Fiedel, N. (2022). Palm: Scaling language modeling with pathways.

Church, K. W., & Hanks, P. (1989). Word association norms, mutual information, and lexicography. *27th Annual Meeting of the Association for Computational Linguistics*, 76–83.

Clark, R. E., Feldon, D., van Merriënboer, J. J. G., Yates, K., & Early, S. (2008). Cognitive task analysis. In J. M. Spector, M. D. Merrill, J. J. G. van Merriënboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd, pp. 577–593).

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*. https://arxiv.org/abs/2110.14168

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), 4171–4186.

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Giorno, A. D., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Behl, H. S., Wang, X., Bubeck, S., Eldan, R., Kalai, A. T., Lee, Y. T., & Li, Y. (2023). Textbooks are all you need.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding [arXiv:2009.03300]. *Proceedings of the International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2009.03300

Javaheripi, M., Bubeck, S., Abdin, M., Aneja, J., Bubeck, S., Mendes, C. C. T., Chen, W., Giorno, A. D., Eldan, R., Gopi, S., Gunasekar, S., Javaheripi, M., Kauffmann, P., Lee, Y. T., Li, Y., Nguyen,

A., de Rosa, G., Saarikivi, O., Salim, A., ... Zhang, Y. (2023, December). Phi-2: The surprising power of small language models.
Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavril, T., Lachaux, M.-A., Massiceti, D., Rio, J., Lambert, R., Bhosale, S., Aminov, S., Kool, W., Everett, R., ...Calandriello, J. (2023). Mistral 7b. *arXiv preprint arXiv:2310.06825*. https://arxiv.org/abs/2310.06825
Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models.
Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., ... Kasneci, G. (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. https://doi.org/https://doi.org/10.1016/j.lindif.2023.102274
Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, 36(5), 757–798. https://doi.org/https://doi.org/10.1111/j.1551-6709.2012.01245.x
Litman, D. (2016). Natural language processing for enhancing teaching and learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1). https://doi.org/10.1609/aaai.v30i1.9879
Moore, S., Schmucker, R., Mitchell, T., & Stamper, J. (2024). Automated generation and tagging of knowledge components from multiple-choice questions. *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, 122–133. https://doi.org/10.1145/3657604.3662030
Olney, A. M., Chounta, I.-A., Liu, Z., Santos, O. C., & Bittencourt, I. I. (Eds.). (2024). *Artificial Intelligence in Education: 25th International Conference*, AIED 2024, Recife, Brazil, July 8–12, 2024, Proceedings, Part II (Vol. 14830). Springer Nature Switzerland.
OpenAI. (2022, November). Chatgpt: Optimizing language models for dialogue [Accessed: March 2025]. https://openai.com/blog/chatgpt/
OpenAI. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
Pardos, Z. A., & Bhandari, S. (2023). Learning gain differences between chatgpt and human tutor generated algebra hints.
Prinsloo, P., & Slade, S. (2017). An elephant in the learning analytics room: The obligation to act. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 46–55.
Reich, J., & Ito, M. (2017). From good intentions to real outcomes: Equity by design in learning technologies. In W. Fitzgerald, J. Burns, N. Sonwalkar, & J. Urry (Eds.), The digital learning challenge: Obstacles to educational uses of copyrighted material in the digital age (pp. 1–42). *The Digital Media; Learning Research Hub*. https://clalliance.org/publications/good-intentions-real-outcomes-equity-design-learningtechnologies/
Sarsa, S., Denny, P., Hellas, A., & Leinonen, J. (2022). Automatic generation of programming exercises and code explanations using large language models. *Proceedings of the 2022 ACM Conference on International Computing Education Research* - Volume 1, 27–43.
Shermis, M., & Burstein, J. (2013). Handbook of automated essay evaluation: Current applications and new directions. *Journal of Writing Research*, 5(2), 239–243.
Stamper, J. & Koedinger, K. (2011). Human-Machine Student Model Discovery and Improvement Using DataShop. In Biswas, G., Bull, S., Kay, J., and Mitrovic, A., eds., Artificial Intelligence in Education, 353–360. Berlin, Heidelberg: Springer. ISBN 978-3-642-21869-9.
Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., ... Zhu, W. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*. https://arxiv.org/abs/2307.09288
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30 (nips 2017)* (pp. 5998–6008). Curran Associates, Inc.
Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models [Survey Certification]. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=yzkSU5zdwD
Wei, Y., Carvalho, P., & Stamper, J. (2025). KCluster: An LLM-based Clustering Approach to Knowledge Component Discovery. In Mills, C., Alexandron, G., Taibi, D., Bosco, G. L., and Paquette, L., eds., Proceedings of the 18th International Conference on Educational Data Mining, 228–240. Palermo, Italy: International Educational Data Mining Society.

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

SRL profiles in math problem-solving: The Essential Role of Monitoring and Translating for Outcomes and Beliefs

Nidhi NASIAR[a*], Jaclyn OCUMPAUGH[b] , Ryan S. BAKER[c], Caitlin MILLS[d], Samuel RHODES[e], Sheela SETHURAMAN[e]
[a]*University of Pennsylvania, USA*
[b]*University of Houston, USA*
[c]*Adelaide University, Australia*
[d]*University of Minnesota, USA*
[e]*CueThink Inc.,USA*
*nasiar@upenn.edu

Abstract: Self-regulation positively impacts learning, which has prompted efforts to detect and foster SRL strategies. We used Gaussian Mixture Modeling to cluster students on five SRL strategies in problem solving (PS)—measured via SRL detectors grounded in Winne's SMART model (2017). Seven distinct SRL profiles emerged from the data, which were examined for their relation to math beliefs, anxiety, and PS measures. Findings show that SRL profiles with high proficiency across multiple SRL skills achieved higher accuracy, spent more time on pre-tests, and reported more opportunities to share their math thinking. These students also frequently engaged in monitoring and translation. Notably, profiles with strong assembling skills underperformed peers who balanced SRL strategies with relatively higher monitoring and translating. Overall, these results highlight the dynamic relationship between SRL and math beliefs in PS, and suggest emerging profiles to design tailored interventions.

Keywords: Self-regulated learning, clustering, mindset, math beliefs

1. Introduction

Self-regulation is important for effective learning and academic achievement. Engaging in self-regulated learning (SRL) improves outcomes; yet, fostering SRL skills in students remains challenging and requires targeted, personalized support (Schunk, 2001). Recent work on automated SRL detection has leveraged trace data to capture fine-grained behaviors (Azevedo et al., 2012), with an increasing shift in focus from measuring to fostering SRL skills (Bouchet et al., 2013; Ting-Chia, 2024). Automated detectors can track individual SRL behaviors in real time and inform adaptive scaffolds. However, designing separate scaffolds for each SRL skill detector is resource-intensive and may overlook how strategies naturally co-occur in problem-solving (PS). Instead, grouping learners based on similar SRL behavior profiles offers a scalable approach to personalized support.

One way to identify learners with common SRL behaviors, as explored in multiple projects, is through clustering techniques (Gasevic et al., 2017), which has provided insights that advanced both theoretical and empirical understanding of SRL. A growing body of work employs various clustering methods to form groups for targeted support (Bouchet et al., 2013; Ting-Chia, 2024). These clustering approaches have shown connections to learner engagement and achievement, with profiles characterized by SRL difficulties correlating with poorer outcomes (Farhana et al., 2021). Grounding clustering analyses in established SRL theories enhances the effectiveness and relevance of the research and informs scaffold design. Winne's SMART model (2017) with cognitive behaviors of searching, monitoring,

assembling, rehearsing, and translating, is widely applied in empirical SRL studies (Azevedo et al., 2012; Farhana et al., 2021). Our study uniquely integrates a suite of automated SRL detectors based on Winne's (2017) model to capture students' SRL behaviors, and then clusters detectors' outputs to identify distinct learner subgroups for targeted support.

The relationship between math beliefs, anxiety, and SRL is complex and cyclic (Maloney & Beilock, 2012). Effective PS involves more than cognitive strategies; it is heavily influenced by learners' beliefs and affect (Pekrun et al., 2017). Students' epistemic beliefs, such as viewing math merely as rule memorization, shape their SRL behaviors, where positive beliefs improve persistence and strategizing while negative ones reduce them (Dweck, 2014). Math anxiety further disrupts strategy selection and progress monitoring (Maloney & Beilock, 2012). However, limited work has examined combinations of SRL skills necessary for effective PS and math beliefs. This study seeks to address that gap by investigating how SRL behaviors co-occur in math PS and their interaction with students' math mindset and beliefs within CueThink, an online math platform. The study is driven by two main questions: i) What distinct SRL profiles emerge in math PS (based on distinct SRL strategies' usage)?, ii) How are these SRL profiles associated with measures of PS, math beliefs, and anxiety?

2. Methods

2.1. Learning Platform: CueThink

This study is based on CueThink, an online platform that scaffolds math PS and promotes regulation. In CueThink, students complete "Thinklets" math problems that follow 4 phases based on Winne (2017): in the Understand phase, they identify what's given and asked, and note observations; in Plan, they select strategies and outline steps for solving; in Solve, they record screencast explanations using mathematical tools; and in Review, they refine their solutions via a guided checklist. Students move freely between phases as needed. Data was collected from 230 7^{th}-8^{th} grade students at a suburban Eastern U.S. middle school in 2022–2023 for 768 problems. Pre- and post-tests were administered at the start and end of the year.

2.2. SRL Detectors used for Clustering

SRL strategies were detected from Zhang et al.'s (2022a) detectors developed using qualitatively coded SRL behaviors from trace data, aligned with Winne's SMART model (2017). They were validated for generalizability using 10-fold student-level cross-validation, using Area Under the ROC Curve (AUC) as primary metric (chance AUC = 0.5; perfect = 1.0), and checked for algorithmic bias (see Zhang et al., 2022). The SRL detectors were developed for i) numerical representation (NR; AUC = 0.894), showing learners' understanding of numerical components, and ii) contextual representation (CR; AUC = 0.813) for noting contextual details-- both classified as Assembling. iii) Data transformation (DT; AUC =0.815), as students manipulate data to find a solution (Translating). iv) Following plans (FP; AUC = 0.808) as correctly including chosen strategies into the plan (Monitoring), and v) incorporating information (II; AUC = 0.803) as correctly including previously assembled information into the plan (Rehearsing). These SRL detectors identified students' SRL processes in the Understand and Plan phase within each problem they solved, using their interaction data. The usage of the 5 SRL strategies was calculated for every student across all their solved problems based on the probability outputs of the 5 SRL detectors, which was then used as input to clustering.

2.3. Research Instruments

Three PS measures were used: i) *Accuracy Score*, with one PS item given per trimester, with items from a standard curriculum. ii) *Normalized learning gains* (LG), calculated as (post-pre)/(1-pre) when post>pre and (post-pre)/pre when post<=pre. iii) *Duration of Pre & Post test*,

seconds spent from the start to the end time on test. We measured *math beliefs* using three SRL subscales (0–100) from the abbreviated Indiana Math Beliefs Scale (Cipora et al., 2015): Speed Orientation (4-items), the degree to which a student values thinking deeply (100) over the speed of solving math problems (0); Nature of Math (7-item), whether students see math from a constructionist lens (100) or as rigid and non-creative (0); Math Share, a single-item asking about the opportunities that students have to talk about math. *Math anxiety* was assessed with the 9-item Modified Abbreviated Math Anxiety Scale (Carey et al., 2017), slightly modified for American English, scored 0–100, with two factors—Learning and Evaluation.

2.4. Gaussian Mixture Modeling (GMM) clustering

We used Gaussian Mixture Modeling (GMM) clustering, with oval-shaped covariance and an outlier threshold of 0.7, to cluster the proportion of strategy usage based on each student's 5 SRL detector outputs. The function returned probabilities of a data point belonging to each cluster. We evaluated different N of clusters using the Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC), and Silhouette score. Overall, 7 clusters performed best.

Differences in PS and survey measures across SRL clusters were tested using the Kruskal–Wallis (KW) H test, a nonparametric test for non-normal continuous data. A significant result indicated that at least one cluster differed. For those measures, pairwise Mann–Whitney U tests identified the cluster pairs that differed, limiting comparisons to reduce Type I error. Benjamini–Hochberg correction was applied to control the false discovery rate and mean ranks were used to determine which cluster exhibited higher/lower values for the measures.

3. Results

3.1. RQ1: Examining SRL Profiles from GMM Clustering

Students were clustered by their SRL strategy use; Table 1 lists each cluster's median of variables (NR, CR, DT, DP, II). Each median was classified relative to its overall detector distribution percentiles (Table 2): low(<25th), moderately low(25th to <50th), intermediate(50th), moderately high(>50th to <75th), and high(>=75th). Cluster names reflect these labels—"advanced" for high or moderately high levels, "intermediate" for median levels, and "beginner" for moderately low levels—omitting skills at low levels except in clusters with uniformly minimal use. E.g., A's NR median(0.87) is between 50th(0.77) and 75th(0.95) percentiles, and is labeled moderately high. Clusters are ordered from the most advanced to emerging SRL strategists.

Table 1. *Cluster Medians for 5 SRL strategies, shaded as per detector distribution of each strategy (see Table 2), with darker shading indicating higher values*

		Assemble		Translate	Monitor	Rehearse
	N (%)	NR	CR	DT	FP	II
Cluster A	82 (36.1)	0.87	0.85	0.84	0.99	1
Cluster B	45 (19.8)	0.73	0.69	0.75	0.95	0.99
Cluster C	52 (22.9)	0.77	0.61	0.64	0.78	0.92
Cluster D	9 (4.0)	0.92	0.91	0.2	0.44	0.93
Cluster E	6 (2.6)	0.93	0.2	0.31	0.52	0.85
Cluster F	27 (11.9)	0.63	0.57	0.37	0.54	0.91
Cluster G	6 (2.6)	0.22	0.02	0.06	0.7	0.11

Table 2. *Detector output distribution*

	Assemble		Translate	Monitor	Rehearse
Level	**NR**	**CR**	**DT**	**FP**	**II**
High.	x ≥ 0.95	x ≥ 0.87	x ≥ 0.85	x ≥ 0.99	x=1
Mod high	0.77 ≤ x <0.95	.69 ≤ x<0.87	0.7 ≤ x<0.85	0.93 ≤ x <0.99	0.99 ≤ x <1
Mod low	0.63 ≤ x <0.77	0.51 ≤x <0.69	0.47 ≤ x <0.7	0.74 ≤ x <0.93	0.9 ≤ x < 0.99
low	x < 0.63	x < 0.51	x < 0.47	x < 0.74	x < 0.9

Cluster A (Advanced SRL strategists) is characterized by high levels of CR, FP, and II, and moderately high usage of NR, and DT. Learners have high usage of SRL, including rehearsing (using information from their plan), monitoring (applying selected strategies), and assembling contextual information. While overall SRL engagement is high, some variability was found in NR, CR, and DT. All students consistently show high levels of FP and II.
Cluster B (Advanced Monitors and Translators with intermediate contextual assembling and rehearsing) is characterized by moderately high DT and FP, medium levels of CR and II, and moderately low NR. These students engage in data manipulation (translation) and follow their plan. Strengthening their assembling skills for NR could create a more balanced SRL profile for these students. Low variability is observed across students.
Cluster C (Intermediate Numerical Assemblers, beginner for other strategies) is characterized by medium values of NR with moderately low CR, DT, FP, and II. These students engage in average amounts of NR math word problems (assembling), but less monitoring, rehearsing, and translating. Students had low variability across strategies.
Cluster D (Advanced Assemblers with beginner rehearsing) is characterized by high CR, moderately high NR, moderately low II, and low DT and FP. These students frequently engage in problem representation and assembling, and sometimes use rehearsing strategies.
Cluster E (Advanced Numerical Assemblers only) is characterized by moderately high NR but low levels of all other strategies. Students in this group often assemble numerical information from math problems, but they do not independently show contextual understanding, DT, FP, and II. Low to moderate variability is observed across SRL strategies, other than for DT.
Cluster F (Beginner Numerical and Contextual Assemblers) is characterized by moderately low NR and CR, and low DT, FP, and II. Low variability is seen across strategies, except II.
Cluster G (Emerging SRL Strategists) is characterized by the lowest usage of all SRL strategies. This group represents students who are failing to use these SRL strategies in math PS. Low variability is observed across all strategies except NR.

Overall, Clusters A and B show well-balanced and comprehensive SRL profiles, with a range of key processes of assembling, monitoring, rehearsing, and translating. Clusters C, D, and E exhibit a mixed pattern of strategy utilization, with learners exhibiting strengths in some strategies but struggling to use others (or not realizing they are useful). Clusters F and G show minimal usage of SRL strategies, and are likely require support to develop SRL skills.

3.2. RQ2: Association of PS, beliefs, and anxiety measures to SRL profiles

The KW H statistic (Table 3) is used to test differences in the survey measures among clusters.

Table 3. *Results from Kruskal-Wallis (KW) H test for measures. Significant values are shaded*

Type	Measures	Pretest		Posttest	
		H	p	H	p
Problem Solving (PS)	Duration (Dur)	20.91	0.002	7.38	0.288
	Accuracy (Acc)	9.21	0.162	15.35	0.010
	Norm. Learning Gains (LG)	NA	NA	6.48	0.372
Anxiety	mAMAS Anxiety Learning (AxL) mAMAS	5.80	0.446	4.07	0.67
	Anxiety Evaluation (AxEV)	1.41	0.965	2.18	0.90
Beliefs	IMBS Speed Orientation (Speed)	3.04	0.804	18.79	0.005
	IMBS Nature of Math (Nature)	7.04	0.317	12.90	0.045
	IMBS Math Share	6.23	0.398	13.35	0.038

The PS measure of duration of the pre-test was significantly different (p=0.002) across at least one cluster pair. By the post-test, differences in duration between clusters were not observed. There was no difference among clusters in accuracy in the pre-test, however, there was a significant difference between clusters for the post-test (p=0.010).

No cluster differences were observed for learning gains or either form of anxiety. However, all three math belief constructs -- measure of speed orientation, nature of math, and math share -- had significant post-test differences among clusters (p=0.005, 0.045, 0.038) but not the pre-test (p=0.804, 0.317, 0.398), meaning the SRL usage clusters had different beliefs

on what math was about, growth mindset in math, sharing math ideas, and on importance of thinking deeply over quickly on a math problem. Differences at post but not pre suggest that these beliefs may have been influenced by students' experience of using SRL strategies to solve problems in CueThink for a year, although a between-conditions comparison would be needed to be certain. Next, pairwise Mann-Whitney U tests were conducted for significant KW test measures (i.e. pre-test duration, and post-test measures of accuracy, speed, and nature).

Table 4. *Pairwise Mann-Whitney U tests only with p <0.05; significant values after B&H correction are shaded.*

	PS (Pre)		**PS (Post)**		**Beliefs (Post)**					
	Dur		**Acc***		**Share**		**Speed**		**Nature**	
	U	p	U	p	U	p	U	p	U	p
A vs. B	-	-	-	-	1042.5	0.03	983.5	0.01	-	-
A vs. C	2555.5	0.002	1354.5	0.036	-	-	-	-	1073	0.01
A vs. D	445	0.04	-	-	-	-	397.5	0.02	-	-
A vs. E	370	0.015	116	0.049	-	-	101	0.03	-	-
A vs. F	1391	0.001	583.5	0.027	-	-	-	-	-	-
B vs. D	-	-	213	0.012	-	-	231.5	0.002	210	0.017
B vs. F	689.5	0.03	-	-	685.5	0.003	625	0.036	-	-
C vs. D	-	-	228	0.016	-	-	240.5	0.008	248	0.004
C vs. F	-	-	-	-	703.5	0.021	-	-	-	-
D vs. E	-	-	4	0.018	-	-	0	0.001	-	-
D vs. F	-	-	111	0.02	-	-	-	-	-	-
E vs. F	-	-	-	-	-	-	34	0.04	-	-

Pre Duration. Significant differences were found in the time taken to finish pre-test between clusters A and C, and A and F (see Table 4). Students in cluster A spent more time on the pre-test than those in C and F, suggesting that advanced SRL strategists invested more time during initial assessment. In contrast, students with lower and intermediate SRL strategy usage spent less time, possibly indicating less engagement or cognitive effort . These effects were not found for the duration of the post-test, however, perhaps due to familiarity with the test, or due to practice with similar math problems in CueThink for a year.

Accuracy at Post. There were significant post-test accuracy differences between clusters B-D, C-D, D-E, and D-F. Students in C, B, and F achieved significantly higher post-test accuracy than in D, indicating that students who balance assembling with other SRL strategies like translation and monitoring (Clusters C and B), or even beginners (Cluster F), have higher accuracy than those focusing solely on assembling skills (Cluster D). Conversely, Cluster D had higher post-test accuracy than E, suggesting that contextual understanding of the problem may support better performance.

Speed Orientation Post. For this measure, a higher value indicated greater preference for deep thinking over speed wile solving problems. Significant differences were found in the post-test between Clusters C-D, B-D, and E-D. Based on their mean ranks (B=23, C=26, D=4.5, E=3.5), students in C and B valued deep thinking over speed more than D, suggesting that C and B cluster students prioritize thoughtful PS while engaging deeply with concepts. In turn, D valued deep thinking more than E. Cluster E students, frequently engaged in NR but had low usage of other SRL strategies and preferred solved problems more quickly.

Math Share Post. Significant differences were found in the post-test Math Share scores between Cluster B and Cluster F (p=0.03). Students in B reported sharing their mathematical thinking in class more than Cluster F, based on their mean ranks (B=23, F=14). This suggests that students with lower levels of SRL strategy usage may perceive fewer opportunities to share their mathematical ideas in class.

Nature of Math Post. Significant differences were observed in post-test perceptions of the nature of mathematics between Clusters C and D (p=0.004), and Clusters B and D (p=0.017). Clusters C and B hold more constructivist views of mathematics (highlighting creative, exploratory, and interconnected nature of math) than Cluster D.

4. Discussion and Conclusion

This study identified 7 student SRL profiles, revealing differences across the use of 5 SRL strategies, math problem solving (PS), and beliefs. Key findings show that students in clusters with a higher proficiency and balance of SRL skills (cluster A and B) have better PS accuracy, spent more time on the pre-tests and reported sharing of their math thinking than clusters with lower usage of fewer SRL skills (cluster E and F). Results further show that these clusters A and B integrate considerable usage of monitoring and translating strategies, processes critical for learning (Winne et al., 2017), whereas other students (in cluster E and F) primarily engaged in NR and sometimes in CR (assembling) but not as much in other SRL strategies. Notably, even students with strong assembling (cluster D) underperformed peers who balanced SRL strategies with a relatively higher monitoring and translating (B, C, and F), illustrating the importance of a well-rounded and diverse SRL profile in general, and specifically these two SRL strategies, for math problem solving. Post-test differences in beliefs (e.g., cluster D showed a lower preference for deep thinking over speed in solving math than clusters B and C, but higher than cluster E; and cluster D exhibited more fixed-mindset, maladaptive math beliefs than B and C, who showed a growth mindset) further highlight the nuances of how SRL experiences shape math beliefs over time.

A limitation of the study is seen in the small sample sizes for Clusters D, E, and G, suggesting these results should be replicated in the future with larger datasets across more types of content and grade levels. Also, the absence of significant links between SRL strategies and anxiety contrasts with prior work (Pekrun et al., 2017), and merits further investigation. Longitudinal analyses could also clarify whether cluster membership shifts as students gain experience. Correlating to additional motivational and affective variables (e.g., self-efficacy) could enrich interpretation. Overall, this study highlights the dynamic interplay between SRL and evolving math beliefs, with potential to inform future studies for personalized learner support based on their SRL gaps and math beliefs. In addition, the ML-detector–based clustering pipeline can be used to automate support for students with distinctive SRL profiles.

References

Azevedo, R., Behnagh, R.F., Duffy, M., Harley, J.M., Trevors, G., 2012. Metacognition and self-regulated learning in student-centered learning environments, in: Theoretical foundations of learning environments. Routledge, pp. 171–197.

Bouchet, F., et al., 2013. Clustering and profiling students according to their interactions with an intelligent tutoring system fostering self-regulated learning. J. Educational Data Mining 5, 104–146.

Carey, E., Hill, F., Devine, A., Szűcs, D., 2017. The modified abbreviated math anxiety scale: A valid and reliable instrument for use with children. Frontiers in Psychology 8, 11.

Cipora, K., Szczygieł, M., Willmes, K., Nuerk, H.C., 2015. Math anxiety assessment with the abbreviated math anxiety scale: Applicability and usefulness. Frontiers in Psychology 6, 1833.

Dweck, C., 2014. Mindsets and math/science achievement .

Farhana, E., Potter, A., Rutherford, T., Lynch, C.F., 2021. Feedback and self-regulated learning in science reading. *International Conference on EDM, 820-826.*

Gasevic, D., Jovanovic, J., Pardo, A., Dawson, S., 2017. Detecting learning strategies with analytics: Links with self-reported measures and academic performance. J. Learning Analytics,113-128.

Maloney, E.A., Beilock, S.L., 2012. Math anxiety: Who has it, why it develops, and how to guard against it. Trends in cognitive sciences 16, 404–406.

Pekrun, R., Lichtenfeld, S., Marsh, H.W., Murayama, K., Goetz, T., 2017. Achievement emotions and academic performance: Longitudinal models of reciprocal effects. Child development 88, 1653–1670.

Schunk, D.H. 2001. Self-regulation through goal setting. ERIC Document Service, 462671.

Ting-Chia, H. S. U. (2024). Effects of the Self-Regulated Learning and Motivation on Learning Achievements of the Programming Courses. In *International Conference on Computers in Education*.

Winne, P.H., 2017. Learning analytics for self-regulated learning. Handbook of LA, 754, 241–249.

Zhang, J., Andres, J.M.A.L., et al., 2022a. Using machine learning to detect smart model cognitive operations in mathematical problem-solving process. J. Educational Data Mining 14, 76–108.

Effects of Gen AI in Enhancing Pre-service Teachers' Instructional Design Competency: A Quasi-Experimental Study

Le Zhang[a], Hongchao Peng[b*] & Mingrui Zhang[a]
[a]*Department of Educational Administration, East China Normal University, China*
[b]*School of Open Learning and Education, East China Normal University, China*
*hcpeng@dec.ecnu.edu.cn

Abstract: This study investigates the impact of Generative Artificial Intelligence (GenAI) on pre-service teachers' instructional design skills, distinguishing between immediate task performance and the development of lasting competency. Employing a quasi-experimental design, this study tracked 30 pre-service teachers from a university in China over a four-week period. Data on their instructional design tasks were collected via expert ratings and self-assessments at three stages: a pre-test, a post-test one week later with GenAI-supported intervention, and a delayed post-test one week later without the tool. The results revealed three key findings: (1) The GenAI intervention significantly enhanced participants' immediate instructional design performance.(2) These gains were sustained in the delayed post-test, suggesting a genuine development of underlying instructional design competency.(3) This positive effect, however, was not mediated by the participants' AI teaching self-efficacy. These findings provide crucial empirical evidence that GenAI can be an effective tool for fostering durable professional skills in teacher education, beyond simply aiding in-the-moment task completion.

Keywords: Gen AI, instructional design competency, pre-service teachers, effects of Gen AI

1. Introduction

Generative Artificial Intelligence (GenAI) represents a paradigm shift in computing, with many strong claims about its revolutionary impact on education (Chiu, 2024; Hodges & Kirschner, 2024). A primary area of interest is its application in instructional design, where research suggests GenAI can act as a powerful scaffold, particularly for novice teachers (Berg & du Plessis, 2023). It can support the entire design lifecycle, encompassing key stages such as learner analysis, objective setting, activity and material design, and instructional evaluation (Sailer et al., 2023; Hilliger et al., 2022).

However, a critical tension exists between this potential and the practical impact on teacher development. This is due to significant challenges, including the unreliability of AI outputs and the risk of fostering over-reliance, which may diminish users' critical thinking (Ruiz-Rojas, 2023; Chan & Hu, 2023). This issue is particularly salient for pre-service teachers and raises a crucial distinction: while GenAI may enhance immediate task performance, it is uncertain whether it fosters genuine professional competency (Hodges & Kirschner, 2024). A recent study highlighted this gap, finding performance differences between groups using GenAI but failing to address the underlying development of competency (Krushinskaia, 2024). Therefore, a key unresolved question is whether GenAI acts merely as a temporary support or as a catalyst for durable skill development. To explore why this distinction might exist, we turn to self-efficacy theory (Bandura, 1997). In education, teacher self-efficacy is a known predictor of instructional quality (Zee & Koomen, 2016), and recent studies have linked AI-specific self-efficacy to teachers' willingness to adopt and persist with AI tools (Yao & Wang, 2024). This suggests AI teaching self-efficacy could be a key mediating factor in this process.

Accordingly, this research addresses three core questions: Does the use of GenAI improve pre-service teachers' instructional design performance? If so, does this improvement persist after the technology is removed, indicating a genuine enhancement of instructional design competency? Does AI teaching self-efficacy mediate the impact of GenAI on the development of this competency?

2. Research Methodology

2.1 Participants and procedures

Participants were 30 pre-service chemistry education majors (aged 19–21) from a university in Shanghai, China. This homogenous cohort was selected because all had previously completed the same "Instructional Design and Practice in Information Technology" course, ensuring baseline familiarity with instructional design principles and AI concepts. Participants were randomly assigned to an experimental (n=15) or a control group (n=15).

2.2 Quasi-Experimental Research Design and Procedures

The research employed a three-phase, longitudinal quasi-experimental design (pre-test, intervention, delayed post-test) conducted over three weeks. The independent variable was the availability of GenAI support, while the primary dependent variable was instructional design competency, measured by expert ratings and self-assessments of participants' lesson plans. AI teaching self-efficacy was examined as a potential mediator.

The procedure unfolded in three distinct phases:

Week 1 (Pre-Test): Both groups completed an initial instructional design task without any GenAI support to establish baseline competency.

Week 2 (Intervention): Following a one-week interval, both groups received a brief, conceptual training on GenAI. This was done to ensure equal awareness and mitigate motivational biases (e.g., perceived resource inequality), thereby isolating the intervention's effect to the application of GenAI, not mere knowledge of it. The control group received no hands-on practice. Immediately after this training, all participants completed the AI teaching self-efficacy scale. Subsequently, the experimental group completed the second design task with GenAI support, while the control group completed it without.

Week 3 (Delayed Post-Test): After another one-week interval, both groups completed a final design task, again without access to the GenAI tool. This phase aimed to measure the retention of skills and assess whether competency gains persisted after the technological scaffold was removed.

2.3 Experimental Setup and Data Collection

We assessed pre-service teachers' instructional design competency through three inquiry-based tasks centered on core high school chemistry topics: "Detecting Iron in Food," "Measuring the pH of Rainwater," and "Understanding Automotive Energy Sources." The content and difficulty of these tasks were validated by an expert panel using the Delphi method. The tasks were intentionally designed to be challenging, creating a genuine need for GenAI support. The selected GenAI tool was Eduaide.AI (ai.eduku.cn), chosen for its specialization in instructional design and ease of access. To structure the design process, participants were provided a template based on the 5E Inquiry-Based Instructional model (Engage, Explore, Explain, Elaborate, and Evaluate), which included relevant tips and guidelines.

2.4 Measurement and Measuring tools

To evaluate instructional design competency, we utilized the 5E Inquiry-Based Instructional Design Evaluation Scale (5E ILP), which was developed by Goldston et al (2013). This

validated instrument, recommended for use with the U.S. Next Generation Science Standards (Goldston et al., 2010) , assesses inquiry-based lesson plans across five dimensions: Engage, Explore, Explain, Elaborate, and Evaluate. Each dimension is rated on a 5-point scale (1 = "very poor" to 5 = "excellent"). The scale demonstrated good reliability in this study, with a Cronbach's alpha of 0.77.

AI teaching self-efficacy was measured using a scale developed by Chou et al. (2024). This instrument assesses teachers' self-efficacy in AI-supported teaching through 25 items across five dimensions: Self-Affirmation, Teaching Enthusiasm, Diligence and Perseverance, Negative Awareness, and Positive Beliefs. Items were rated on a 5-point Likert scale (1 = "Strongly Disagree" to 5 = "Strongly Agree"). The scale showed good reliability, with a Cronbach's alpha of 0.73.

To ensure fidelity, both scales were translated into Chinese using a back-translation procedure (Duda et al., 1998).

3. Data Analysis and Results

3.1 Comparison of Pre-service Teachers' Baseline Instructional Design Competency

An independent samples t-test on pre-test scores confirmed that the experimental and control groups were equivalent at baseline. There were no significant differences in their instructional design competency, as measured by either expert ratings (t=1.60, p=.12) or self-assessments (t=0.28, p=.39). This established a valid foundation for comparing the effects of the post-intervention outcomes.

3.2 Exploring the Effect of Gen AI on Pre-service Teachers' Instructional Design

An independent samples t-test was conducted on the post-intervention scores to assess the impact of the GenAI intervention. The results revealed that the experimental group scored significantly higher than the control group on both expert ratings (M=76.41 vs. M=63.28, p= .008) and self-assessments (M=67.87 vs. M=56.40, p=.007). Furthermore, the magnitude of these differences was substantial. Large effect sizes were observed for both expert ratings (Cohen's d=1.77) and self-assessments (Cohen's d=2.36), indicating that the GenAI intervention had a strong, positive effect on improving the instructional design competency of pre-service teachers.

3.3 Verification of Substantive Enhancement in Instructional Design Competency

To assess the retention of skills, a delayed post-test was conducted one week later, during which neither group had access to GenAI. An independent samples t-test revealed that the experimental group continued to significantly outperform the control group on both expert ratings (M=74.63 vs. M=66.26, p=.008) and self-assessments (M=64.00 vs. M=55.47, p=.016). This lasting effect was substantial, with large effect sizes observed for both expert ratings (Cohen's d=1.07) and self-assessments (Cohen's d = 0.93). These findings indicate that the competency gains from the GenAI intervention were durable and persisted even after the technological support was removed.

3.4 *Mediating Role of AI Teaching Self-Efficacy*

To test whether AI teaching self-efficacy mediated the relationship between the GenAI intervention and instructional design competency, we conducted a mediation analysis using Hayes' PROCESS Model 4 with 2000 bootstrap resamples.

Table 1. Mediating Effect of AI Teaching Self-Efficacy

Path	c Total Effect	a	b	a*b (Indirect Effect)	Boot SE (Indirect Effect)	95% Boot CI (Indirect Effect)	c' (Direct Effect)
Gen AI (X) → AI Teaching Self-Efficacy (M) → Expert Rating (Y)	13.12**	12.67**	0.24	3.04	0.11	-0.01 ~ 0.44	10.07**
Gen AI (X) → AI Teaching Self-Efficacy (M) → Self-Assessment (Y)	11.47**	12.67**	-0.01	-0.14	0.08	-0.19 ~ 0.16	11.61**

**p<0.05 ** p<0.01*

As shown in Table 1, the analysis revealed no significant indirect effect of the intervention on competency through AI teaching self-efficacy. For expert ratings, the 95% bootstrap confidence interval for the indirect effect (a*b = 3.04) was [-0.012, 0.443], which contains zero. Similarly, for self-assessments, the confidence interval for the indirect effect (a*b = -0.14) also included zero [-0.190, 0.162]. Therefore, while the GenAI intervention had a significant direct effect on improving competency (see c' path in Table 1), the hypothesis that this effect is mediated by AI teaching self-efficacy was not supported.

4. Conclusion and Discussion

4.1 GenAI's Impact on Instructional Design Performance and Competency

The findings first confirm that GenAI significantly enhances the immediate instructional design performance of pre-service teachers. This aligns with research suggesting that large language models improve task performance by serving as a powerful knowledge repository and an interactive cognitive partner (Van Den Berg & Du Plessis, 2023). Through human-AI collaboration, participants engaged in knowledge retrieval and reconstruction, effectively supporting their design process (Baidoo-Anu & Ansah, 2023).

More importantly, this study demonstrates that these gains are sustainable, persisting even after the technological support was removed. This indicates a genuine development of underlying instructional design competency, a crucial distinction from mere performance enhancement. Therefore, this research provides crucial evidence that GenAI can be framed not merely as a performance tool, but as a developmental accelerator. It facilitates a transition from novice performance to expert-like competency by simultaneously modeling expertise and demanding active metacognitive participation from the learner.

4.2 The Unexpected Role of AI Teaching Self-Efficacy

Contrary to our hypothesis, AI teaching self-efficacy did not emerge as a significant mediator in the relationship between GenAI use and competency development. One plausible explanation is that the scaffolding provided by the GenAI tool was so direct and effective that it enhanced skills without needing to be channeled through the psychological construct of self-efficacy in this short-term context (Kirschner et al., 2006). It is also possible that the influence of self-efficacy requires a longer period to manifest or was overshadowed by the novelty and strength of the technological intervention itself (Mok & Moore, 2019).

4.3 Implications for Pre-service Teacher Education

Our findings offer significant implications. Theoretically, they highlight the critical difference between technology-aided performance and technology-driven competency development.

Practically, they suggest that the pedagogical approach to integrating GenAI is paramount. Merely providing access to GenAI is insufficient; teacher education programs should focus on training pre-service teachers in prompt engineering, critical evaluation of AI outputs, and the reflective integration of AI-generated content into their pedagogical frameworks. This approach is key to fostering durable professional skills rather than dependency.

4.4 Limitations and Future Directions

Building on this study's findings, several promising avenues for future research emerge to advance the understanding of GenAI's role in teacher education.

A primary direction for future research is to establish the generalizability of these findings. This would involve conducting replication studies with larger, more diverse participant samples across various institutional and cultural contexts. Furthermore, longer-term longitudinal studies are needed to track the development of both instructional competency and AI teaching self-efficacy over an entire academic term.

Another critical area for enhancement is methodological rigor, which would allow for a clearer isolation of causal mechanisms and a deeper understanding of the learning process. Such refinements could include: Incorporating pre-intervention surveys to assess participants' GenAI familiarity. This data would allow for this variable to be controlled as a covariate or used for subgroup comparisons (e.g., novices versus experienced users). Employing methods like self-report logs, think-aloud protocols, or system usage data to account for activities occurring between formal sessions, thus providing a more granular view of skill development.Utilizing a three-group design (e.g., GenAI use vs. conceptual training only vs. no-contact control) to more effectively disentangle the effects of active tool use from mere conceptual awareness.

Pursuing these directions will be crucial for developing a nuanced, evidence-based framework for leveraging GenAI to foster genuine, durable skills in the next generation of educators.

Acknowledgements

We gratefully acknowledge the support of the Smart Education Multimodal Research Laboratory. This research was funded by the Humanities and Social Science Fund of the Ministry of Education of China (Grant No. 24YJC880105), and in part by the Shanghai University Student Innovation and Entrepreneurship Training Program (Project No. 202510269119S). This work was also supported by the Special Fund of East China Normal University for Graduate Students to Attend International Conferences.

References

Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. Journal of AI, 7(1), 52-62. https://doi.org/10.61969/jai.1337500

Bandura, A., & Wessels, S. (1997). Self-efficacy (pp. 4-6). Cambridge: Cambridge University Press.

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. International Journal of Educational Technology in Higher Education, 20(1), 43. https://doi.org/10.1186/s41239-023-00411-8

Chiu, T. K. F. (2024). The impact of Generative AI (GenAI) on practices, policies and research direction in education: A case of ChatGPT and Midjourney. Interactive Learning Environments, 32(10), 6187–6203. https://doi.org/10.1080/10494820.2023.2253861

Chou, C.-M., Shen, T.-C., Shen, T.-C., & Shen, C.-H. (2024). Developing and validating an AI-supported teaching applications' self-efficacy scale. Research and Practice in Technology Enhanced Learning, 19, 035. https://doi.org/10.58459/rptel.2024.19035

Duda, J.,& Hayashi, C. T.(1998). Advances in Sport and Exercise Psychology Measuresment. Morgantown: Fitness Information Technology:471-483

Goldston, M. J., Dantzler, J., Day, J., & Webb, B. (2013). A Psychometric Approach to the Development of a 5E Lesson Plan Scoring Instrument for Inquiry-Based Teaching. Journal of Science Teacher Education, 24(3), 527–551. https://doi.org/10.1007/s10972-012-9327-7

Goldston, M. J., Day, J. B., Sundberg, C., & Dantzler, J. (2010). Psychometric analysis of a 5E learning cycle lesson plan assessment instrument. International Journal of Science and Mathematics Education, 8(4), 633–648. https://doi.org/10.1007/s10763-009-9178-7

Hilliger, I., Muñoz-Merino, P. J., De Laet, T., Ortega-Arranz, A., & Farrell, T. (Eds.). (2022). Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption: 17th European Conference on Technology Enhanced Learning, EC-TEL 2022, Toulouse, France, September 12–16, 2022, Proceedings (Vol. 13450). Springer Nature. https://doi.org/10.1007/978-3-031-16290-9

Hodges, C. B., & Kirschner, P. A. (2024). Innovation of Instructional Design and Assessment in the Age of Generative Artificial Intelligence. TechTrends, 68(1), 195–199. https://doi.org/10.1007/s11528-023-00926-x

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why Minimal Guidance During Instruction Does Not Work: An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and Inquiry-Based Teaching. Educational Psychologist, 41(2), 75–86. https://doi.org/10.1207/s15326985ep4102_1

Konak, A., & Clarke, C. (2023). Augmenting Critical Thinking Skills in Programming Education through Leveraging Chat GPT: Analysis of its Opportunities and Consequences. 2023 Fall Mid Atlantic Conference: Meeting Our Students Where They Are and Getting Them Where They Need to Be Proceedings, 45117. https://doi.org/10.18260/1-2--45117

Krushinskaia, K., Elen, J., & Raes, A. (2024). Effects of generative artificial intelligence on instructional design outcomes and the mediating role of pre-service teachers' prior knowledge of different types of instructional design tasks. In International Conference on Artificial Intelligence in Education (pp. 395-400). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-64312-5_49

Lau, S., & Guo, P. (2023). From" Ban it till we understand it" to" Resistance is futile": How university programming instructors plan to adapt as more students use AI code generation and explanation tools such as ChatGPT and GitHub Copilot. In Proceedings of the 2023 ACM Conference on International Computing Education Research-Volume 1 (pp. 106-121). https://doi.org/10.1145/3568813.3600138

Mok, M. M. C., & Moore, P. J. (2019). Teachers & self-efficacy. Educational Psychology, 39(1), 1–3. https://doi.org/10.1080/01443410.2019.1567070

Ruiz-Rojas, L. I., Acosta-Vargas, P., De-Moreta-Llovet, J., & Gonzalez-Rodriguez, M. (2023). Empowering Education with Generative Artificial Intelligence Tools: Approach with an Instructional Design Matrix. Sustainability, 15(15), 11524. https://doi.org/10.3390/su151511524

Sailer, M., Bauer, E., Hofmann, R., Kiesewetter, J., Glas, J., Gurevych, I., & Fischer, F. (2023). Adaptive feedback from artificial neural networks facilitates pre-service teachers' diagnostic reasoning in simulation-based learning. Learning and Instruction, 83, 101620. https://doi.org/10.1016/j.learninstruc.2022.101620

Van Den Berg, G., & Du Plessis, E. (2023). ChatGPT and Generative AI: Possibilities for Its Contribution to Lesson Planning, Critical Thinking and Openness in Teacher Education. Education Sciences, 13(10), 998. https://doi.org/10.3390/educsci13100998

Yao, N., & Wang, Q. (2024). Factors influencing pre-service special education teachers' intention toward AI in education: Digital literacy, teacher self-efficacy, perceived ease of use, and perceived usefulness. Heliyon, 10(14), e34894. https://doi.org/10.1016/j.heliyon.2024.e34894

Zee, M., & Koomen, H. M. (2016). Teacher self-efficacy and its effects on classroom processes, student academic adjustment, and teacher well-being: A synthesis of 40 years of research. Review of Educational research, 86(4), 981-1015

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

An Exploration of the Impact of Using ChatGPT on Students' On-Task Performance and Task Success

Arnel OCAY [ab*] & Maria Mercedes RODRIGO [a]
[a]*Ateneo de Manila University, Philippines*
[b]*Urdaneta City University*
*arnelocay@ucu.edu.ph

Abstract: Learning programming can be difficult, especially for beginners. This study examines the potential of ChatGPT to influence students' programming performance by analyzing self-reported on-task behaviors. Sixty-two first-year IT and CS students were randomly assigned to a control group and an experimental group. Participants of the study completed a hands-on GUI-based programming task using Java Swing. Findings showed significant differences in help-seeking behaviors between groups, with ChatGPT users seeking less help from instructors. While self-reported task success was significantly higher among ChatGPT users, on-task performance metrics like error count and code execution attempts were not associated with task success. These findings suggest that while ChatGPT may support performance on programming tasks, its pedagogical implications require careful consideration to scaffold learning effectively and avoid over-reliance on the tool.

Keywords: ChatGPT, On-Task Performance metrics, Experimental study, Programming performance

1. Introduction

Many students in introductory programming courses, particularly those without prior coding experience, face significant challenges in learning and understanding programming concepts (Javier, 2021; Muntanga et al., 2023). These difficulties extend beyond writing code to include reading, interpreting, and debugging programs (Msane et al., 2020; Thuné & Eckerdal, 2018; Robins, Rountree, & Rountree, 2003). For example, McCracken et al. (2001) found that most first-year students were unable to produce syntactically and logically correct code, while Lister et al. (2004) observed that novice programmers struggled with code comprehension. These recurring issues have been widely recognized (Lahtinen, Ala-Mutka, & Järvinen, 2005; Kaila et al., 2010; Höök & Eckerdal, 2015), often leading students to develop negative perceptions of programming courses (Lai et al., 2022).

To address these challenges, various tools have emerged to support students in acquiring programming skills. One of these is the use of large language models (LLMs), such as ChatGPT, which have shown promise in automating code generation and offering real-time support to novice programmers (White et al., 2023; Takerngsaksiri et al., 2023). ChatGPT, a neural network-based generative AI model, can generate complex responses and has been widely applied in domains like programming and software development (Petrovska et al., 2024; Kadir et al., 2023). Several studies have documented the usefulness of ChatGPT and similar tools in improving productivity and workflow among professional developers (Sauvola et al., 2024; Deniz et al., 2023; Benaich & Hogarth, 2021). These tools are now being adopted in educational settings, particularly in computer science education (Silva et al., 2024), where they have been utilized for code generation (Denny et al., 2024), error correction (Sobania et al., 2023), and syntax assistance.

Despite growing integration of these, there is still limited empirical evidence on how tools like ChatGPT influence novice programmers' problem-solving behaviors, particularly during hands-on programming activities. While several existing studies have focused on

student perceptions (Yilmaz & Yilmaz, 2023; Sun et al., 2024), performance comparisons (Xue et al., 2024; Kozar et al., 2024), or theoretical overviews (Kasneci et al., 2023; Lo, 2023), few have investigated the nature of students' on-task behaviors while using AI assistance in programming. In particular, the relationship between tool usage, task success, and behavioral indicators such as help-seeking, debugging efforts, and persistence remains underexplored.

This study contributes to the discourse on generative AI in computing education by analyzing self-reported on-task performance metrics during a programming task with or without ChatGPT assistance. Focusing on performance behaviors, it examines engagement and interaction strategies through indicators such as completion time, code executions, and error frequencies (Robins et al., 2003). While not direct measures of cognitive gains, these metrics offer valuable proxies for understanding how students navigate programming challenges in real time. This study is part of a threefold research initiative exploring the role of Gen AI in computing education. An overview of these interrelated studies was published in LCNS Augmented Cognition HCII 2025 (Ocay and Rodrigo, 2025).

Specifically, the study aims to answer the following research questions:

RQ1: What is the difference in self-reported on-task performance between students who used ChatGPT and those who did not?

RQ2: What is the relationship between the self-reported on-task performance metrics and the overall self-reported task success of the students?

RQ3: Is there a significant difference in overall self-reported task success between students who used ChatGPT and those who did not?

2. Methods and Procedures

2.1 Participants and Experimental Design

Sixty-four first-year students from two Philippine universities participated in the study. After excluding two students due to incomplete data, 62 participants remained (Male=46, 74.2%; Female=16, 25.8%). All were enrolled in an introductory Java programming course and had no exposure to GUI programming using Swing components.

Participants were randomly assigned to a control group (n=31), which used PDFs, slides, and videos, or an experimental group (n=31), which used ChatGPT. Both groups received a 22-minute video tutorial on Java Swing, then completed a 40-minute programming task, that is to build a basic calculator app with GUI components. The experimental group used the same native ChatGPT version throughout the activity for assistance.

2.2 Measure: On-Task Performance Metrics

To evaluate ChatGPT's impact on students' engagement and outcomes in programming, we adopted on-task performance metrics grounded in prior studies on novice programmers (Robins, Rountree, & Rountree, 2003). Participants self-reported these metrics using a web-based interface that logged task-specific actions during the hands-on session (see Figure 1).

2.2.1 Error Count (ErrorCount)

Error count is an indicator of problem-solving strategy in programming, where a high frequency suggests ineffective or underdeveloped approaches (Lahtinen et al., 2005; Robins et al., 2003; Winslow, 1996). This refers to the total number of errors a student records during the coding task, incremented with each occurrence using the logging system. Only positive values are accepted.

2.2.2 Number of Attempts to Execute the Code (NumAttemptsExecCode)

A common strategy in problem-solving is trial and error (Iyagba, 2020). Often used by novice programmers through repeated code execution attempts (Lister et al., 2004), which reflects iterative refinement leading to solutions (Lahtinen et al., 2005; Jadud, 2006). This metric tracks the frequency a student executes code during the task, with each attempt logged as a single increment.

2.2.3 Number of Times Sought Help from Teachers (NumTimesHelp)

Help-seeking is widely recognized as a sign of engagement and effective problem-solving (Nelson-Le Gall, 1985; Newman & Schwager, 1995). This metric tracks how often a student requested help during the task, with each instance logged as a single increment.

2.2.4 Task Success (TaskSuccess)

A binary metric that captures whether the student completed or solved the programming problem. Students report this by indicating “Yes”, which means the task was completed, or “No”, which means the task was not completed at the end of the programming task.

2.3 Data Collection Procedure

To ensure ethical compliance, the study obtained clearance from the Ateneo Research Ethics Office. Participation was strictly voluntary.

Data on students’ on-task performance was collected using a custom-built web application called the Study User Interface (SUI, see Figure 1), developed with JavaScript, HTML, and CSS, and using Firebase as the backend. This platform allowed students to record performance logs during and after the hands-on task. This includes the on-task performance metrics mentioned in section 2.3.



Figure 1. User Interface for Students during Hands-on Task.

During the 40-minute task, participants interacted with the Student User Interface (SUI) to record task-related actions in real time. Selecting START initiated the session timer, while each error and debugging event was incrementally logged at the moment of occurrence. Upon selecting FINISH, participants provided self-assessment data, including the number of help requests, perceived code quality, and task success. Built-in input validation ensured data consistency and accuracy, thereby minimizing memory bias.

2.4 Statistical Analysis

An independent non-parametric test was used to compare on-task performance metrics between groups. Descriptive statistics, including mean and frequency analysis, were employed to identify group differences. A point-biserial correlation analysis was conducted to assess relationships between task success and each performance metric. Fisher’s exact test was used to examine whether task success rates significantly differed between the experimental and control groups.

3. Discussion of Findings

3.1 (RQ1) Comparison of Self-Reported On-Task Student Performance

A Mann-Whitney U test was conducted to compare self-reported on-task performance between groups. Among the three indicators, only the number of help-seeking instances differed significantly (U=606.50, p=.024), with the control group (M=1.29, SD=2.75) requesting help more frequently than the experimental group (M=0.16, SD=0.37). This suggests that students without AI support relied more on external instructional help, aligning with prior work highlighting novice learners' reliance on external scaffolds (Lister et al., 2004; Vygotsky, 1978).

No significant differences were observed for ErrorCount or NumAttemptsExecCode. However, mean comparisons showed that students using ChatGPT had slightly more executions (M=5.65 vs. M=4.71) and more errors (M=3.81 vs. M=3.16), possibly reflecting more trial-and-error attempts encouraged by immediate AI feedback. While this may indicate iterative engagement, it does not confirm improved understanding. Prior research suggests novice programmers often iterate without grasping underlying logic (Lahtinen et al., 2005; Lister et al., 2004). These findings suggest that ChatGPT may support sustained engagement and reduced instructor dependence but do not confirm cognitive gains or deeper learning.

3.2 (RQ2) Relationship between Self-Reported On-Task Metrics and Task Success

A point-biserial correlation showed that self-reported task success was not significantly related to ErrorCount (rpb=.099, p=.445), NumAttemptsExecCode (rpb=.004, p=.976), or NumTimesHelp (rpb=.056, p=.664). This indicates that the amount of effort or frequency of activity did not directly translate to task success.

However, correlations were found among the behavioral metrics themselves. Students who executed code more often tended to encounter more errors (rpb=.465, p<.001), and those who executed more also sought help more often (rpb=.265, p=.037). These patterns reflect typical novice strategies that use trial-and-error and help-seeking (Lahtinen et al., 2005; Puustinen & Rouet, 2009; Vygotsky, 1978; Lister et al., 2004) in navigating problems and tasks.

These results suggest that while students may demonstrate high activity levels, these do not reliably indicate success. Educational designers may need to design interventions that promote more reflective, strategic problem-solving behaviors rather than relying solely on behavioral persistence, such as trial-and-error strategy.

3.3 (RQ3) Differences in Self-Reported Task Success Rates

A significant difference in task success was observed between groups. Fisher's exact test (p<.001) and chi-square analysis ($\chi^2(1)=18.645$, p<.001) showed that students with ChatGPT reported higher success rates than those without. While encouraging, these are self-reported success outcomes and should not be interpreted as evidence of improved programming competence.

These findings are consistent with prior work highlighting the potential of generative AI tools to support novice learners (Kasneci et al., 2023). However, the increased success may reflect perceived progress or reliance on AI-generated solutions rather than a deeper understanding of the programming concepts.

4. Conclusions

With the growing interest in AI's impact on programming education, this study examined ChatGPT's influence on students' performance behaviors through self-reports. Findings show that ChatGPT users sought less human assistance and executed their code more often, suggesting that AI use may encourage more independent, iterative problem-solving. However, behaviors such as error count, code execution, and help-seeking showed no significant link to

reported task success. This indicates that high engagement does not necessarily translate to deeper conceptual learning.

Students in the experimental group reported higher success rates, which may reflect greater confidence in task completion. While this aligns with industry observations of AI tools improving productivity, it raises important pedagogical questions about overreliance and surface-level learning. As generative AI tools become more accessible in educational contexts, thoughtful integration into programming pedagogy is essential.

5. Limitations and Future Work

This study used self-reported data, which remain prone to bias despite real-time entry. Future work should triangulate these with ChatGPT logs, including chat transcript analysis to analyze student behaviors further, and use pretest and posttest analyses to determine learning gains.

Acknowledgements

We thank Ateneo de Manila University, Ateneo Laboratory for the Learning Sciences, and the Ateneo University Research Office for their support. We also extend our gratitude to all participants and their schools, Pangasinan State University-Urdaneta Campus, and Saint Louis University-Baguio City. Also, our gratitude to Dr. Jennifer Ramos-Miguel for her guidance in the statistical design of this work.

References

Benaich, I., Hogarth, N. (2021) State of AI Report. Accessed at https://www.stateof.ai/2021-report-Launch

Deniz, B. K., Gnanasambandam, C., Harrysson, M., Hussin, A., & Srivastava, S. (2023). Unleashing developer productivity with generative AI. McKinsey Digital.

Denny, P., Prather, J., Becker, B., Finnie-Ansley, J., Hellas, A., Leinonen, J., Luxton-Reilly, A., Reeves, B., Santos, E. A., Sarsa, S. (2024) Computing Education in the Era of Generative AI. Commun. ACM 67, 2, 56–67. https://doi.org/10.1145/3624720

Höök, J., & Eckerdal, A. (2015) On the bimodality in an introductory programming course. An analysis of student performance factors. *In Proceedings of the 2015 Learning and Teaching in Computing and Engineering Conference, Taipei, Taiwan pp. 79-86.* https://doi.org/doi:10.1109/LaTiCE.2015.25

Iyagba, P. W., (2020). Learning and Problem Solving: The Use of Problem-Solving Method To Achieve learning in Pupils. African Social and Educational Journal.

Jadud, M. C. (2006). Methods and tools for exploring novice compilation behavior. *ACM SIGCSE Bulletin, 38(3), 1–5.* https://doi.org/10.1145/1140123.1140124

Javier, B. (2021). Understanding their Voices from Within: Difficulties and Code Comprehension of Life-Long Novice Programmers. *1. 53-76*.

Kadir, E., Rahman, T., & Barman, S. (2023). Exploring the Competency of ChatGPT in Solving Competitive Programming Challenges. Int. J. Adv. Trends Comput. Sci. Eng., vol. 13, no. 1, pp. 13–23, 2023, https://doi.org/doi:10.30534/ijatcse/2024/031312024.

Kaila, E., Rajala, T., Laakso, M., & Salakoski, T. (2010). Effects of course-long use of a program visualization tool. Australasian Computing Education Conference, 97–106. https://doi.org/10.5555/1862219.1862234

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., …Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, Article 102274.

Kozar, T., Ostojic, D., Liu, Y., Mernik, M. (2024) Computer Science Education in ChatGPT Era: Experiences from Experiment in a Programming Course for Novice Programmers. *Mathematics. 12, 629*.

Lahtinen, E., Ala-Mutka, K., & Järvinen, H. M. (2005). A study of the difficulties of novice programmers. *ACM SIGCSE Bulletin*, 37(3), 14–18.

Lai, C., Chen, Y., Wang, Y., & Liao, H. (2022). The Study of Learning Computer Programming for Students with Medical Fields of Specification—An Analysis via Structural Equation Modeling. International Journal of Environmental Research and Public Health, 19(10), 6005. https://doi.org/10.3390/ijerph19106005

Lister, R., Adams, E., Fitzgerald, S., Fone, W., Hamer, J., Lindholm, et al. (2004). A multi-national study of reading and tracing skills in novice programmers. *Inroads, 36(4),* 119–150.
Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. Education Sciences, 13(4), 410.
McCracken, M., Almstrum, V., Diaz, D., Guzdial, M., Hagan, D., Kolikant, Y.B., Laxer, C., Thomas, L., Utting, I., & Wilusz, T. (2001). A multi-national, multi-institutional study of assessment of programming skills of first-year CS students. *Inroads, 33(4)*, 125–140. https://doi.org/10.1145/572133.572137
Msane, J., Mutanga, B., & Chani, T. (2020). Students' perception of the effect of cognitive factors in determining success in computer programming: a case study. *International Journal of Advanced Computer Science and Applications, 11(7).* https://doi.org/10.14569/ijacsa.2020.0110724M.
Mutanga, M. B., Piyose, P. X., & Ndovela, S. (2023). Factors Affecting Career Preferences and Pathways: Insights from IT Students. *Journal of Information Systems and Informatics, 5(3), 1111–1122.* https://doi.org/10.51519/journalisi.v5i3.556
Nelson-Le Gall, S. (1985). Help-seeking behavior in learning. Review of Research in Education, 12, 55-90.
Newman, R.S., & Schwager, M.T. (1995). Students' help seeking during problem solving: Effects of grade, goal, and prior achievement. American Educational Research Journal, 32, 352-376.
Ocay, A., & Rodrigo, M.M. (2025). A Field Study on the Use of Gen AI to Support Computing Education. In: Schmorrow, D.D., Fidopiastis, C.M. (eds) Augmented Cognition. HCII 2025. Lecture Notes in Computer Science, vol 15778. Springer, Cham. https://doi.org/10.1007/978-3-031-93724-8_17
Petrovska, O., Clift, L., Moller, F., & Pearsall, R. (2024). Incorporating Generative AI into Software Development Education. *ACM. United Kingdom.*
Puustinen, M., & Rouet, J. F. (2009). Learning with new technologies: Help seeking and information searching revisited. Computers & Education, 53(4), 1014–1019. https://doi.org/10.1016/j.compedu.2008.07.002
Robins, A., Rountree, J., & Rountree, N. (2003). Learning and Teaching Programming: A Review and Discussion. *Computer Science Education, 13(2), 137–172.* https://doi.org/10.1076/csed.13.2.137.14200
Sauvola, J., Tarkoma, S., Klemettinen, M., Riekki, J., Doermann, D. (2024). Future of software development with generative AI. *Automated Software Engineering 31:25.* https://doi.org/10.1007/s10515-024-00426-z
Silva, C.A.G.d., Ramos, F.N., de Moraes, R.V., & Santos, E.L.d. (2024). ChatGPT: Challenges and Benefits in Software Programming for Higher Education. *Sustainability 2024, 16,1245.* https://doi.org/10.3390/su16031245
Sobania, D., Briesch, M., Hanna, C., & Petke, J. (2023). An Analysis of the Automatic Bug Fixing Performance of ChatGPT. arXiv:2301.08653.
Sun, D., Boudouaia, A., Zhu, C., Li, Y. (2024) Would ChatGPT-facilitated programming mode impact college students' programming behaviors, performances, and perceptions? An empirical study. *Int J Educ Technol High Educ. 21:14.*
Takerngsaksiri, W., Warusavitarne, C., Yaacoub, C., Hou, M. H. K., & Tantithamthavorn, C. (2023, October 31). Students' perspective on AI code completion: benefits and challenges. arXiv.org. https://arxiv.org/abs/2311.00177
Thuné, M., & Eckerdal, A. (2018). Analysis of Students' learning of computer programming in a computer laboratory context. European Journal of Engineering Education, 44(5), 769–786. https://doi.org/10.1080/03043797.2018.1544609
Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Harvard University Press.
White, J., Hays, S., Fu, Q., Spencer-Smith, J., & Schmidt, D. C. (2023). ChatGPT prompt patterns for improving code quality, refactoring, requirements elicitation, and software design. arXiv.org. https://arxiv.org/abs/2303.07839
Winslow, L. E. (1996). Programming pedagogy—a psychological overview. *SIGCSE Bulletin, 28(3), 17–22.* Retrieved at https://dl.acm.org/doi/pdf/10.1145/234867.234872
Xue, Y., Chen, H., Bai, G., Tairas, R., Huang, Y. (2024). Does ChatGPT Help with Introductory Programming? An Experiment of Students Using ChatGPT in CS1. *Int Conf on Software Engineering. IEEE/ACM.*
Yilmaz, R., Yilmaz, F. G. (2023) The effect of generative artificial intelligence-based tool use on students' computational thinking skills, programming self-efficacy and motivation. *Computers and Education: Artificial Intelligence 4*

RoboTuB: Retrieval-Augmented Tutoring for Adaptive Learning in STEM MOOCs

Avijit PANDEY[a], Gaurav MISHRA[b], Sunny Prakash PRAJAPATI[a] & Syaamantak DAS[a*]
[a]*Centre for Educational Technology, Indian Institute of Technology Bombay, India*
[b]*Department of Mechanical Engineering, Indian Institute of Technology Bombay, India*
*syaamantak.das@iitb.ac.in

Abstract: Educational AI chatbots have shown promise in enhancing learning, yet they often lack adaptive reasoning and pedagogical scaffolding for domain-specific subjects. This paper presents RoboTuB, a Retrieval-Augmented Generation (RAG) powered tutoring chatbot, designed to support team-based learning in STEM education. Unlike traditional question-answering (QA) systems, RoboTuB retrieves structured knowledge from domain-specific corpora, dynamically generating explanations, simulations, and code examples tailored to learners' needs. We evaluate RoboTuB in a two-phase study: (1) a usability assessment with novice learners using System Usability Scale (SUS) and NASA-TLX cognitive load metrics, and (2) a comparative analysis in a MOOC setting. Results from a two-phase evaluation showed that students using RoboTuB demonstrated substantial learning gains and reduced cognitive load compared to Q&A forums. These findings suggest that retrieval-augmented tutoring can provide meaningful support for multi-step problem-solving tasks. We discuss implications for NLP-driven educational chatbots and propose future directions for multimodal AI tutors.

Keywords: question answering, query processing, chatbot, Retrieval Augmented Generation (RAG)

1. Introduction

Large Language Models (LLMs) and retrieval-based architectures have led to significant advances in question-answering (QA) and dialogue systems. However, when applied to educational contexts, especially in STEM domains, they often fall short of delivering sustained, adaptive, and pedagogically grounded learning experiences. Learners require more than factual answers; they need scaffolded support that adapts to their evolving understanding, handles conceptual dependencies, and provides multimodal guidance (e.g., code, simulations, and theoretical context). In this work, we focus on multi-step learning tasks, tasks that require learners to sequentially navigate through multiple interdependent knowledge units. For instance, in robotics education, a student must first understand a physical system's dynamics, simulate it with appropriate models, and then implement a real-time control algorithm. Current QA systems often treat these subtasks in isolation, leading to fragmented or incoherent guidance.

To address this gap, we introduce RoboTuB (Robotics Tutoring Bot), a Retrieval-Augmented Generation (RAG) chatbot tailored to support learners in a MOOC setting focused on robotics and control systems. RoboTuB is designed to function as an AI-powered tutor, offering structured, context-aware explanations that evolve with the learner's progress. It integrates three pedagogically relevant modalities: theory (conceptual foundations), simulation (interactive experimentation), and code (implementation-level examples). For this study, we explore the following research questions:

RQ1: Can a retrieval-augmented LLM-based tutor improve learners' understanding and performance on multi-step tasks in a robotics MOOC?

RQ2: Does using such a system reduce cognitive load compared to traditional QA forums?

RQ3: How do learners perceive the usability and dialogue quality of an AI tutor designed with pedagogical scaffolding in mind?

For these research questions, we evaluated the effectiveness of the RoboTub using a two-phase study design with learners in lab-based and MOOC-based settings.

2. Related Work

Research on intelligent tutoring systems (ITS) and AI-driven educational technologies has grown substantially in recent years. While early ITS models focused on rule-based dialogue or domain-specific pipelines (VanLehn, 2011), recent advances in large language models (LLMs) and retrieval-augmented architectures (Lewis et al., 2020) have made scalable, adaptable AI tutors more feasible. However, translating these architectures into pedagogically effective systems remains a challenge. Prior work has highlighted the limitations of large language models in educational settings, including hallucination, a lack of alignment with curriculum goals, and difficulty sustaining scaffolding across multi-step reasoning (Kasneci et al., 2023). At the same time, emerging studies have begun exploring retrieval-augmented LLMs for dialogue tutoring and knowledge grounding in education (Henkel et al., 2024). Our work builds on this line of research by introducing a layered retrieval architecture (theory, simulation, code) and evaluating its educational impact in authentic learning environments.

2.1 Educational NLP and Scaffolding

Scaffolding refers to instructional strategies that provide learners with just-in-time support and fade assistance as learners gain competence (Wood, Bruner & Ross, 1976). In the context of AI tutors, this implies tailoring explanations, offering intermediate steps, and adjusting complexity dynamically. Prior works like (Winkler & Söllner, 2018; Stahl et al., 2024) emphasizes the importance of aligning chatbot responses with scaffolding principles to enhance conceptual understanding and promote cognitive engagement. Despite these advances, few systems offer retrieval-based scaffolding across multiple modalities such as theory, simulation, and code.

2.2 Retrieval-Augmented Generation for Educational Tasks

Retrieval-Augmented Generation (RAG) has shown promise in QA and summarization tasks by enhancing factuality and grounding generative outputs in external corpora (Lewis et al., 2020; Karpukhin et al., 2020). In education, RAG can reduce hallucinations and provide students with more domain-consistent explanations (Béchard & Ayala, 2024). However, most prior implementations focus on single-turn responses or factoid-style QA, which limit their pedagogical utility in supporting multi-step reasoning or personalized tutoring.

2.3 Dialogue Quality and Evaluation in Educational Chatbots

Evaluating AI tutors requires more than standard NLP metrics (e.g., BLEU or ROUGE). Usability frameworks such as SUS (Brooke, 1996) and workload metrics like NASA-TLX (Hart & Staveland, 1988) are essential to assess learners' cognitive and affective experiences. Additionally, the PARADISE framework (Walker et al., 1997) enables analysis of dialogue success based on task completion, efficiency, and user satisfaction. In their work, (Istrate, & Velea, 2024) argue that aligning chatbot evaluation with educational learning outcomes is crucial for real-world applicability, especially in MOOCs and self-paced learning environments.

2.4 Generative AI in STEM Education

The use of generative models in STEM instruction is on the rise, yet most systems either rely on static prompts or offer minimal interactivity. Khanmigo (Khan Academy, 2023) and Google's Socratic App provide lightweight tutoring experiences, but lack deep integration with structured

pedagogical frameworks. Our work builds upon this gap by using a structured retrieval pipeline and LLM prompting strategy to simulate domain-specific scaffolding. RoboTuB distinguishes itself through its focus on multi-modal support and dialogic adaptability tailored to robotics education.

3. Methodology

3.1 System Architecture and Retrieval Pipeline

RoboTuB follows a modular RAG architecture. The core system consists of:

- Corpus Segmentation: The knowledge base is divided into three thematic layers: (1) theory (concepts, equations), (2) simulation (code snippets with explanations), and (3) programming (Python/C++ implementations). Each document is chunked into 100–250 token passages and embedded using MiniLM-L6 sentence-transformers.
- Indexing and Retrieval: Chunks are indexed using FAISS (Facebook AI Similarity Search) (Johnson and Jégou, 2019) for approximate nearest neighbor retrieval. For each query, top-5 passages from each layer are retrieved and passed through a reranker using cosine similarity with the dialogue context embedding.
- Context Construction: Retrieved texts are concatenated using a delimiter scheme and prepended to the system prompt. Summarized multi-turn history (using GPT-3.5) ensures conversational coherence.

3.2 Prompt Engineering and Generation

We use GPT-3.5 in a zero-shot setting with domain-adaptive prompts that dynamically insert context-relevant scaffolding instructions. Prompt templates were iteratively refined with the following goals:

- Encourage stepwise explanations (“Explain like a tutor to a novice”).
- Format responses using sections: Conceptual Overview, Simulation Walkthrough, Code Illustration.
- React to prior dialogue context (“Based on what we discussed before…”).

Example prompt prefix: “*You are a robotics tutor assisting a student working on PID control. Given the following materials and prior questions, provide a clear explanation including theory, a simulation example, and runnable code.*”

3.3 Evaluation Design and Controls

- **Phase 1** involved 19 graduate students using RoboTuB in a lab setting. SUS and NASA-TLX were administered after 30-minute tutoring sessions.
- **Phase 2** used data from two editions of a MOOC on control systems. Edition 1 (n = 92) used a traditional Q&A forum; Edition 2 (n = 91) used RoboTuB.
- Both editions followed an identical curriculum, grading, and teaching support. Prior academic background, declared programming experience, and control exposure were collected to ensure demographic balance.
- Data from platform logs, post-test scores, and chatbot interactions were anonymized and analyzed using Welch’s t-test.

It is important to note certain methodological limitations. First, while RoboTuB leverages a retrieval-augmented pipeline, our evaluation did not include an LLM-only baseline or a systematic ablation of retrieval layers and prompt templates. We chose to focus on the contrast between RoboTuB and the discussion forum baseline, as this reflects learners’ existing support environment in the MOOC. Future work will isolate the contribution of individual system components through controlled ablation studies. Second, while the system is adaptable to different learner queries, personalization in the present study was limited to

prompt-template scaffolds (e.g., encouraging step-by-step reasoning). More advanced personalization techniques, such as learner modeling, adaptive difficulty, or long-term tracking, were outside the scope of this study, but remain promising directions.

4. Results & Analysis

4.1 RQ1: Performance on Multi-Step Tasks

To address RQ1, we analyzed learners' task accuracy and post-test performance across two MOOC cohorts. Edition 1 (n = 92) used a traditional discussion forum, while Edition 2 (n = 91) interacted with RoboTuB. Post-test scores revealed a 15.8% average increase in problem-solving accuracy for the RoboTuB group compared to the forum group. While this indicates a substantial improvement, we did not conduct fine-grained statistical subgroup analysis, and therefore interpret the result as an observed trend rather than a conclusive effect.

4.2 RQ2: Cognitive Load Comparison

To evaluate RQ2, we administered the NASA Task Load Index (NASA-TLX) after 30-minute tutoring sessions with 19 graduate students during Phase 1. Results indicated reductions in both mental demand (Mean = 34.8) and frustration (Mean = 31.5) compared to baseline scores from Q&A-based learning. Learners reported that RoboTuB's structured explanations and multimodal output helped reduce cognitive burden, particularly when transitioning between theory and code. Learners with lower perceived workload generally performed better in follow-up tasks, though this relationship was not statistically tested.

4.3 RQ3: Usability and Dialogue Quality

For RQ3, we assessed learner perceptions using the System Usability Scale (SUS) and qualitative feedback. RoboTuB received a mean SUS score of 76.58, indicating strong usability. Participants appreciated the multi-step explanations and adaptability of responses. Manual inspection of conversations suggested that RoboTuB responses were more often complete and relevant compared to peer replies in the discussion forum. Learner feedback during SUS administration and open-ended reflections suggested satisfaction with the chatbot's clarity and instructional helpfulness. Based on an inspection of interaction logs, most learner problems appeared to be resolved within a few conversational turns (typically 2–4), although turn counts were not systematically recorded.

5. Discussion & Future Work

The findings from our evaluation suggest that RoboTuB can meaningfully enhance the learning experience for students engaged in complex STEM topics. By combining retrieval-based grounding with pedagogically aligned generation, the system improved usability, cognitive load, and learning outcomes. However, these contributions must be interpreted with consideration of several design constraints.

5.1 Pedagogical Implications

RoboTuB's effectiveness stems from its ability to simulate instructional scaffolding, drawing on retrieval to tailor explanations and context across multiple modalities. The observed improvements in performance and satisfaction, especially among novice learners, highlight the value of adaptive support in MOOCs. These results support recent literature advocating for dialogic, context-aware AI tutors that move beyond factoid-level QA. These observed improvements should be interpreted with caution, as subgroup analyses were not conducted.

5.2 Limitations

The system was evaluated within a specific MOOC focused on robotics and control systems, and the transferability to other subjects remains unverified. The study took place during the availability of GPT 3.5 in late 2023 to early 2024, without the opportunity to utilize advanced LLM models like GPT 4/Claude. Although no fine-tuning was applied, prompt engineering proved to be complex and may necessitate task-specific efforts to achieve generalization. Additionally, some responses tended to be overly verbose or misaligned with learners' intent, highlighting a need for real-time adaptation of responses.

5.3 Future Research Directions

Future work will explore integrating RoboTuB with adaptive learner models to enable more fine-grained personalization, as well as conducting controlled ablation studies to evaluate retrieval and prompting strategies in isolation. Richer qualitative analysis of learner interactions, and replication across multiple MOOCs, will further validate the system's generalizability. In conclusion, this study demonstrates that a retrieval-augmented tutoring system can enhance learners' performance, reduce perceived cognitive load, and provide a usable and pedagogically meaningful alternative to traditional Q&A forums in STEM MOOCs.

6. Conclusion

This paper introduced RoboTuB, a retrieval-augmented tutoring system that supports multi-step learning in a robotics MOOC. Through two-phase evaluation, we showed that RoboTuB improves problem-solving accuracy, reduces perceived workload, and enhances learner satisfaction. Our contributions include a modular RAG pipeline for multimodal tutoring, a validated prompt engineering framework, and evidence-based evaluation. While limitations exist around generalizability and verbosity, the system lays groundwork for future AI tutors that are pedagogically aware, contextually grounded, and student-adaptive. By refining how LLMs engage with domain knowledge, RoboTuB demonstrates the potential of NLP to deliver personalized, structured educational support in open-access environments like MOOCs.

References

Brooke, J. (1996). SUS: A quick and dirty usability scale. In P. W. Jordan, B. Thomas, & B. A. Weerdmeester (Eds.), Usability Evaluation in Industry (pp. 189–194). CRC Press.

Béchard, P., & Ayala, O. M. (2024). Reducing hallucination in structured outputs via Retrieval-Augmented Generation. arXiv preprint arXiv:2404.08189..

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. Advances in Psychology, 52, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Henkel, O., Levonian, Z., Li, C., & Postle, M. (2024). Retrieval-augmented generation to improve math question-answering: Trade-offs between groundedness and human preference. In Proceedings of the 17th International Conference on Educational Data Mining (pp. 315-320).

Istrate, O., & Velea, S. (2024). Guidelines for online and blended learning: Design, delivery, assessment, evaluation of study programmes. Premises of academic curriculum digitalisation. September, 1-113.

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3), 535-547.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and individual differences, 103, 102274.

Karpukhin, V., Oguz, B., Min, S., Lewis, P. S., Wu, L., Edunov, S., ... & Yih, W. T. (2020, November). Dense Passage Retrieval for Open-Domain Question Answering. In EMNLP (1) (pp. 6769-6781).

Khan Academy. (2023). Khanmigo: AI-powered tutoring for learners and educators. Retrieved from https://www.khanacademy.org/khanmigo

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Küttler, H., ... & Riedel, S. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems (NeurIPS), 33, 9459–9474.

Stahl, M., Biermann, L., Nehring, A., & Wachsmuth, H. (2024). Exploring LLM prompting strategies for joint essay scoring and feedback generation. arXiv preprint arXiv:2404.15845.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4), 197–221. https://doi.org/10.1080/00461520.2011.611369

Walker, M. A., Litman, D. J., Kamm, C., & Abella, A. (1997). PARADISE: A framework for evaluating spoken dialogue agents. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL), 271–280. https://doi.org/10.3115/976909.979652

Winkler, R., & Söllner, M. (2018, July). Unleashing the potential of chatbots in education: A state-of-the-art analysis. In Academy of management proceedings (Vol. 2018, No. 1, p. 15903). Briarcliff Manor, NY 10510: Academy of Management.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem-solving. Journal of Child Psychology and Psychiatry, 17(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Idea Space Change Process Visualization System for Supporting Interpretation of Thinking Strategy in Research Instruction

Kota KUNORI[a*] & Tomoko KOJIRI[b]
[a]*Graduate School of Science and Engineering, Kansai University, Japan*
[b]*Faculty of Engineering Science, Kansai University, Japan*
*k194388@kansai-u.ac.jp

Abstract: In research activities, it is difficult for instructors to provide guidance to learners' thinking strategies. Since thinking depending on the thinking strategy, it is possible to infer the strategy from the thinking process for ideas (idea space change process). Therefore, in this study, we propose a system that visualizes an idea space change process from the thinkings obtained by the learners. As the result of evaluation experiment, it was confirmed that the number of instructional contents for thinking strategies increased by using system.

Keywords: Research instruction support, idea space change process, visualization

1. Introduction

In research activities, instructors often guide learners through the research outputs they produce. Research outputs do not describe the strategies that generated the thinking (thinking strategies). A thinking strategy refers to a policy of thinking to proceed with research, such as “When considering a research problem, consider various candidates.” It is desirable for instructors to provide guidance for learners’ thinking strategies in order to enable them to conduct research on their own.

Research consists of elements such as background, objectives, solutions, and experiments. The research output is a specific topic that corresponds to each of these elements. Since the thinking strategy indicates how to proceed with thinking in order to obtain the research output, to grasp the sequence of learners’ thinking from a global viewpoint will help to understand their thinking strategy. In this study, we aim to develop a system that can help instructors become aware of learners’ thinking strategies by visualizing the sequence of learners’ thought.

Some systems provide information which enable learners or teachers to interpret learners’ thinking strategies by estimating them. (Yoshioka, Hayashi, & Seta, 2019; Tian, et al., 2019). These systems only summarize a few consecutive thinkings, and it is difficult to estimate the strategy that contributes to the learners over all thinking. A method for visualizing process of creating outputs has been proposed to provide a global viewpoint of the entire thinking process.

A system that automatically generates Linkograph, a method that enumerates the components of outputs and expresses which components are integrated with which components in a hierarchical structure, has been proposed (Smith, et al., 2025). This representation method cannot show the order of integration. Since the thinking strategy is expressed in the process of creating outputs, it can be estimated by visualizing the process of creating outputs.

In this study, we propose a method of visualizing how the semantic clusters of topics (idea), which are candidates for outputs, have changed in a graph structure. We also aim to support instructors’ estimation of learners’ thinking strategies by constructing a system with a

function to visualize the process of creating research outputs according to the proposed method.

2. Approach

2.1 Idea in Research Activity

People have the elements to be considered in research (research knowledge), and they differ depending on the research domain and methods. Therefore, it is considered that people use multiple types of research knowledge according to their research domains and methods.

Research outputs are the information obtained through investigation, analysis, and inference (topics), which are mapped to the research knowledge. In the process of deriving research outputs, several elements of research knowledge and topics are considered as a meaningful group, and the candidate outputs are created by developing and selecting the group. In this research, we call such a meaningful group an idea. Figure 1 shows the relationship among research knowledge, topics, and ideas.

For example, consider a researcher who has research knowledge such as “background,” “problem,” “method,” and “technology.” If the topic “programming learning” is corresponding to “background” and the topic “difficulty in understanding the grammar” is corresponding to “problem,” the researcher is considered to have the idea of “research to solve the difficulty in understanding the grammar in programming learning.” An idea can be only research knowledge or only a topic. For example, if a researcher is thinking of “background” because “background” is necessary for his/her research, he/she has an idea that exists only for “background,” and if he/she wants to do something using generative AI because generative AI is popular, he/she has an idea for “generative AI” only as a topic. In this way, we can say that an idea is a collection of research knowledge and topics that researchers focus on in order to create a candidate outputs.

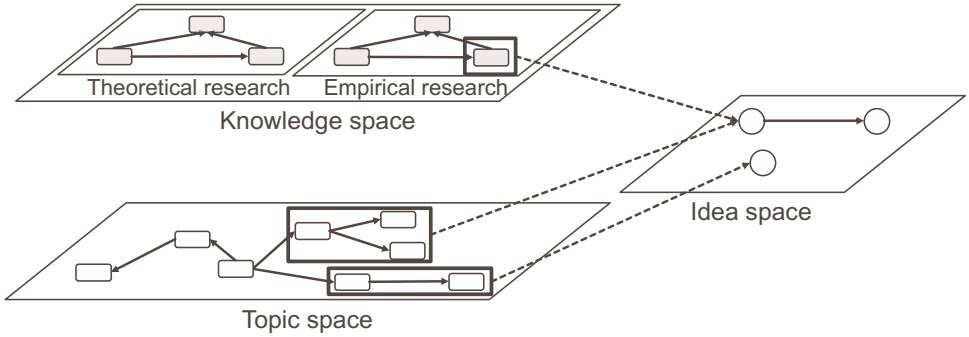


Figure 1. Relationship among Research Knowledge, Topics, and Ideas

2.2 Thinking Strategy and Its Estimation

In order to derive the research outputs, it is important to know what kind of thinking is to be done. In this study, a thinking strategy is a policy of thinking to achieve an objective. The objectives include generating certain elements of research knowledge and topic that corresponds to an element of research knowledge, such as “deriving a research approach” or “determining a research problem.” Thinking strategies include divergent thinking and convergent thinking (Guilford, 1967), which are strategies for expanding ideas, and data-driven thinking and goal-directed thinking (Russell & Norvig, 2010), which are strategies for how to proceed with reasoning.

Thinking transforms the idea space by updating or deriving ideas. Since thinking is carried out according to a thinking strategy, we can estimate the thinking strategy by looking at the process of changes of idea space by the implemented thinking strategy. For example, when thinking is done in such a way that a number of derived ideas are created in response to certain existing idea, it can be estimated that divergent thinking was used as a thinking

strategy. If the thinking that updates the existing ideas one after another is performed, it can be estimated that convergent thinking was used as a thinking strategy. If we can visualize the idea space change process, we can support the estimation of the thinking strategy.

2.3 Research Instruction Support Environment

In this study, we construct a research instruction support environment consisting of a thinking externalization system, an idea space change process visualization system, and a thinking database. The thinking externalization system provides an interface that enables learners to organize their ideas while expressing their own research knowledge and topics. The idea space change process visualization system has an idea space map generation function that generates a graph representing the idea space change process (idea space map) based on the learner's thoughts, and an interface for displaying the generated idea space map. The thinking database is a database that stores the learner's thinking.

The following describes the procedure for using the research instruction support environment. The learner expresses his/her research knowledge and topics on the interface of the thinking externalization system. The operations of expressing research knowledge and topics are stored in the thinking database. The instructor selects a learner whose thinking strategy he/she wants to estimate on the interface of the idea space change process visualization system. The idea space change process visualization system acquires the selected learner's operation series from the thinking database and generates idea space map with the idea space map generation function. The interface of the idea space change process visualization system displays the generated idea space map. The instructors estimates the learner's thinking strategy based on the visualized idea space map and provided guidance.

3. Thinking Externalization System

The thinking externalization system provides an interface for representing one's own research knowledge (knowledge externalization interface) and an interface for representing topics and mapping topics to research knowledge represented by the knowledge externalization interface (thinking externalization interface). Since knowledge and topics can be expressed in terms of elements and their relations, they are represented in a graph structure with the elements of research knowledge and topics as nodes and their relations as edges. The knowledge externalization interface is operated in the same way as the thinking externalization interface, so only the thinking externalization interface is described below.

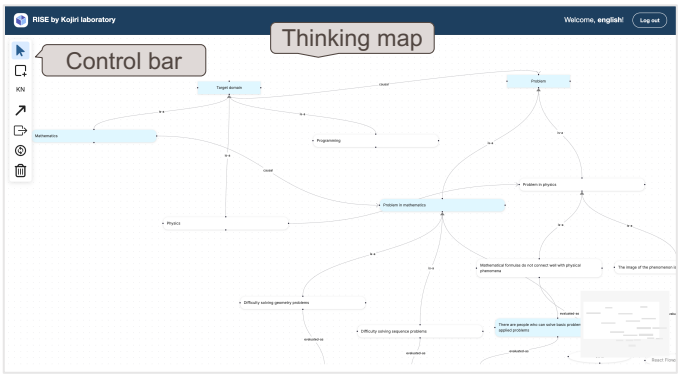


Figure 2. Thinking Externalization Interface

A thinking externalization interface must be able to represent and relate topics and research knowledge. As thinkings for representing topics and research knowledge, we have prepared operations for creating topic nodes and research knowledge nodes, and topic nodes with edges. The creation of a node with an edge is an operation to add a new topic node that is connected to an existing topic node by a specific relation. As mapping topics and research knowledge, we have prepared an operation to add edges between topic nodes and research knowledge nodes. As types of relations between topics, we define the part-of relation, which

is a detailing relation, the causal relation, which is a premise-conclusion relation, and the is-a relation, which is an abstraction relation. The is-a relation, which represents correspondence, defined as a relation between a topic node and a research knowledge node.

As evaluating topics, we have prepared operations for creating evaluation criterion nodes and evaluation value nodes. The evaluation criterion node describes from what point of view the topic is evaluated, and the evaluation value node describes the evaluation value from that point of view. As a selection of topics, an operation to label the topic nodes as adopted or rejected is provided.

Figure 2 shows the thinking externalization interface. The thinking externalization interface consists of a thinking map and a control bar. On the thinking map, research knowledge and topics are represented as nodes, and the relations between them are represented as edges. Nodes and edges are created and edited using the control bar.

4. Idea Space Change Process Visualization System

The idea space change process can be represented as a graph structure in which ideas are nodes and changes of ideas are edges. This graph is called an idea space map. The idea space map supports the estimation of thinking strategies from the shape of the graph structure by representing the idea space change process as a shape of the graph structure. There are two types of idea change: derivation and renewal. The operation of creating a different idea from an existing idea is called "derivation," and the operation of updating an existing idea by adding new elements is called "renewal." To express these two types of changes, we prepare two types of edges in the idea space map: derivation and renewal.

In order to generate idea space map, it is necessary to identify how operations on the thinking map transform the idea space. Operations on the thinking map are performed on specific research knowledge nodes or topic nodes (target node). The target node is included in one of the ideas. Therefore, we can say that the operations on the thinking map are operations that change the ideas that contain the target nodes. The idea space map can be created by three steps: 1. Identifying the nodes and edges targeted by the operations on the thinking map, 2. Identifying the ideas that contain those nodes and edges, 3. Adding new ideas to the idea space map and adding edges that represent changes between the ideas that contained the target nodes and edges. When an operation is performed that does not target other ideas, such as “adding a topic node,” a new idea is added to the idea space map.

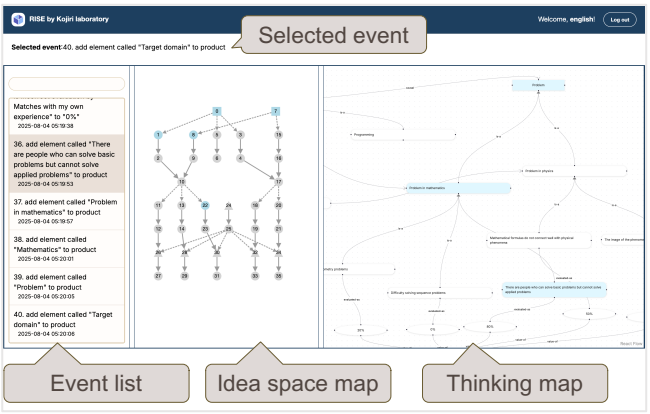


Figure 3. Interface of Idea Space Change Process Visualization System

The interface of the idea space change process visualization system is shown in Figure 3. This interface consists of three parts: an event list that displays the learner’s operations in one glance, a thinking map created by the learner, and an idea space map that displays the idea space change process generated from the learner’s thinking process. The nodes on the idea space map represent the ideas, and the edges represent the change between ideas caused by operations on the idea. Nodes are assigned the order of the operations that caused their state. By checking the node orders, the instructor can see in what order the operations were performed.

5. Evaluation Experiment

5.1 Experiment Settings

We have conducted an experiment whether the proposed idea space change process externalization system is effective in supporting research instruction. The effectiveness of research instruction is evaluated by comparing the instruction without using the proposed system with that when considering instruction using the proposed system.

The participants in the experiment were 9 graduate students (A-I) who can provide research instruction to undergraduate students. The following is the experimental procedure. First, the participants were presented with the research output that the authors had prepared in advance and the thinking map on which the research output was based, in a setting in which they had been created by a fictional learner. The prepared research output and the process for creating it included parts that may require instruction on thinking strategies, such as “deep consideration of the research topics, but neglecting to consider the consistency between the topics.” The research output was a slide in which the adopted parts of the thinking map ware listed.

The participants were asked to review the research output and thinking map provided to them and consider instructional contents (Step 1). Participants were instructed to think not only the presentation of research, but also the contents of research and the thinking behind them. Next, we asked them to consider instructional contents using the proposed idea space change process visualization system (Step 2). The idea space map corresponding to thinking map is shown in Figure 4. In this idea space map, thinkings about the “research background” and thinkings about the “research approach” are carried out independently, without any consideration of the consistency between them. In Step 2, in addition to instructional contents they had thought up in Step 1, we asked them to think of new instructional contents using the idea space map. In addition, if the participants felt that the instructional contents, they had considered in Step 1 were unnecessary or needed to be changed based on Step 2, they were asked to delete or change them. Finally, they filled out a questionnaire asking what they had been conscious of when considering the instructional content.

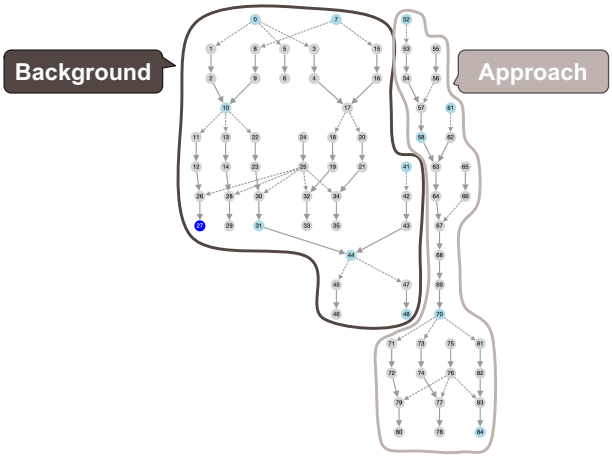


Figure 4. Idea Space Map Used in Experiment

5.2 Result and Discussion

To evaluate the effectiveness for supporting instruction for thinking strategies, we categorized the type of instructional contents. Since the research output provides instruction on the presentation of research, the thinking map provides instruction on the contents of research, and the idea space map provides instruction on thinking strategies, the classification was made into three types: 1. Instructional contents for presentation of research, 2. Instructional contents for contents of research, and 3. Instructional contents for thinking strategies. The number of instructional contents by type in Step 1 and Step 2 is shown in Table 1. Since Step

2 newly uses an idea space map, it is expected that the number of instructional contents for thinking strategies increased in Step 2.

A t-test was conducted to check whether there were significant differences between the number of instructional contents in Step 1 and Step 2 for each type. There was no significant difference in the number of instructional contents for presentation of research ($p > 0.05$), but there was a significant difference in the number of instructional contents for contents of research and thinking strategies ($p < 0.05$). This suggests that the idea space map provided insight into contents of research and thinking strategies. Example of instructional contents by referring to the idea space map included, “When thinking about your research approach, you should consider whether it is likely to fit the research problem,” and “The learner now has some understanding of what should be considered and in what order in research.” This can be said to be an example of how we were able to provide instructors with an awareness of learners’ thinking strategies.

Table 1. *The Number of Instructional Contents by Type in Step 1 and Step 2*

Type	Step	A	B	C	D	E	F	G	H	I	Avg.
1	1	2	1	1	0	7	0	4	1	0	1.78
	2	2	1	1	0	7	0	6	1	0	2.00
2	1	5	11	5	5	1	6	3	2	6	4.89
	2	6	15	7	7	1	10	5	2	7	6.67
3	1	1	0	2	1	2	0	0	1	0	0.78
	2	2	0	4	2	3	0	0	2	2	1.67

6. Conclusion

In this paper, we proposed a system that can express a learner’s thinking and a system that visualizes learner’s idea space change process as a graph structure to help instructors provide guidance to learner’s thinking strategies in research activities. The result of the evaluation experiment showed that by referring to the idea space map, the number of instructional contents for thinking strategies increased.

In the evaluation experiment, instruction was given to fictional learner. In order to verify whether proposed system can also provide an awareness of learner’s thinking strategies, and whether the derived instructional contents are useful to leaners, we need to conduct experiments with real learners. In the future, we will conduct an evaluation experiment in situations where instructors actually provide instruction to real learners.

Acknowledgements

This work was supported by JST SPRING, Grant Number JPMJSP2150.

References

Guilford, J. P. (1967). *The Nature of Human Intelligence.* New York: McGraw-Hill Inc.

Russell, S., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach* (3rd ed.). Upper Saddle River: Peason Education.

Salvati, G., Santoni, C., Tibaldo, V., & Pellacini, F. (2015). MeshHisto: Collaborative Modeling by Sharing and Retargeting Editing Histories. *ACM Transactions on Graphics, 34*(6), 1-10.

Smith, A., Anderson, B. R., Otto, J. T., Karth, I., Sun, Y., Chung, J. J. Y., Roemmele, M., & Kreminski, M. (2025, 2 7). *Fuzzy Linkography: Automatic Graphical Summarization of Creative Activity Traces.* Retrieved 3 2025, from arXiv: http://arxiv.org/abs/2502.04599

Tian, F., Yue, J., Chao, K., Qian, B., Shah, N., Li, L., Zhu, H., Chen, Y., Zeng, B., & Zheng, Q. (2019, 6 4). *Modeling e-Learners' Cognitive and Metacognitive Strategy in Comparative Question Solving.* Retrieved 3 2025, from arXiv: http://arxiv.org/abs/1906.03074

Yoshioka, M., Hayashi, Y., & Seta, K. (2019). Reflection Support System in Ill-defined Problem Solving. *The 27th International Conference on Computers in Education*, (pp. 405-413).

Audience Robot Control System Adapted to Presenter's Skills for Presentation Practice

Yuya KISHIMOTO[a*] & Tomoko KOJIRI[b]
[a] *Graduate School of Science and Engineering, Kansai University, Japan*
[b] *Faculty of Engineering Science, Kansai University, Japan*
*k504134@kansai-u.ac.jp

Abstract: During a presentation, a presenter must adapt his/her style by observing audience reactions. To help develop this skill, we have developed a system with an audience robot that reacts according to the presenter's non-verbal expressions. However, since there are multiple non-verbal expression, it is difficult for a presenter to modify all of them at once. The objective of this research is to develop a system that allows presenters to practice multiple non-verbal expressions step-by-step, by recommending specific expressions tailored to his/her skill level. The system uses the robot's sensors to assess the presenter's proficiency with each non-verbal expression. It then assigns target expressions for practice and configures the robot to react solely to those expressions, ensuring a focused training session.
Keywords: Presentation support, communication robot, non-verbal expression, presentation skill, training

1. Introduction

A presentation is an activity in which a presenter conveys his/her point of view to an audience. The presenter effectively communicates with audiences by utilizing verbal and non-verbal expressions. Since an understandable presentation differs depending on the audience and the environment, it is not always possible to give a good presentation. Therefore, it is necessary to observe the audience's reactions and modify one's own presentation according to the audience's reaction (this action is referred to "modify behavior").

Researches exist that support modifying the presenter's behavior during the presentation. (Kopf, Schön, Guthier, Rietsche, & Effelsberg, 2015) (Schneider, Börner, Rosmalen, & Specht, 2015). They evaluate whether the presenter's behaviors are in keeping with predefined good verbal and non-verbal expressions and provide direct feedback whether presenter's behaviors are appropriate in real time. Such researches encourage presenters to have predefined behavior, but it is not possible to cultivate the modify behavior. During the modify behavior, the presenter observes the audiences and modifies behavior dynamically according to the audience's reactions. The verbal and non-verbal expressions that audiences perceive as effective vary among individuals. Therefore, it is difficult to define these factors in predefined.

Palmas et al. deployed a virtual audience in a VR space that provides real-time feedback in reaction to the presenter's non-verbal expressions (Palmas, Cichor, Plecher, & Klinker, 2019). The necessity of special equipment to utilize VR space makes this system difficult to use casually. In addition, there are possibility of inducing VR sickness.

In order to solve such problems, we have developed an environment in which presenters can practice their presentations anytime, anywhere by developing an audience robot (Kato, Kunori, & Kojiri, 2024). This audience robot changes its internal state, which expresses emotion, in reaction to non-verbal expressions acquired by its sensors, and expresses its behavior in reaction to its internal state. Using this robot, it is possible to practice the activity of modifying non-verbal expressions in the audience robot's reactions. However, since there are multiple non-verbal expressions, it may be difficult for the presenter to modify all the non-verbal expressions at once. Therefore, the objective of this research is to develop the system, which we name audience robot control system, that determines non-verbal

expressions to be practiced according to the presenter's skill in each non-verbal expression. The target non-verbal expressions and their order are determined so that the presenter can start practicing with one non-verbal expression and increase the number of non-verbal expressions to be considered step by step until he/she finally becomes master of all the non-verbal expressions. In addition, this research extends the audience robot so as to adjust its reactions based on the determined non-verbal expressions for each practice session.

In the field of the intelligent tutoring systems, learning materials are determined based on the learner's understanding state and skills (Wenger, Brown, & Greeno, 1987). The system has materials, which are the contents of the learning, and the system selects content from the learning materials based on a learner model that represents the learner's state of skills. The audience robot control system we develop is regarded as an intelligent tutoring system.

2. Approach

Figure 1 shows the configuration of the audience robot control system and the audience robot.

We have developed an audience robot that reacts according to the quality of the presenter's non-verbal expressions (Kato, Kunori, & Kojiri, 2024). The robot uses various sensors to detect non-verbal expressions with its non-verbal expression detection function. Emotion generation function updates the emotion model of the audience robot based on the non-verbal expressions detected by the non-verbal expression detection function, and the reaction corresponding to the internal model is expressed by the reaction generation function.

In order for the audience robot to react to different non-verbal expressions depending on the presenter's skill, the emotion generation function should be able to update its model only from specific non-verbal expressions. Therefore, we introduce evaluation method determination function that dynamically changes the evaluation function of the non-verbal expressions used in the emotion generation function. In addition, sensor setting function is introduced to adjust sensor list to detect only the non-verbal expression of the target practice.

The evaluation function is set for each non-verbal expression and has a list of threshold values that represent the maximum and minimum of the range of having positive emotions, respectively. A positive emotion indicates when the value of the presenter's non-verbal expression is between the maximum and minimum values of the threshold, and a negative emotion shows the value is greater than the maximum or less than the minimum value. The audience robot gives positive or negative behaviors according to its emotion. In order to practice modify behavior, audience robot needs to give negative reaction for the target non-verbal expression. Therefore, the evaluation method determination function adjusts the range of maximum and minimum values according to the presenter’s usual non-verbal expression, which we call baseline behavior. That is, the range does not include the baseline behavior.

On the other hand, the audience robot control system maintains practice contents that consists of set of non-verbal expressions to be practiced and their order relations. It also has a user model that represents presenter’s skills. The user model is developed by evaluating the practice history with audience robot using user model development function.

The audience robot control function determines the non-verbal expressions targeted by the audience robot and sends the maximum or the minimum value values. The audience robot control function has two phases. First, a preparation phase that operates at the beginning of practice. Second, a practice phase that operates during the practice. The preparation phase is for gathering the baseline behavior (item 1) and for checking whether the presenter can perform the modifier behavior (item 2). For item 1, it makes presenter practice with audience robot which does not give any reaction. For item 2, it asks presenter to practice with audience robot with sensors of all non-verbal expressions and give reactions to all non-verbal expressions. In practice phase, it controls the audience robot by referring to the user model and the practice contents.

Currently, eye contact, speaking speed and volume of voice are the target non-verbal expressions.

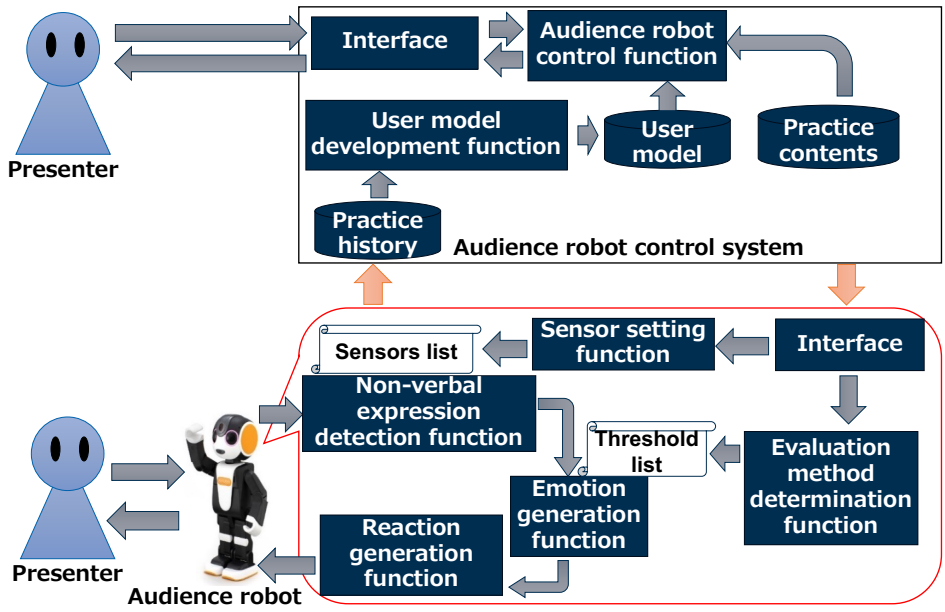


Figure 1. System Configuration of Audience Robot Control System

3. User Model Development Function

The user model development function identifies whether the presenter could control the non-verbal expressions based on the presenter's behavior during the presentation with the audience robot. In order to identify the well-controlled non-verbal expressions, it is necessary to grasp whether they are trying to modify the non-verbal expressions when the audience's reaction is negative, or whether they are trying to keep the non-verbal expressions when the audience's reaction is positive.

If the presenter tries to modify the inappropriate behavior or keeps the good behavior for most of the time, the presenter is regarded to perform the modify behavior. The user model development function calculates the average and standard deviation of the behavior in the previous 30 seconds and sets the range of the baseline behavior as *average – standard deviation* to *average + standard deviation*. Then, it detects the behavior as different from the baseline behavior if it is outside the range of the average and standard deviation. For the first 30 seconds, the average and standard deviation of the preparation phase behavior are considered as the baseline behavior.

On the other hand, if the presenter is trying to keep his/her own behavior, the good behavior will be kept.

The presenter usually does not always monitor the audiences and change his/her behavior, so if the presenter controls non-verbal expression, such as keeps or modifies the target non-verbal expression, more than 50 % of the presentation, the presenter is succeeded in modifying behavior. When there are more than one target non-verbal expression, if more than one non-verbal expression is well-controlled, the presenter is regarded to master the modify behavior of the target non-verbal expressions.

4. Practice Contents

The practice contents consists of a set of non-verbal expressions to be practiced (*contents*) by presenter and their relations. Since the final goal of the presenter is to control non-verbal expressions in the modify behavior, practice with all non-verbal expressions is the last contents to be practiced. Practicing with targeting only one non-verbal expression is the easiest, and difficulty level increases as the number of non-verbal expressions increases. Based on these considerations, all possible sets of non-verbal expressions are defined as contents and their inclusive relation was attached between them.

Figure 2 shows an example of the practice contents when eye contact, volume of voice, and speaking speed were selected as target non-verbal expressions.

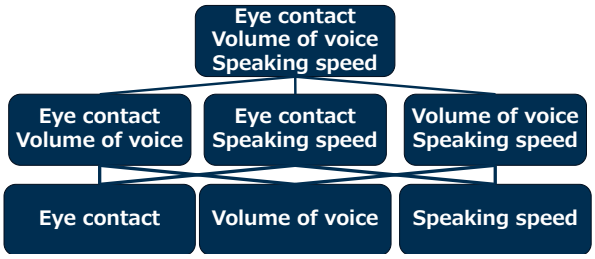


Figure 2. Example of Practice Contents

5. Audience Robot Control Function

Audience robot control function identifies a sequence of contents to be practiced by the presenter from the materials based on the user model.

5.1 User Model

The User model maintains the non-verbal expressions that presenter could control, his baseline behavior identified in the preparation phase such as their average and standard deviations of non-verbal expressions. In addition, the model maintains the contents used for practice and their achievement status.

5.2 How to Decide Which Non-verbal Expression to Practice

Figure 3 shows a flowchart for determining the practice contents. If the presenter could control all non-verbal expressions to modify behavior, the practice ends. If there are non-verbal expressions which the presenter is not able to control, the content consisting only of that non-verbal expression is first selected as practice content. If there are ancestor nodes for each practice content, they are selected as the practice content. Since the parent nodes of different non-verbal expressions may be the same node, the content is selected only once. The selected contents are provided, starting from the lower layer to the upper.

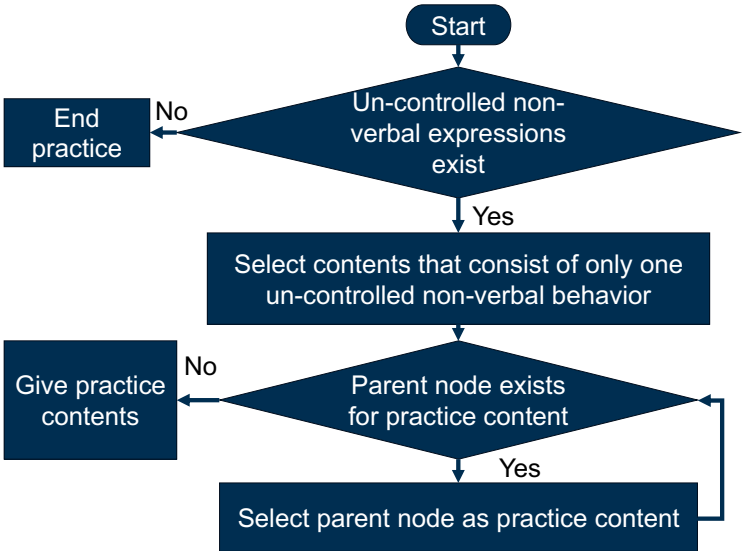


Figure 3. Practice Contents Selection Algorithm

6. Evaluation Experiment

6.1 Overview

An experiment was conducted to evaluate the effectiveness of the presentation practice environment developed in this research. 10 undergraduate and graduate students (A~J) participated in the experiment.

The experimental steps were as follows.

1. Participants give a presentation in the preparation phase (pre-presentation)
2. Participants were divided into two groups, WITH audience robot control function (*Group E*) and WITHOUT audience robot control function (*Group C*), and each group practiced 7 times.
3. Participants give a presentation in the preparation phase (post-presentation)

In this experiment, the participants were asked to give a presentation to an audience robot. Slides and dialogue for the presentation were prepared by authors. The presentation

themes and the number of slides used in the experiment are shown in Table 1. The lengths of the slides were about 3 to 4 minutes. Different slides were used for the pre-presentation, practice, and post-presentation to avoid the influence of the theme. Slides and dialogue were given to the participants before the presentation, but they were allowed to change the dialogue. During the presentation, we explained that the audience robot would evaluate the presentation and express its reaction according to the values of the non-verbal expressions displayed on the interface, and instructed the participants to change their non-verbal expressions when they felt that the audience robot's reaction was negative so as to improve the audience robot's reaction.

In step 1, the presentation is performed in the preparation phase. At the end of Step 1, the audience robot control system identifies the presenter's skill using the data from Step 1. In Step 2, the number of participants in the group WITH and WITHOUT the audience robot control function was equal. For both groups, participants were asked to give presentation for seven times. For participants without the audience robot control function, the three combinations of non-verbal expressions were presented every time. In Step 3, the presentation was conducted again in preparation phase to identify the presenter's skills.

Table 1. *Presentation Theme and the Number of Slides*

Presentation theme	The number of slides
1: Introduction of Kansai University	7
2: What is Chat GPT?	6
3: Introduction of Kojiri Laboratory	5

6.2 Result

Let's evaluate the effectiveness of the audience robot control function in item 1. Table 2 shows the percentage of participants who were able to perform the modify behavior in the pre-presentation and the post-presentation. A *t-test* was conducted between the pre-condition and the post-condition. The result was $p>0.05$ ($p=0.768$), indicating that there was not a significant difference. Table 3 shows the practice contents of *Group E* in all seven practice sessions, where "E" is eye contact, "V" is volume of voice, and "S" is speaking speed. All experimental participants in *Group E* could not reach to the final practice contents in seven sessions. Therefore, when the number of presentation practice is limited, the participants were not always able to perform all practice contents and, thus, were not guaranteed to improve their modify behavior.

Next, we evaluate the load on practice using the audience robot control function. If the practice is appropriate for the participants, the percentage of participants who could perform modify behavior increases. are controlled should maintain or improve with each practice session. For this purpose, we investigated the percentage of improvement in the target non-verbal expressions between exercises of the same practice contents in *Group E*. Table 4 shows the number of practice that participants continued the same practice contents and the ratio of achieving modify behavior during these practice. The *t-test* was conducted to examine whether there was a significant difference between *Group E* and *Group C*. The result was $p=0.047$ ($p<0.05$), indicating a statistically significant difference. This indicates that the load of modify behavior was lowered in *Group E* by practicing step-by-step.

The participant in *Group E* commented, "By observing the feedback from the audience robot control system and the reactions of the audience robot, I was able to notice that the speaking speed was changing unconsciously," and "By practicing in a step-by-step practice, it became easier to control the non-verbal expression and modify behavior." These results suggest that the audience robot control function is effective in acquiring modify behavior.

Table 2. *Percentage of Participants Who were Able to Perform Modify Behavior*

Group	Participants	Pre-presentation	Post-presentation
E	A	0.667	0.800
	B	0.833	1.000
	C	0.300	0.333

	D	0.846	0.600
	E	0.571	0.700
C	F	0.750	0.929
	G	0.600	0.625
	H	0.286	0.667
	I	0.667	0.300
	J	0.909	0.667

Table 3. *Practice Contents of Group E*

	1	2	3	4	5	6	7
A	E	E	E	V	S	S	S
B	E	E	S	S	S	S	S
C	E	V	V	V	V	V	V
D	E	S	S	S	S	EV	EV
E	V	V	V	V	S	S	S

Table 4. *The Number of Times Contents are Repeated and Rate of Improvement*

Group	Participants	The number of times practice contents are repeated	Improvement rate
E	A	4	0.75
	B	5	0.60
	C	5	1.00
	D	3	0.75
	E	5	0.80
C	F	6	0.67
	G	6	0.50
	H	6	0.50
	I	6	0.50
	J	6	0.75

7. Conclusion

In this research, we proposed a framework that determines the combination of non-verbal expressions to practice according to the presenter's non-verbal expression skills and adjusts the non-verbal expressions that the audience robot reacts in a presentation practice environment. In this research, the goal of learning is to present with controlled of all non-verbal expressions. In order to do so, the timing and the type of the non-verbal expressions to use important. It is considered that appropriate non-verbal expression to use depends on the intention to be conveyed by the non-verbal expressions. In the future, we need to analyze the contents that can be conveyed by each set of non-verbal expressions and expand the training environment to have the audience robot select the appropriate non-verbal expressions based on the type of reaction.

References

Kato, S., Kunori, K., & Kojiri,T. (2024). Development of Audience Robot for Supporting Presentation Rehearsal of Nonverbal Expression. Proceedings of EdMedia + Innovate Learning, 799-808.

Kopf, S., Schön, D., Guthier, B., Rietsche, R., & Effelsberg, W. (2015). A Real-time Feedback System for Presentation Skills. *Proceedings of In EdMedia+ Innovate Leaning: Association for the Advancement of Computing in Education*, 1686-1693.

Palmas, F., Cichor,J.E., Plecher,D. A., & Klinker, G., (2019). Acceptance and Effectiveness of a Virtual Reality Public Speaking Training. 2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), 363-371.

Schneider, J., Börner, D., Rosmalen, P. v., & Specht, M. (2015). Presentation Trainer, your Public Speaking Multimodal Coach. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 539-546.

Wenger, E., Brown, J. S., & Greeno, J. (1987). *ARTIFICIAL INTELLIGENCE and TUTORING SYSTEMS Computational and Cognitive Approaches to the Communication of Knowledge.* Morgan Kaufmann Publishers, Inc.

A Human-AI Collaborative Assessment of AI-Generated vs. Human-Created MCQ Distractors

Zifeng LIU[a*], Priyadharshini Ganapathy PRASAD[a], Bach NGO[b], Xinyue JIAO[c] & Wanli XING[a]

[a]*University of Florida, the United States*
[b]*The Frazer School, the United States*
[c] *New York University, the United States*
*liuzifeng@ufl.edu

Abstract: Multiple-choice questions (MCQs) are a widely used and effective assessment method, with the quality of distractors being crucial for their effectiveness. Recent studies have explored the use of large language models (LLMs) to generate distractors. However, generating high-quality distractors remains challenging in subjects such as computer science and mathematics that requiring strong reasoning. While some research has investigated AI-generated distractors in these fields, evaluating their quality is difficult due to existing metrics primarily focusing on surface semantics and failing to capture the necessary reasoning. This study introduces a human-AI collaborative assessment approach to evaluate distractor quality. We applied this method to compare AI-generated and human-created distractors in two high school courses: programming (N = 349) and statistics (N = 576). The findings suggest that AI-generated distractors can be competitive with human-created ones in programming courses, but significant differences exist in the understand, analyze, and evaluate types of MCQs in statistics ($p < 0.01$). This study provides a practical and scalable solution for integrating and evaluating AI-generated distractors in educational assessments.

Keywords: Distractor, generative AI, evaluation, human-AI collaboration, multi-choice question

1. Introduction

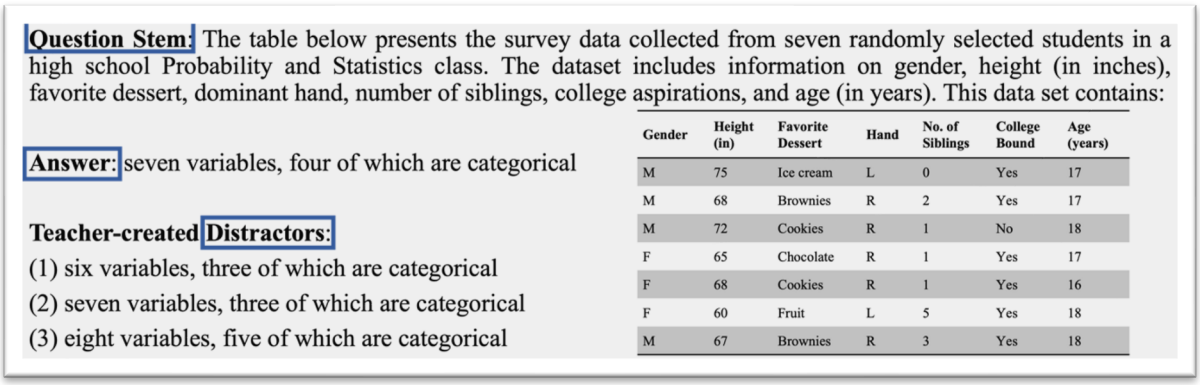
The rapid advancement of generative artificial intelligence (AI) has revolutionized assessments in education, where AI-driven approaches have been increasingly applied to automate generate test questions (Rodrigues et al., 2024) or distractors in multi-choice questions (MCQ) (Doughty et al., 2024). Recent AI methods for distractor generation include deep neural networks (e.g., Zhou et al., (2020)) and pre-trained language models (e.g., Bulathwela et al., (2023)). However, these methods have predominantly focused on generating distractors for reading comprehension and language learning assessments using techniques such as fine-tuning and prompting.

Recent research has explored various neural approaches to generating contextually appropriate multiple-choice question (MCQ) distractors. For example, the BERT-based Distractor Generation (BDG) model leverages multi-task learning and negative answer training to produce diverse and instructionally meaningful distractors (Chung et al., 2020). Building on this line of work, T5-based models have been adapted using techniques such as closed-book question answering and binary classification to improve the semantic relevance of distractors (Lelkes et al., 2021). Beyond fine-tuning, prompting has emerged as a lightweight yet powerful alternative, enabling pre-trained language models to generate distractors through textual instructions without the need for additional labeled data. Large language models (LLMs) like GPT-3, GPT-4, and ChatGPT have shown strong performance in this regard, using strategies such as zero-shot, few-shot, single-stage, and multi-stage prompting (Bitew et al., 2023;

McNicholas et al., 2023). For instance, zero-shot prompting involves providing the model with only the question stem and correct answer, asking it to generate plausible incorrect options. Few-shot prompting enhances the quality of distractors by incorporating several annotated examples within the prompt. In addition, chain-of-thought (CoT) prompting has been used to guide LLMs in simulating common student misconceptions, thereby improving the pedagogical effectiveness of generated distractors (Wei et al., 2022).

Despite the success of fine-tuned and prompted models in generating high-quality distractors, challenges remain. While AI-generated distractors can significantly reduce educators' workload by serving as an initial draft, human review and refinement remain necessary to ensure their pedagogical effectiveness. Evaluating the quality of AI-generated distractors is an identified challenge of this research area. Distractor evaluation methods are typically categorized into automatic and human evaluation techniques. Automatic evaluation employs metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) to assess similarity between generated and human-crafted distractors. Ranking-based and n-gram-based metrics further quantify the quality and relevance of distractors by analyzing their position within a ranked list or their lexical overlap with ground truth distractors (Lavie & Denkowski, 2009; Papineni et al., 2002). While these metrics offer scalable evaluation, they may not fully capture the pedagogical effectiveness of distractors.

To address these challenges, this study proposes a human-AI collaborative assessment framework to compare distractor quality from AI and from human teachers. We apply this framework to programming and statistics courses by comparing AI-generated distractors with human-created ones, enabling a more structured and scalable evaluation process. Specifically, our approach integrates fine-tuning and prompting-based AI models, explores advanced evaluation metrics that extend beyond surface-level similarity, and incorporates human-in-the-loop validation to improve both the accuracy and pedagogical value of AI-generated distractors. By systematically assessing the effectiveness of different AI-assisted techniques in MCQ design, this study contributes to a deeper understanding of how AI can enhance educational assessments.



Question Stem: The table below presents the survey data collected from seven randomly selected students in a high school Probability and Statistics class. The dataset includes information on gender, height (in inches), favorite dessert, dominant hand, number of siblings, college aspirations, and age (in years). This data set contains:

Gender	Height (in)	Favorite Dessert	Hand	No. of Siblings	College Bound	Age (years)
M	75	Ice cream	L	0	Yes	17
M	68	Brownies	R	2	Yes	17
M	72	Cookies	R	1	No	18
F	65	Chocolate	R	1	Yes	17
F	68	Cookies	R	1	Yes	16
F	60	Fruit	L	5	Yes	18
M	67	Brownies	R	3	Yes	18

Answer: seven variables, four of which are categorical

Teacher-created Distractors:

(1) six variables, three of which are categorical

(2) seven variables, three of which are categorical

(3) eight variables, five of which are categorical

Figure 1 An Example of the collected data

2. Method

2.1 Data Description

Data was collected on Florida Virtual School from two high school courses: Foundations of Programming and Foundations of Statistics. The programming course contains 349 MCQs, while the statistics course includes 576 questions. Typically, in this dataset, each MCQ consists of a question stem with textual supporting materials (e.g., tables), a correct answer, and three distractors manually created by teachers, shown in Figure 1.

Table 1. The Bloom's taxonomy levels of the MCQs in the dataset

Course	R	U	App	Ana	E	C	Total

Programming	91	110	87	40	18	3	349
Statistics	4	43	185	269	71	4	576
Total	95	153	272	309	89	7	925

R: Remember, U: Understand, App: Application, Ana: Analysis, E: Evaluation, C: Create

2.2 Bloom's Taxonomy Levels

To investigate AI's ability to generate distractors for different types of questions, we categorized all the MCQs (N = 925) based on Bloom's taxonomy (Krathwohl, 2002). All MCQs were classified into six cognitive levels: Remember, Understand, Apply, Analyze, Evaluate and Create. Two researchers both with expertise in educational assessment and taxonomy-based categorization, independently assigned taxonomy levels to each MCQ. The Inter-Rater Reliability (using Cohen's Kappa) is 0.67. For cases where their classifications differed (N = 309), a third researcher who was also an expert in educational assessment reviewed the discrepancies and selected the more accurate classification. The final distribution of categorized MCQs is presented in Table 1.

2.3 Distractor Generation

To generate distractors for MCQs, this study employed GPT-4 API, a state-of-the-art large language model that have been frequently used in recent studies to automate the creation of distractors or answers (Doughty et al., 2024; Rodrigues et al., 2024). To ensure consistency and relevance, a standardized prompt structure was designed, incorporating four components (i.e., the question stem, the correct answer, the corresponding Bloom's taxonomy level, and specific instructional constraints). The inclusion of Bloom's taxonomy categories is to guide GPT-4 in generating distractors that match the intended cognitive level of each question. The prompt explicitly instructed GPT-4 to generate three incorrect but plausible distractors in a structured format. The primary objective was to ensure that the generated distractors maintained semantic and cognitive alignment with the question stem and correct answer while varying in difficulty levels according to Bloom's taxonomy. GPT-4 was accessed through OpenAI's API with hyperparameters set to optimize creativity and reliability, specifically using a temperature of 0.7 and top-p of 0.9 to balance diversity and control in response generation.

Following the generation process, all distractors underwent a two-step validation process to ensure quality and reliability. First, an automated filtering step was applied to identify and eliminate instances of empty or erroneous generations. In cases where the generated output was incomplete or contained errors, GPT-4 was re-invoked to regenerate the distractors. Second, a human expert review was conducted to refine the dataset. Any extraneous content beyond the intended distractors was identified and removed. The validated distractors were then incorporated into the final MCQ dataset for further analysis. Distractors that duplicate teacher-created distractors are included.

2.4 Human-AI Collaboration Distractor Evaluation

To evaluate the quality of AI-generated distractors in comparison to teacher-created distractors, we employed a human-AI collaborative framework involving both human experts and state-of-the-art generative AI models (i.e., DeepSeek-V3 (DeepSeek-AI et al., 2024) and GPT-4o (OpenAI et al., 2024)). This hybrid evaluation approach was designed to ensure a unbiased assessment of distractor plausibility and pedagogical effectiveness while optimizing human resource allocation. The evaluation process consisted of three key steps. First, for each MCQ, a total of six distractors (three generated by teachers and three by GPT-4) were randomly shuffled and anonymized to prevent evaluators from identifying their source. Then, each human expert and GenAI model independently assessed all six distractors and selected the three most plausible and pedagogically effective options. The GenAI models were instructed using the following prompt: "You are a programming/statistics course instructor

evaluating distractors for multi-choice questions. Select the three most effective distractors based on plausibility, similarity to the correct answer, and ability to challenge the test-taker." After collecting the selections, the final set of three distractors was determined through a ranking process. The distractor with the highest selection frequency across all evaluators was chosen first (e.g., Distractor 4 in Figure 2). Among the remaining distractors with the same selection frequency, those endorsed by both AI and human evaluators were prioritized (e.g., Distractor 2 and Distractor 3). This selection process ensured that the final distractors represented the most effective choices from human and AI.

To determine whether GPT-4-generated distractors were of comparable or superior quality to teacher-created ones, we aggregated the selection results across all evaluators. The frequency with which each distractor was selected served as an indicator of its quality. If AI-generated distractors were chosen more frequently than teacher-created distractors across multiple MCQs, it suggested that AI could generate plausible and effective distractors. Conversely, if teacher-created distractors were selected more often, it indicated that human-authored distractors remained more pedagogically valid. Finally, a Wilcoxon Signed-Rank Test was conducted at the six Bloom's taxonomy levels to assess whether the differences in selection rates between AI-generated and teacher-created distractors were significant.

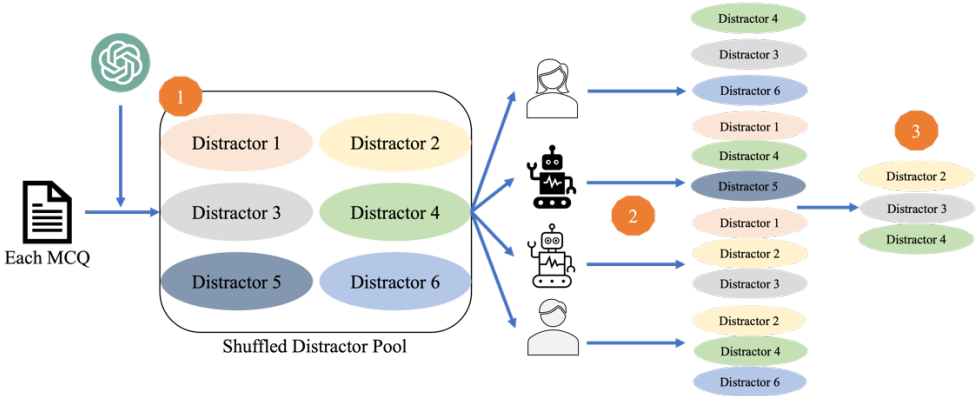


Figure 2 The Human-AI collaboration evaluation process.

3. Results

3.1 AI-Generated Distractors

A descriptive analysis of the GPT-4-generated distractors is shown in Table 2. Length represents the number of words in each distractor. Flesch Reading Ease and Flesch-Kincaid Grade Level measure readability, and the results indicate similar readability levels across both courses. Distractor Similarity was calculated using the Jaccard similarity between the three generated distractors for each question. For Answer Similarity, the Levenshtein distance between each distractor and the correct answer was computed, where a larger distance indicates greater dissimilarity. Due to the presence of numeric distractors, the minimum Levenshtein distance across both courses ranged from 2 to 3, while the average distance was between 40 and 68, indicating a reasonable level of distinction from the correct answers.

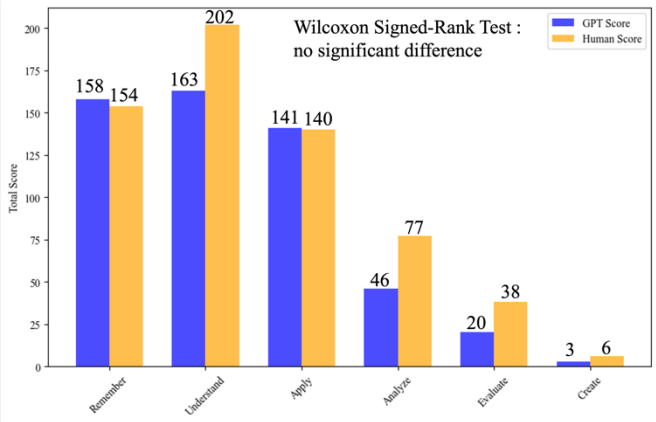
Table 2 Description of the AI-generated distractors

	Programming			Statistic		
Metric	Min	Mean (SD)	Max	Min	Mean (SD)	Max
Length	3	33.9 (26.2)	95	3	52.2 (50.7)	352
Reading	31.3	68.4 (38.2)	121.2	29.2	78.9 (34.3)	121.2
FKGL	3.5	5.1 (5.6)	16.5	3.5	4.3 (5.7)	19.2
J Similarity	-	0.01 (0.06)	1	-	0.01 (0.08)	1
L Similarity	3	41.0 (27.2)	210	2	67.4 (54.7)	351

Reading: Flesch Reading Ease, FKGL: Flesch-Kincaid Grade Level, J Similarity: Flesch-Kincaid Grade Level, L Similarity: Answer Similarity (Levenshtein)

3.2 AI-Generated vs Human-created Distractors

From Figure 3, it is evident that for programming courses, ChatGPT and AI perform consistently on MCQs related to Remember, Apply, and Create levels of Bloom's taxonomy. However, for other types such as Understand, Analyze, and Evaluate, human distractors score higher. Further analysis using the Wilcoxon signed-rank test reveals no significant differences in scores between human and AI for these types of MCQs. The table indicates that there are nine questions each where human and AI scores are equal for Remember and Understand questions. For statistics courses, AI and human teachers perform similarly on Remember, Apply, and Create questions. However, human teachers perform better in the other three dimensions. Further Wilcoxon analysis shows significant differences in Understand, Analyze, and Evaluate, with the following results: ($p = 0.002$, with a small effect size = 0.17), ($p < 0.001$, with a large effect size = 0.58), and ($p = 0.008$, with a medium effect size = 0.32). This indicates that AI has a noticeable gap in generating distractors for Analyze-type MCQs compared to human teachers.



Bloom Level	Count Equal	Count GPT_Higher	Count GPT_Lower
Analyze	1	11	28
Apply	4	39	44
Create	0	1	2
Evaluate	0	5	13
Remember	9	42	40
Understand	9	41	61

Figure 3 The evaluation results

4. Discussion

Advancements in AI-driven distractor generation offer promising opportunities for automating MCQ creation. Our findings highlight both the strengths and limitations of AI-generated distractors compared to human-crafted ones. GPT-4-generated distractors performed comparably to human-created ones in lower-order cognitive skills (Remember, Apply, and Create), aligning with prior research on pre-trained language models' ability to generate diverse, context-aware distractors (Bulathwela et al., 2023; Lelkes et al., 2021). Low distractor similarity scores further suggest AI-generated options are varied and non-redundant. However, challenges arise in higher-order cognitive skills (Understand, Analyze, Evaluate). While no significant differences emerged in programming MCQs, which aligns with previous work focus on programming distractors (Doughty et al., 2024), statistics MCQs showed substantial gaps in Understand, Analyze, and Evaluate MCQs. The large effect size in Analyze suggests AI struggles with nuanced reasoning, consistent with prior studies on the difficulty of generating pedagogically effective distractors for complex tasks (Bitew et al., 2023; Wei et al., 2022). This limitation likely stems from AI's challenges in semantic reasoning and conceptual understanding. Though strategies like few-shot and chain-of-thought prompting aim to improve distractor plausibility (McNichols et al., 2023), our results indicate their shortcomings in abstract reasoning-heavy disciplines like statistics. Building on previous studies (Atchley et al., 2024; Doughty et al., 2024), this study proposes the human-AI collaborative assessment of distractor quality, which to some extent alleviates the workload of fully manual evaluation. Future research should refine hybrid methods integrating fine-tuning and prompting, develop

more robust evaluation metrics, and explore human-in-the-loop validation. Additionally, instructional MCQs, which demand multi-modal analysis, warrant further investigation. By improving AI-assisted question design, this research supports scalable, effective assessment tools for educators.

Acknowledgements

The project is funded by the National Science Foundation (NSF) of the United States under grant numbers 1503196 and 2105695.

References

Atchley, P., Pannell, H., Wofford, K., Hopkins, M., & Atchley, R. A. (2024). Human and AI collaboration in the higher education environment: Opportunities and concerns. *Cognitive Research: Principles and Implications, 9*(1), 20. https://doi.org/10.1186/s41235-024-00547-9

Bitew, S. K., Deleu, J., Develder, C., & Demeester, T. (2023). Distractor generation for multiple-choice questions with predictive prompting and large language models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 48–63). Springer. https://doi.org/10.1007/978-3-031-74627-7_4

Bulathwela, S., Muse, H., & Yilmaz, E. (2023). Scalable educational question generation with pre-trained language models. In *International Conference on Artificial Intelligence in Education* (pp. 327–339). Springer. https://doi.org/10.1007/978-3-031-36272-9_27

Chung, H.-L., Chan, Y.-H., & Fan, Y.-C. (2020). A BERT-based distractor generation scheme with multi-tasking and negative answer training strategies. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 4390–4400). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.393

DeepSeek-AI, Liu, A., Feng, B., Xue, B., et al. (2024). Deepseek-v3 technical report. *arXiv*. https://arxiv.org/abs/2412.19437

Doughty, J., Wan, Z., Bompelli, A., Qayum, J., Wang, T., Zhang, J., Zheng, Y., Doyle, A., Sridhar, P., Agarwal, A., Bogart, C., Keylor, E., Kultur, C., Savelka, J., & Sakr, M. (2024, January 29–February 2). A comparative study of AI-generated (GPT-4) and human-crafted MCQs in programming education. In *Proceedings of the Australian Computing Education Conference (ACE 2024)* (p. 10). ACM. https://doi.org/10.1145/3636243.3636256

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice, 41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Lavie, A., & Denkowski, M. J. (2009). The METEOR metric for automatic evaluation of machine translation. *Machine Translation, 23*, 105–115. https://doi.org/10.1007/s10590-009-9059-4

Lelkes, A. D., Tran, V. Q., & Yu, C. (2021). Quiz-style question generation for news stories. In *Proceedings of the Web Conference 2021* (pp. 2501–2511). https://doi.org/10.1145/3442381.3449892

Lin, C. Y. (2004, July). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out* (pp. 74–81). Association for Computational Linguistics.

McNichols, H., Feng, W., Lee, J., Scarlatos, A., Smith, D., Woodhead, S., & Lan, A. S. (2023). Exploring automated distractor and feedback generation for math multiple-choice questions via in-context learning. *CoRR*. arXiv:2308.03234. https://doi.org/10.48550/arXiv.2308.03234

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., et al. (2024). GPT-4 technical report. *arXiv*. https://arxiv.org/abs/2303.08774

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Rodrigues, L., Pereira, F. D., Cabral, L., Gašević, D., Ramalho, G., & Mello, R. F. (2024). Assessing the quality of automatic-generated short answers using GPT-4. *Computers and Education: Artificial Intelligence, 7*, 100248. https://doi.org/10.1016/j.caeai.2024.100248

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35*, 24824–24837. https://doi.org/10.5555/3600270.3602070

Zhou, X., Luo, S., & Wu, Y. (2020). Co-attention hierarchical network: Generating coherent long distractors for reading comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*, 9725–9732. https://doi.org/10.1609/aaai.v34i05.6522

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

AI Support for Translanguaging in International School Classrooms

Alina GUHA*, Gyeongri KIM, Shamya KARUMBAIAH, Esther BETTNEY HEIDT & Mariana CASTRO
University of Wisconsin - Madison, USA

Abstract: Translanguaging, a pedagogical approach that utilizes multilingual learners' full linguistic repertoires, is especially relevant in linguistically diverse international school contexts. While translanguaging holds potential to improve academic outcomes and support students' linguistic identities, it remains difficult to implement due to language barriers. Artificial Intelligence (AI) can support translanguaging practices by providing in-the-moment language support and translations. This study explores the connections between international school teachers' beliefs and practices around translanguaging, multilingualism, technology and AI. We compare the survey responses from 105 international school teachers and 232 U.S.-based teachers, analyzing the correlation of scores derived from Likert-scale items. Our findings indicate that international school teachers' translanguaging beliefs and multilingual beliefs are significantly correlated with their reported translanguaging classroom practices, international school teachers' beliefs on decoloniality are significantly correlated with their beliefs on translanguaging, multilingualism and their translanguaging practices, and there were no significant correlations between teachers' AI attitudes and their other reported beliefs and practices. In contrast, U.S. teachers had no significant correlation between translanguaging beliefs and translanguaging practices and had a significant correlation between their translanguaging beliefs and their AI attitudes. We discuss how the structural and cultural contexts of international schools may enable more organic translanguaging, teachers' nuanced views on decoloniality and AI and how these contextual differences may impact the design of AI tools to support equitable, multilingual education through translanguaging.

Keywords: Translanguaging, Survey, Teacher Beliefs, Current Practices, AI

1. Introduction

Translanguaging is a theory from bilingual education that encourages students to utilize their entire linguistic repertoires (Garcia & Wei, 2014). This means going beyond socially and politically defined categories of languages (Otheguy, et. al., 2015) and moving fluidly between all languages in the classroom. Translanguaging is a natural practice in the daily lives of multilingual speakers, but has not been explicitly accepted in many educational contexts where stricter language separation may be the norm. This is documented in the case of U.S. classrooms, where research is being done on how students and teachers feel about translanguaging (Ticheloven, et. al., 2021) and education has historically been rooted in monoglossic ideologies (Garcia & Wei, 2014; Marsh, 2018;). Beyond the U.S., in the context of international schools or schools in different countries, there may be more factors at play which allow for translanguaging practices in the classroom.

In the case of international schools, educators may be following educational policies and structures that are influenced by Western perspectives that may not value students' natural multilingual abilities (Bettney Heidt & Nordmeyer, 2023) or there may be a focus on national languages in the curriculum that do not leave room for the complexities of their multilingual and multicultural student populations thus ignoring minority languages (Cenoz, 2017).

Translanguaging as a pedagogical framework aims to address these linguistic injustices in the classroom, by viewing students' nuanced multilingual abilities as an asset and making an active effort to bring these into the classroom. This can lead to better

academic outcomes where students are able to use their entire linguistic repertoires to support meaning making and communicate their understanding (Garcia & Wei, 2014) and support the development of students' linguistic identities thus increasing their self confidence and autonomy (Vogel & Garcia, 2017). Especially in international contexts, it would be beneficial to incorporate this framework into students' regular coursework to have their classroom practices reflect their daily lives as multilingual learners.

While translanguaging can be helpful for multilingual learners in theory, it can be difficult for teachers to put this theory into practice (Ticheloven, et. al., 2023). There may be school or national policies that prevent teachers from trying to incorporate translanguaging in their classrooms (Cenoz, 2017). Even if teachers are allowed to practice translanguaging, there may be logistical reasons that make this difficult to do - such as the linguistic diversity of the classroom or teachers' language backgrounds. If students speak different languages from each other and teachers are also not fluent in all of these languages, it can be challenging to allow and encourage multilingual learners to draw fluidly on their languages and to have these contributions be communicated and accepted by all participants in the classroom.

Although there may still be a need for deeper shifts around mindsets and policies about translanguaging, in terms of the logistical challenges, AI can be a helpful tool to facilitate translanguaging in classrooms where it may have been challenging in the past (Donley, 2024; Yang & Lin, 2025). Translation tools, generating text in different languages, and AI as a conversational partner are some of the ways in which AI, or Multilingual Large Language Models (MLLMs), can support teachers in encouraging students to utilize their entire linguistic repertoires.

In order to effectively implement these technological solutions, it is important to collaborate with teachers and consider their current perceptions and practices. Teacher perception has a significant impact on how technology is used in the classroom and what effect it has on the students (Farjon, et. al., 2017). In this paper, we focus on international school teachers' beliefs on translanguaging, multilingualism, their technology and translanguaging practices, their attitudes on AI, and how all of these might relate to each other. This information can be used as a basis for future work on AI implementation to support translanguaging in international classrooms. We will be exploring the following research questions:

1. *What are the connections between international school teachers' reported beliefs and practices about translanguaging, multilingualism, technology, and AI?*
2. *How do their beliefs and practices compare with U.S.-based teachers?*

2. Methods

2.1 *Survey Design and Development*

We designed a survey on translanguaging and AI inspired from past surveys on these topics (Samperio, 2017; Leonet, et. al., 2020; Dockrell, et. al., 2022) and relevant literature to fill in the gaps (Garcia & Wei, 2014; Vogel & Garcia, 2017; Alharbi, 2023). We conducted four think-alouds with current or past teachers of multilingual learners and two expert reviews with experts in education for multilingual learners and refined the survey further based on feedback from these. The survey took around 25 - 30 minutes to complete.

The Qualtrics survey was distributed via Prolific and University of Wisconsin-Madison's mailing list of teachers in June 2024. We received 232 complete responses to this survey from teachers of multilingual learners across the U.S. teaching a range of grade levels, with different years of experience, and teaching various subjects.

This survey was then adapted for teachers in an international school context. There were eight Likert items on decoloniality added that were relevant to language pedagogies and translanguaging especially in an international context. We drew from literature on multilingual practices and decoloniality (Bettney Heidt 2024, Bettney Heidt & Nordmeyer 2023) to develop these Likert items. Some examples are '*I would like to encourage the use of minority languages (e.g., Quechua, Sami) in my classroom*', '*It is important for students to*

become proficient in standard and accepted English'. We also modified other Likert items from the U.S. survey to account for the difference in context - this included changing terms (e.g., bi/multilingual learners to multilingual learners) and considering the language dynamics in international schools compared to local schools in the U.S.. We conducted three teacher think alouds with current international school teachers and one review with an expert in international education to check for any questions that would not translate to this context and any missing elements that might be particular to international schools. Many changes were made to the demographic section to consider teachers' backgrounds and their school and classroom contexts.

2.2 *Survey Distribution*

The survey was distributed in collaboration with the Multilingual Learning Research Center's (MLRC), a research center within the UW-Madison School of Education. The MLRC collaborates with over 100 international schools worldwide through the MLRC School Network, a research-practice partnership focused on exploring common problems of practice in teaching multilingual learners. The MLRC distributed the survey through its mailing list and social media accounts. This intentional sampling allowed for a focus on international school teachers which matched our participation requirements but it is important to note it may contribute to sampling bias, as the teachers in this network may have different beliefs and practices than the general international school teacher population.

We received 105 complete responses: 55% of the teachers' schools were based in Asia (including China, Singapore, India etc.), 14% from European countries (including the Netherlands, Germany, UK etc.), 11% from North America (all of these were based in the U.S.), 11% from South America (including Brazil, Guatemala etc.) and the few remaining respondents were teaching in Morocco, Mozambique or Dubai. 70% of the teachers were from MLRC schools and the rest were from other international schools. The majority were from teachers with more than 5 years of experience working in international education (11 or more years: 56; 6 to 10 years: 32; 3 to 5 years: 13; 1 to 2 years: 2; Less than 1 year: 2), most reported having more than 50% multilingual students in their classroom (More than 50%: 85; Around 50%: 17; Less than 30% : 3), and they taught a variety of age groups (2 to 4 years: 6; 5 to 10 years: 54; 11 to 13 years: 55; 14 to 16 years: 46; 17 to 19 years: 24).

2.3 *Data Analysis and Score Curation*

We curated scores to approximate teachers' perceptions and practices based on their responses to the relevant Likert scale questions in the survey. This allowed us to find patterns across teachers and search for any relevant correlations, using Pearson's correlation analysis on these calculated scores.

For the scores determining their alignment with theories of translanguaging, decoloniality, beliefs on multilingualism, and attitudes on AI - we first assigned a positive and negative valence to each Likert item in the relevant section of the survey. The scores for each respondent were calculated based on their Likert scale response by adding a value between -2 and +2 for each item based on the valence and relevant theory. Figure 1 provides examples of positive and negative valence items for the translanguaging score and how they would be added to the final score for each respondent. These scores were used to aggregate information from respondents allowing us to analyze teachers' overall perceptions of these theories and how they might relate to one another as well as their classroom practices.

ML (+)	Item: Speaking more than one language is an advantage for learning at school
ML (-)	Item: Teaching students in more than one language (e.g., English and Spanish) results in language confusion
TL (+)	Item: I would feel comfortable if my students switched between their language(s) in my classroom
TL (-)	Item: It is best to keep languages separate when teaching

Figure 1. Sample Likert items and valences assigned. ML = Beliefs about multilingualism. TL = Beliefs about translanguaging.

For the score calculating the reported frequency of translanguaging practices and technology practices amongst students in the classroom, we used the questions that addressed this directly. There were five Likert items on allowing students to translanguage in different contexts, these were added up to determine a score between 0 to 20. There were 11 Likert items on the same scale on teachers' reported use of technology in different contexts and these were added up to determine a score between 0 and 44.

3. Results

3.1 *RQ1: What are the connections between international school teachers' reported beliefs and practices about translanguaging, multilingualism, technology and AI?*

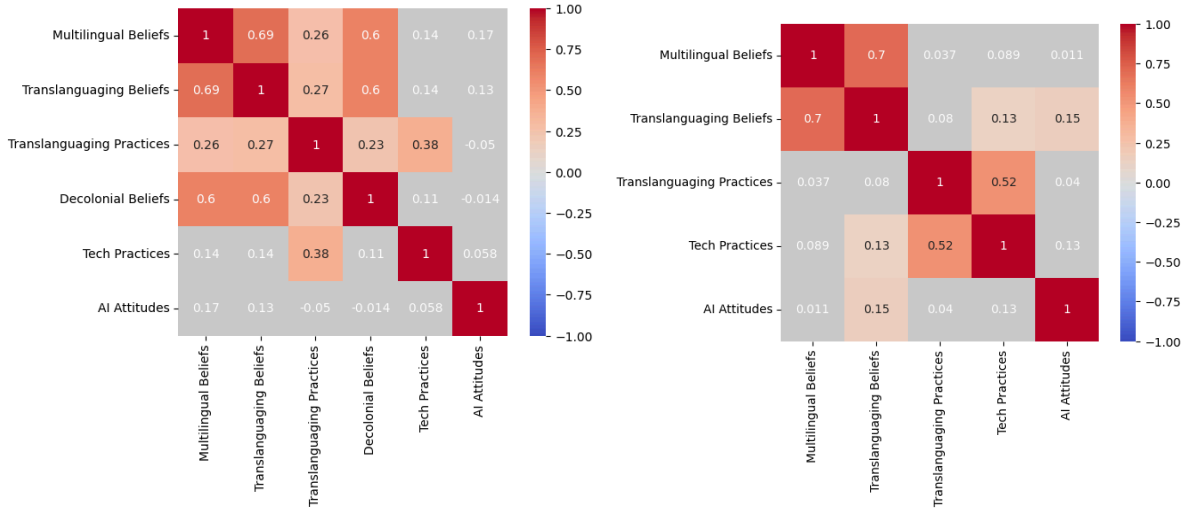


Figure 2. Correlation heatmap of constructed beliefs and practices of teachers from international (*left*; n = 105) and U.S.-based (*right*; n = 232) schools. Gray values are not significant ($p > 0.05$).

We found a significant positive correlation between teachers' beliefs about translanguaging and multilingualism (0.69, $p < 0.01$). This was expected, given that both theories reflect asset-based understandings of students' linguistic abilities.

We also found a significant positive correlation between teachers' beliefs about translanguaging and their reported translanguaging practices (0.27, $p < 0.01$) as well as a significant positive correlation between teachers' beliefs about multilingualism and reported translanguaging practices (0.26, $p < 0.01$). This indicates that teachers' classroom practices may be slightly aligned with their own theoretical or pedagogical beliefs, such that teachers that have positive views towards translanguaging and multilingualism may report higher frequencies of translanguaging practices in their classrooms.

There were significant correlations across different beliefs and practices with teachers' reported beliefs on decoloniality. There was a significant positive correlation between teachers' beliefs on decoloniality and their beliefs on translanguaging (0.6, $p < 0.01$) as well as with their beliefs on multilingualism (0.6, $p < 0.01$). This indicates that teachers that have positive views on these theories also tend to adopt critical views on coloniality in the context of international education. There was a significant positive correlation between teachers' beliefs on decoloniality and their translanguaging practices (0.23, $p < 0.01$), showing that teachers' anti-colonial views might also have an effect on students' language practices in the classroom.

There was a significant positive correlation between teachers' reported translanguaging practices and their technology practices (0.38, $p < 0.01$). Some hypotheses for this correlation may be a shared open-mindedness - where teachers are willing to adopt both new technologies and new pedagogies - or it may be an increased use of translation tools that accounts for the positive correlation - where teachers that have a lot students translanguaging need to use technology more often in order to communicate.

There were no significant correlations between teachers' AI attitudes and any of the other scores. Whether teachers have positive or negative views on translanguaging, multilingualism, and decoloniality do not seem to imply positive or negative views on AI, as there was a wide variety of qualitative responses. For example, some teachers who showed positive attitudes about translanguaging and multilingualism had concerns about AI. One teacher said, "When translating some words or expressions with specific cultural connotations, [AI] may not be able to accurately convey their meanings, leading to misunderstandings among multilingual students…" On the other hand, there were teachers who had positive attitudes about translanguaging and multilingualism and positive AI attitudes: "I do see that, when we gave students access to a bot to help them improve their writing, several students who are quieter made significant use of it to get feedback, which was great to see."

Connected to the decoloniality perspective, some respondents had a negative attitude toward AI while they stressed the importance of multilingual and multicultural backgrounds. They viewed AI as a tool that contains bias and is linguistically and culturally colonialized. One teacher said "AI systems trained on data primarily from Western cultures may exhibit biases towards "WEIRD" (Western, Educated, Industrialized, Rich, Democratic) perspectives, neglecting the views and experiences of minority groups, women, and underprivileged communities." Another teacher wrote, "There could be a lack of cultural sensitivity in AI-generated content, which may mislead or offend multilingual students."

AI attitudes did not have any significant correlation with reported technology practices (i.e., use of laptops, iPads, translation tools, Google search, classroom management systems, etc.). Teachers had more nuanced views on AI (i.e., chatbots, AI used for students' writing, creating class materials for teachers, etc.) that do not necessarily apply to their reported technology practices, implying teachers view these technologies as fundamentally different from AI.

3.2 *RQ2: How do their beliefs and practices compare with U.S.-based teachers?*

Teachers in the U.S. and international teachers had a significant positive correlation between their multilingual beliefs and translanguaging beliefs. However in the U.S. survey, there was no significant correlation between their translanguaging beliefs and their translanguaging practices, and no significant correlation between their multilingual beliefs and their translanguaging practices either. This indicates that there is a discrepancy between U.S. teachers' views on translanguaging and whether they are able to enact these views through practices in their own classrooms. Thus it is notable that for international school teachers there is a significant positive correlation between teachers' beliefs and their practices. This might indicate that translanguaging happens more naturally in international classrooms, as reported by teachers. Potential hypotheses for why international school teachers might be able to translate their beliefs to practices while U.S. teachers may find this more challenging could be because of the naturally multilingual nature of international classrooms or because teachers have more autonomy and are not constricted by school policy. When asked about their school policy, several teachers mentioned a positive attitude towards informal translanguaging - "English as the main teaching language. Bilingual communication is encouraged in daily interactions," or "primarily uses English as the language of instruction but we encourage the use of any language that assists in learning."

U.S. teachers' responses showed a significant positive correlation between their translanguaging practices and their technology practices (0.52, $p < 0.01$). International school teachers had a similar correlation, though it was not as strong (0.38, $p < 0.01$). International school teachers had no other significant correlations with their reported

technology practices nor their attitudes on AI. Whereas U.S. teachers had a correlation with their translanguaging beliefs and technology practices (0.13, $p < 0.05$) and their translanguaging beliefs and AI attitudes (0.15, $p < 0.05$).

4. Discussion

We compared survey results for international school teachers and U.S. school teachers and found that international school teachers are able to translate their translanguaging beliefs into classroom practices, teachers' translanguaging beliefs and practices are correlated with their beliefs on decoloniality, and their attitudes towards AI had no significant correlations with any of the beliefs or practices. This can help inform future AI developments to support multilingual learners in international schools: by considering teachers' beliefs and current practices AI design can be better aligned with classroom realities.

4.1 *International school teachers' translanguaging beliefs are positively correlated with their reported translanguaging practices*

In consideration of the stronger correlation between translanguaging beliefs and practices amongst international school teachers compared to U.S. teachers, we suggest this may be related to the higher levels of autonomy available to international school teachers. In contrast to national public schools, which are influenced by multiple layers of national or regional governance, international schools typically function with a degree of independence from local educational authorities. This autonomy frequently results in policy compression, as international schools develop and implement their own programs and policies (Bettney Heidt & Nordmeyer, 2021). Within this context, international school teachers may be able to implement translanguaging practices aligned with their beliefs in ways that are not possible in a more policy-constricted space within the U.S.

Another contributing factor may be that in recent years, there has been a noted increase in linguistic diversity within international schools which may contribute to the necessity for international school teachers to enact translanguaging practices within their classrooms. Recent research proposes that international schools serve mainly students who do not come from English-speaking schools (Tanu 2018, Crisfield 2023). We suggest international school teachers translanguaging beliefs and practices may be more aligned because of the necessity to find ways to support their multilingual students on a daily basis.

4.2 *International school teachers' beliefs on decoloniality are correlated with their positive beliefs on translanguaging and multilingualism and their translanguaging practices*

Our results demonstrate a positive correlation between beliefs about decoloniality, translanguaging and multilingualism. While we do not have comparative data from the U.S. study for this correlation, there is some evidence to suggest an emerging relationship between decolonization and critical understandings of language in international schools. Bettney Heidt and Nordmeyer (2023) link international school educators' growing awareness of critical approaches, such as decolonization, to shifts from monolingual to multilingual mindsets, including embracing translanguaging. However, these changes are not yet widespread, as shown in Bettney Heidt's (2023) case study of a Colombian international school, where beliefs ranged from maintaining English's perceived superiority to critically addressing colonial mindsets, often in shifting and contradictory ways. Our study extends this work by connecting beliefs about decoloniality with critical understandings of languages.

A weaker positive correlation also existed between decoloniality and translanguaging practices themselves. López (2017) notes, "Conceptually there are numerous contributions regarding decolonization and its emancipatory potential in multicultural and multilingual contexts. Notwithstanding, the field is practically virgin regarding recommendations for action and particularly concerning the school and the classroom" (p. 302). Further research is

required to examine how teachers may or may not enact their beliefs about decoloniality through their classroom practices.

4.3 *There is no significant correlation between teachers' AI attitudes and their translanguaging, multilingualism beliefs and technology practices*

No significant correlation was found between international school teachers' attitudes toward AI and any of the other category scores in beliefs and practices of translanguaging, multilingualism, decoloniality and technology practice, which may reflect the diverse and complex nature of teachers' perceptions of AI.

Teachers who hold positive attitudes toward AI in these categories may also be more open to transformative theoretical approaches, such as translanguaging, multilingualism, and decoloniality, suggesting that openness to innovative theory may align with openness to emerging technologies. Conversely, teachers who had negative attitudes toward AI could also have positive attitudes toward translanguaging and multilingual beliefs. This phenomenon could be explained through a decolonial perspective: when AI is perceived as a colonial artifact that can perpetuate colonial power structures (Mohamed et al., 2020; Godwin-Jones, 2025). Teachers' attitude toward AI may be negative even when they are positive about multilingualism and translanguaging.

Moreover, teachers' AI attitudes were not related to their technology practice in general. While AI attitudes did not correlate with translanguaging and multilingual beliefs, there was a significant positive correlation between general technology practice use and translanguaging practices. This may be because technology practice was broadly interpreted to include everyday classroom tools like computers and iPads, which often involve practical functions such as translation apps to address language barriers. Even though such tools involve AI (e.g., Google Translate), teachers may not associate them with the more colloquial notion of AI often represented by chatbots or autonomous systems. Teachers exhibited concerns about using AI in the classrooms due to lack of AI education for students and teachers. This suggests that there is a need for more AI education and research about AI in international school contexts to better support multilingual students.

4.4 *Limitations and Future Work*

This survey is a start to understand what might be the areas of support for international teachers of multilingual learners to encourage translanguaging in their classrooms. Individual interviews may be necessary to understand these findings in more depth. International schools are not homogenous - future studies will need to recognize the contextual differences of different schools such as the country they are in, their educational policies and curriculum, the student body composition etc. as well as the contextual differences within the classroom and the teacher - the number of years they have worked, their language background, their classroom strength and subject. Our quantitative analysis attempts to identify higher-level patterns across survey respondents to get a sense of trends that may be common across these different contexts. It would be necessary to do further qualitative analysis of open-ended responses and analyze responses to individual survey items to understand these trends better.

Additionally, as noted previously, for this pilot study of the international survey, respondents were primarily recruited through the MLRC mailing list and social media accounts. It may be likely that individuals who have chosen to be in contact with the MLRC, whether as members or not, may have been more exposed to translanguaging theories and practices than the general international school teacher population. Hence, our results may not be generalizable for international schools more broadly. In future studies, we may employ more diverse recruitment strategies to explore whether this correlation between translanguaging beliefs and practices is reflective of the general international school teacher population or whether it was more a reflection of particular beliefs and practices of the people who received the survey in the first place.

References

Alharbi, W. AI in the foreign language classroom: A pedagogical overview of automated writing assistance tools. Education Research International, 4253331. (2023). https://doi.org/10.1155/2023/4253331

Bettney Heidt, E. (2024). 6 Translanguaging Pedagogies as a Tool to Decolonize Multilingual Education in International Schools. In R. Zaidi, U. Boz, & E. Moreau (Eds.), Transcultural Pedagogies for Multilingual Classrooms: Responding to Changing Realities in Theory and Practice (pp. 114–128). Multilingual Matters. https://www.degruyterbrill.com/document/doi/10.21832/9781800414419-010/html?lang=en

Bettney, E., & Nordmeyer, J. (2021). Glocal network shifts: Exploring language policies and practices in international schools. Global Education Review, 8, 116-137.

Bettney Heidt, E., & Nordmeyer, J. (2023). Shifting Linguistic Landscapes Within the Context of International Schools: In M. R. Barker, R. C. Hansen, & L. Hammer (Eds.), Advances in Educational Marketing, Administration, and Leadership (pp. 580–604). IGI Global. https://doi.org/10.4018/978-1-6684-8795-2.ch023

Cenoz, J. (2017). Translanguaging in School Contexts: International Perspectives. Journal of Language, Identity & Education, 16(4), 193–198. https://doi.org/10.1080/15348458.2017.1327816

Crisfield, E. (2023). Orientations to linguistic diversity in international schools: Setting a new agenda. RIPE Research Journal. 9,11-23.

Dockrell, J. E., Papadopoulos, T. C., Mifsud, C. L., Bourke, L., Vilageliu, O., Bešić, E., Seifert, S., Gasteiger-Klicpera, B., Ralli, A., Dimakos, I., Karpava, S., Martins, M., Sousa, O., Castro, S., Søndergaard Knudsen, H. B., Donau, P., Haznedar, B., Mikulajová, M., & Gerdzhikova, N. (2022). Teaching and learning in a multilingual Europe: Findings from a cross-European study. European Journal of Psychology of Education, 37(2), 293–320. https://doi.org/10.1007/s10212-020-00523-z

Donley, K. (2024). Teaching with ChatGPT as a linguistically responsive tool for multilingual learners. Technology in Language Teaching & Learning, 6(3), 1719-1719.

Farjon, D., Smits, A., & Voogt, J. (2019). Technology integration of pre-service teachers explained by attitudes and beliefs, competency, access, and experience. Computers & Education, 130, 81–93. https://doi.org/10.1016/j.compedu.2018.11.010

García, O., & Wei, L. (2014). Translanguaging. Palgrave Macmillan UK. https://doi.org/10.1057/9781137385765

Godwin-Jones, R. (2025). Technology integration for less commonly taught languages: AI and pedagogical translanguaging.

Leonet, O., Cenoz, J., & Gorter, D. (2020). Developing morphological awareness across languages: Translanguaging pedagogies in third language acquisition. Language Awareness, 29(1), 41–59. https://doi.org/10.1080/09658416.2019.1688338

López, L., & Sichra, I. (2017). Indigenous Bilingual Education in Latin America. Bilingual and Multilingual Education, Encyclopaedia of Language and Education. Springer.

Marsh, V. L. (2018). Best practices for educating English language learners: History, controversy, and a path forward. Research Brief: Center for Urban Education Success. https://www.rochester.edu/warner/cues/wp-content/uploads/2023/12/ELLS-brief_FINAL-with-header_proofed-TD.pdf

Mohamed, S., Png, M. T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. Philosophy & Technology, 33, 659-684.

Otheguy, R., García, O., & Reid, W. (2015). Clarifying translanguaging and deconstructing named languages: A perspective from linguistics. Applied Linguistics Review, 6(3), 281–307. https://doi.org/10.1515/applirev-2015-0014

Samperio Sanchez, N. (2017). Discovering Students' Preference for Classroom Activities and Teachers' Frequency of Activity Use. Colombian Applied Linguistics Journal, 19(1), 51. https://doi.org/10.14483/calj.v19n1.9292

Tanu, D. (2018). Growing up in transit: The politics of belonging at an international school. Berghahn Books.

Ticheloven, A., Blom, E., Leseman, P., & McMonagle, S. (2021). Translanguaging challenges in multilingual classrooms: Scholar, teacher and student perspectives. International Journal of Multilingualism, 18(3), 491–514. https://doi.org/10.1080/14790718.2019.1686002

Vogel, S., & García, O. (2017). Translanguaging. In S. Vogel & O. García, Oxford Research Encyclopedia of Education. Oxford University Press. https://doi.org/10.1093/acrefore/9780190264093.013.181

Yang, W., & Lin, C. (2025). Translanguaging with generative AI in EFL writing: Students' practices and perceptions. Journal of Second Language Writing, 67, 101181

Bridging Computational Thinking skill gaps with ASSET and Mindspark CT

Ashwini CHANDRASHEKHAR[1], and Aditya Vikram SINGHANIA[2]*
[1,2]Educational Initiatives Private Limited, Bengaluru, Karnataka, India
ashwini.chandrashekhar@ei.study, aditya.singhania@ei.study

Abstract: This paper analyses findings from the ASSET (Assessment of Scholastic Skills through Educational Testing) Computational Thinking (CT) extension, drawing on data from 8,994 Indian students (Grades 3–10). Despite recent reforms advocating CT in curricula, test scores mostly remain between 30% and 50% across four core domains—logical reasoning, pattern recognition, algorithmic thinking, and data interpretation. We spotlight four representative misconceptions that underscore the need for explicit, domain-focused instruction in Indian K–12. Furthermore, we introduce Mindspark CT, an interactive, theme-based program by Educational Initiatives that uses scaffolded puzzles, coding projects, and real-world data analysis to transform students from technology consumers into creators.

Keywords: computational thinking, skill gaps, diagnostic test, ASSET CT, Mindspark CT

1. Introduction

Computational Thinking (CT) has emerged as a critical 21st-century skill, regarded as fundamental for all (Wing, 2006). Many countries have integrated CT into K–12 curricula; for instance, India's National Education Policy 2020 calls for increased emphasis on mathematical and computational thinking through puzzles and games (NEP, 2020).
However, despite this growing recognition, there is a lack of systematic assessments to gauge students' CT skills at scale. Most CT education efforts focus on coding activities or small-scale evaluations, with few large-scale diagnostics outside of contests like Bebras. As a result, educators have limited data on which CT concepts students grasp or struggle with most. Addressing this gap requires structured assessment tools capable of pinpointing specific student skill deficiencies.
Problem Statement: This study addresses the need for systematic CT assessments by employing ASSET CT, a diagnostic test covering four computational thinking skills—Logical Thinking, Pattern Recognition, Algorithmic Thinking, and Data Analysis—to uncover prevalent skill gaps and student misconceptions among Indian learners.

The research questions we intend to answer are:

a) *Which computational thinking concepts and skills do Grades 3–10 students struggle with the most, as evidenced by assessment outcomes?* and
b) *What common misconceptions emerge from students' responses?*

2. Computational Thinking Assessment Framework (ASSET CT)

ASSET CT was developed as part of ASSET (Assessment of Scholastic Skills through Educational Testing), a diagnostic tool designed by Educational Initiatives to evaluate fundamental academic skills systematically. The ASSET diagnostic tests in Math, English and Science have been taken by over 2 million children since 2005.

2.1 Reliability and Validity of ASSET

ASSET employs single-administration methods—specifically Cronbach's alpha—to ensure internal consistency, with alpha values consistently falling within acceptable reliability ranges of 0.7–0.95 (Nunnally, 1978). ASSET utilizes Item Response Theory (IRT), particularly the 2-Parameter Logistic (2PL) model, to ensure items effectively measure intended constructs. Difficulty

parameters beyond ±4, inadequate item discrimination, or poor point-biserial correlations trigger careful review or removal of items.

Although no exact parallel to ASSET exists in India, ASSET scores correlate significantly with international benchmarks. For example, Dubai schools' ASSET scores showed substantial correlations with PISA 2022 results (Reading: 0.81, Math: 0.74, Science: 0.41), underscoring ASSET's validity as a robust diagnostic tool. However, the CT extension of ASSET remains a work-in-progress. Future validation steps, including extensive IRT analyses and continued refinement of CT-specific items, will further enhance its diagnostic precision.

2.2 CT Assessment Framework (ASSET CT)

The design of ASSET CT was informed by key CT frameworks in the literature – notably the programming-focused concepts of Brennan & Resnick (2012), the broader CT facets of Shute et al. (2017), and the K–12 CT practices identified by Weintrop et al. (2016). Our framework defines CT in terms of four main skill categories (**Logical Thinking**, **Pattern Recognition**, **Algorithmic Thinking**, and **Data Analysis**), each subdivided into specific subskills. Based on these existing frameworks, we identified seven core CT sub skills for assessment, spanning traditional programming concepts (e.g. sequencing steps, using conditionals) as well as general reasoning abilities (logical inference, spatial reasoning, etc.).

Each subskill is mapped to an accepted dimension in CT – for example, Brennan & Resnick's coding concepts (sequence, loops, conditionals) , Shute et al.'s six CT facets (**decomposition**, **abstraction**, **algorithm design**, **debugging**, **iteration**, **generalization**), and Weintrop et al.'s emphasis on data and problem-solving practices. By covering a range of subskills, our assessment captures both computational concepts (like sequences and conditionals) and practices (like logical reasoning and pattern identification), consistent with component-based CT assessments in literature (Wiebe et al., 2019). Notably, we include Spatial Reasoning as a subskill given evidence that spatial ability supports CT development (Román-González et al., 2017), even though it is not always explicit in CT curricula.

Table 1. Computational Thinking Skills Taxonomy

ASSET CT Main Skill	ASSET CT Subskill	Brennan & Resnick (2012)	Shute et al. (2017)	Weintrop et al. (2016)
Logical Thinking	Logical Conclusions	Testing & Debugging	Generalization	Problem-Solving Practices
	Problem Solving	Algorithmic Thinking	Decomposition, Debugging	Problem-Solving Practices
Pattern Recognition	Identifying Sequences	Sequences, Patterns	Abstraction	Data Practices
	Spatial Reasoning	Representation	Abstraction	Modelling Practices
Algorithmic Thinking	Sequencing Commands	Sequences	Algorithm Design	Programming Practices
	Conditional Logic	Conditionals, Debugging	Algorithm Design	Programming Practices
Data Analysis	Data Interpretation	Data Representation	Abstraction, Analysis	Data Practices

2.3 Test Structure

Using the above framework, we developed a set of multiple-choice questions targeting these subskills. The test instrument underwent expert review to ensure content validity for each skill area. ASSET CT was then administered in Summer 2024 as a paid diagnostic test to schools across India in Grades 3–10. 8,994 students from 38 schools participated, providing a large dataset of

responses. Each student's test was scored by subskill, enabling analysis of proficiency and misconceptions per skill. Below, we present the findings in relation to our research questions.

2.4 Participants

Each grade-level ASSET CT test (Levels 1–4) included 30 multiple-choice items—except Grades 3–4, which had 25. As 115 unique questions repeated across grade-bands, the dataset recorded 230 total "item records." Items emphasized conceptual proficiency (puzzles, stepwise logic, multi-field data) rather than code syntax. To measure item quality, we computed point-biserial correlation (PBC)—higher PBC (>+0.30) indicates stronger discrimination between high- and low-performers (Frisbie, David A. 1988).

3. Skill Gaps and Illustrative Misconceptions

3.1. Grade-wise scores

Table 2 shows the average percentage score, standard deviation, and highest marks per grade. While Grade 10 achieved 52.3% on average—higher than younger grades— scores typically remain in the 30–50% band across grades, confirming widespread CT skill gaps.

Table 2. ASSET CT Performance by Grade (N=8,994)

Grade	Paper Code	Students	Avg. Score	Std. Dev.	Highest
3	73124	1,241	37.8%	17.4%	25
4	74124	1,129	42.7%	18.4%	24
5	75124	1,154	36.8%	14.1%	30
6	76124	1,211	41.2%	16.0%	28
7	77124	1,055	36.2%	14.5%	29
8	78124	1,308	45.1%	17.5%	28
9	79124	1,032	46.5%	17.9%	29
10	7A124	864	52.3%	18.5%	30

While Grades 5 and 10 had some top scorers achieving 100%, these were exceptions. High standard deviations (up to ~18.5%) reflect wide performance variability, indicating that even within the same class, certain learners can excel, whereas many others demonstrate foundational gaps in multi-step reasoning or data interpretation.

3.2. Skill gaps and misconceptions

The ASSET-CT results reveal significant and persistent gaps across the four core computational thinking skills: Logical Thinking, Pattern Recognition, Algorithmic Thinking, and Data Analysis, and their respective subskills. While simple skills showed some improvement across grades, complex, multi-step tasks exposed consistent difficulties. Below, we discuss the results skill by skill, grounding each finding in specific performance data.

3.2.1 Logical Thinking (problem solving and drawing logical conclusions)

In a Grade 3/4 task, students were tasked to determine the total number of people in a line based on dual ranking clues (5th from left, 15th from right). Fewer than 20% answered correctly; even in Grade 8, the correct rate was only about 28%, with many incorrectly adding the ranks without adjusting for overlap. Similarly, in a Grade 7/8 task involving prediction of bucket motion based on gear connections (direct, open belt, cross belt), a large proportion of students selected wrong answers, indicating challenges in tracing multi-step mechanical causality. These patterns reveal persistent gaps in systematically applying multiple logical rules to solve complex problems.

3.2.2 Pattern Recognition (identifying patterns in numbers, shapes and language)

In a Grade 5 question based on Gauss's summation strategy, students were shown the method for adding numbers from 1 to 100 and asked to apply it for 1 to 50. Only about 32% answered correctly. Common mistakes included incorrect calculation of the number of pairs or errors in summing them, suggesting that students struggled to abstract and transfer the demonstrated strategy to a new situation. Similarly, in visual pattern continuation tasks at Grades 3/4, around 30–40% selected the correct next figure, with many wrong choices reflecting focus on surface features rather than understanding the generative rule.

3.2.3 Algorithmic Thinking (sequencing commands and conditional logic)

In a Grade 7/8 problem, students were tasked to guide a robot mouse to collect nuts in a 50×50 grid, issuing forward, turn, and repeat commands. When asked to compute the total number of commands required if the repeat feature was disabled, approximately 30% of students selected the correct answer. Many students missed counting turns as separate commands or miscounted the exact number of cells to traverse, indicating difficulties in breaking down repetitive tasks into explicit step-by-step instructions. Similarly, in a Grade 9 task involving systematic swaps between cards indexed by positions (swap i with i+3), fewer than 32% answered correctly. Students struggled to predict the outcome of iterative step-by-step execution, suggesting challenges in mentally tracing algorithms across multiple operations.

3.2.4 Data Analysis (representation and interpretation)

Students showed modest successes in simple data retrieval but struggled with relational and multi-condition reasoning. In a historical reasoning task, only 38.5% of students could deduce the correct battle date by integrating multiple timeline clues. In a coin-combination puzzle, fewer than 25% identified the correct grouping, revealing difficulty with exhaustive case analysis and elimination strategies. In a GDP graph interpretation, around 30% correctly isolated Germany's peak year, with many students distracted by overall graph height instead of focusing on the specific country's trend. These findings suggest that while students can read basic data points, they falter when deeper comparison, elimination, or cross-referencing is required.

3.3 Performance across skills

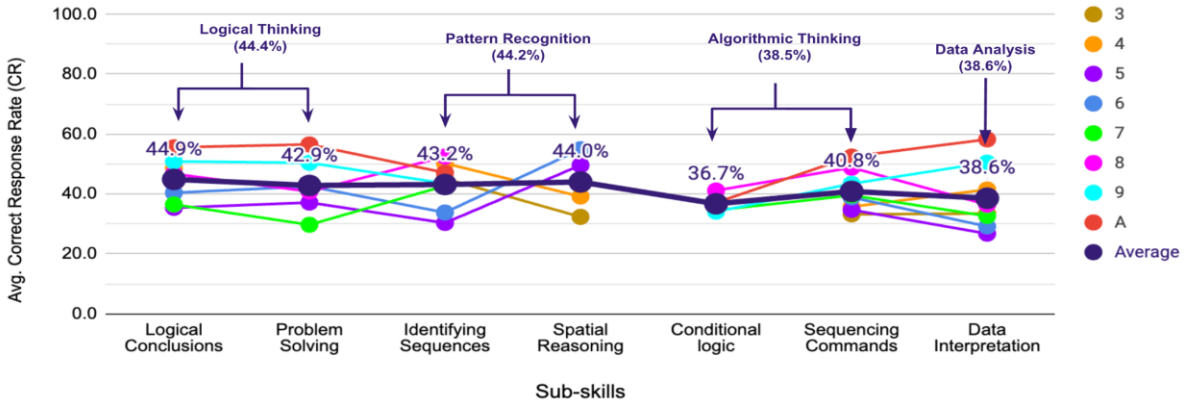


Figure 1. Skill-wise performance across grades 3 to 10. Some grades lack data for certain subskills, e.g., grade 3/4 was not tested on conditional logic problems; grade 9/10 was not tested on spatial reasoning.

Algorithmic Thinking (averaging ~38–39%) and Data Analysis (38.6%) remain the largest gaps overall, whereas Logical Thinking (~43–44%) and Pattern Recognition (~43–44%) perform slightly better. Grade 10 (recorded as 'A') tends to outscore younger grades in most subskills, reinforcing a slow upward trend, but subskills like Conditional Logic remain low across all cohorts. Such tasks require carefully tracing or formulating step-by-step procedures—skills typically not emphasized in standard curricula (Yadav, Hong, & Stephenson, 2016).

4. Future Work: Mindspark CT Intervention and iterations of ASSET CT

To remediate these gaps, a targeted adaptive learning platform called Mindspark CT has been initiated, integrating diagnostic insights directly into instruction. The Mindspark CT pilot provides structured exercises explicitly addressing misconceptions identified by ASSET CT, such as logical reasoning puzzles and conditional logic scenarios. Moreover, detailed item-level validation through advanced IRT analyses is planned for ASSET CT, leveraging now sufficient data from thousands of student responses. This rigorous validation will refine ASSET CT's capability to pinpoint precise skill gaps further and guide targeted instructional intervention.

5. Conclusion

This study demonstrates the value of systematic assessment of computational thinking in the K–12 context. The findings reveal that many students, even up to grade 10, have not mastered core CT skills such as Data Analysis, Pattern Recognition, and Algorithmic Thinking. These gaps are not unique to a single region but reflect broader challenges in CT education – however, our work provides much-needed data from an Indian and global south context, balancing a literature often focused on western settings.

Although limited by its multiple-choice format and preliminary nature, our assessment establishes a valuable baseline for understanding CT competencies across diverse student groups. Future efforts should focus on iterative refinement of both assessment tools and instructional methods, ensuring that educational practices concretely develop strong computational thinking skills for navigating the challenges and opportunities of the 21st century.

References

Barr, V., & Stephenson, C. (2011). Bringing computational thinking to K–12. ACM Inroads, 2(1), 48–54.

Brennan, K., & Resnick, M. (2012). New frameworks for studying and assessing the development of computational thinking. Proceedings of the 2012 Annual Meeting of the American Educational Research Association.

Denning, P. J. (2017). Remaining trouble spots with computational thinking. Communications of the ACM, 60(6), 33–39.

Frisbie, David A. (1988). Reliability of scores from teacher-made tests. Educational Measurement: Issues and Practice, 7(1), 25–35.

Government of India, Ministry of Education. (2020). National Education Policy 2020.

Román-González, M., Pérez-González, J.-C., & Jiménez-Fernández, C. (2017). Which cognitive abilities underlie computational thinking? Computers in Human Behavior, 72, 678–691.

Shute, V. J., Sun, C., & Asbell-Clarke, J. (2017). Demystifying computational thinking. Educational Research Review, 22, 142–158.

Weintrop, D., Beheshti, E., Horn, M., Orton, K., Jona, K., Trouille, L., & Wilensky, U. (2016). Defining computational thinking for mathematics and science classrooms. Journal of Science Education and Technology, 25(1), 127–147.

Wing, J. M. (2006). Computational thinking. Communications of the ACM, 49(3), 33–35.

Yadav, A., Hong, H., & Stephenson, C. (2016). Computational thinking for all. TechTrends, 60(6), 565–568.

Educational Cone Model in Embedding Vector Spaces

Yo EHARA[a]
[a]*Faculty of Education, Tokyo Gakugei University, Japan*
*ehara@u-gakugei.ac.jp

Abstract: Human-annotated datasets with explicit difficulty ratings are essential in intelligent educational systems. Although embedding vector spaces are widely used to represent semantic closeness and are promising for analyzing text difficulty, the abundance of embedding methods creates a challenge in selecting the most suitable method. This study proposes the Educational Cone Model, which is a geometric framework based on the assumption that easier texts are less diverse (focusing on fundamental concepts), whereas harder texts are more diverse. This assumption leads to a cone-shaped distribution in the embedding space regardless of the embedding method used. The model frames the evaluation of embeddings as an optimization problem with the aim of detecting structured difficulty-based patterns. By designing specific loss functions, efficient closed-form solutions are derived that avoid costly computation. Empirical tests on real-world datasets validated the model's effectiveness and speed in identifying the embedding spaces that are best aligned with difficulty-annotated educational texts.

Keywords: Embedding, difficulty, vector spaces

1. Introduction

Datasets that are annotated with difficulty levels by educators ("difficulty-annotated educational datasets") are essential for developing educational support systems (Arase et al., 2022; Hendrycks et al., 2020). Although embedding spaces are widely used to represent semantic similarity and are promising for analyzing such datasets, the abundance of embedding methods (Muennighoff et al., 2022) poses a challenge in selecting suitable methods.

To address this, we propose the Educational Cone Model, which assumes that easier items covering fundamental concepts exhibit lower diversity, while more difficult items are more diverse. This results in a cone-like structure in the embedding space, independent of specific methods. This intuition aligns with the findings of vocabulary acquisition (e.g., Zipf's law), Piaget's developmental stages (Piaget, 1952), and Bloom's taxonomy.
We mathematically show that evaluating alignment with this model is reduced to solving an optimization problem that identifies a "difficulty direction" in the embedding space. By designing appropriate loss functions, we derive closed-form solutions, avoiding computationally expensive operations such as centroid comparisons.

Empirical evaluations with recent sentence embeddings confirm that the proposed model enables efficient selection of embedding methods that are well-aligned with difficulty-annotated datasets.

The contributions of this study are summarized as follows: 1) We propose a geometric model that reflects the assumption that easier items are less diverse in the embedding space. 2) We show that the model can identify a difficulty direction via optimization. 3) We formulate this as a closed-form optimization problem. 4) We demonstrate that this solution requires only mean vector differences between difficulty levels. 5) We empirically validate the effectiveness of the model on real datasets using recent sentence embeddings.

2. Formulation

2.1 Notation

Let $\{x_1, \ldots, x_N\}$ denote a set of N embedding vectors, where $\boldsymbol{x}_i$ is a D-dimensional vector. We assume that all embedding vectors are normalized; that is, $||\mathbf{x}|| = 1$, where $||\mathbf{x}||$ denotes the Euclidean norm of the vector. The proposed method can be applied to both word and sentence embeddings if the aforementioned conditions are satisfied. For simplicity, we refer to both words and sentences as items throughout this paper. We introduce a D-dimensional vector $\mathbf{w} \in \mathbf{R}^D$ to represent the direction in the embedding space, with the aim of determining the coordinates of $\mathbf{w}$.

2.2 Educational Cone Model and Difficulty Direction Search Problem

We first consider the Educational Cone Model and show that, under simple assumptions, it aligns with the problem of searching for the difficulty direction in the embedding space. The Educational Cone Model assumes that simpler items exhibit lower diversity and more difficult items exhibit higher diversity. We interpret the magnitude of diversity in terms of the spatial spread within the embedding space. Assuming that simpler items exhibit lower diversity, their embedding vectors have a smaller spread in the embedding space. By further refining the notion that simpler items exhibit lower diversity, we assume the existence of the "simplest item." Embedding vector spaces are typically structured such that semantically similar items are positioned closer together. Hence, in the embedding vector space, the simplest item is assumed to reside in the least-spread-out region, represented by point $\boldsymbol{e}$. If we consider difficulty as a component of "semantic similarity," items closer to $\boldsymbol{e}$ should be simpler, whereas those farther away should be more difficult. In the Educational Cone Model, suppose that $\boldsymbol{x}_i$ is simpler than $\boldsymbol{x}_j$. Based on the above discussion, $\boldsymbol{x}_i$ is closer to the "simplest item" $\boldsymbol{e}$ than $\boldsymbol{x}_j$. By measuring the distance using the Euclidean distance and assuming that all x are normalized to $||\mathbf{x}|| = 1$, we obtain the following transformations: $||\boldsymbol{x}_i - \boldsymbol{e}|| < ||\boldsymbol{x}_j - \boldsymbol{e}|| \Leftrightarrow \boldsymbol{w}^\top \boldsymbol{x}_i < \boldsymbol{w}^\top \boldsymbol{x}_j$. Here, we define $\boldsymbol{w}$ as $-\boldsymbol{e}$. $\boldsymbol{w}$ represents a direction in the embedding vector space. The expression $\boldsymbol{w}^\top \boldsymbol{x}_i$ implies that items can be arranged in order along this direction. Consequently, the direction $\boldsymbol{w}$ indicates that moving in this direction within the embedding vector space corresponds to increasing difficulty.

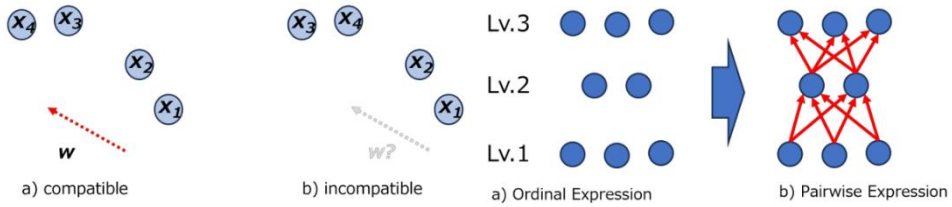


Figure 1. Left: Overview of the proposed method: a) Considering $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3,$ and $\boldsymbol{x}_4$ as two-dimensional (2D) word/sentence embeddings. That is, $\boldsymbol{x}_1$ is annotated as simpler than $\boldsymbol{x}_2$ which, in turn, is simpler than $\boldsymbol{x}_3$, etc. If listing points along direction $\mathbf{w}$ aligns with the annotation, the embedding vector set is defined as "compatible" with the annotation. b) In this case, no direction in the 2D space orders $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3,$ and $\boldsymbol{x}_4$ in the annotated order, so the embedding is defined as "incompatible." Right: Conversion of difficulty annotations into pairwise constraints.

We provide an intuitive interpretation of the difficulty direction $\boldsymbol{w}$. Simply put, the difficulty direction $\boldsymbol{w}$ represents the direction in which all items in the dataset (annotation set) appear, arranged in order of difficulty. Although determining an ideal direction is preferable, it is often unrealistic. To address this, we allow slight deviations in the order of difficulty. To this end, we first introduce the concept of "compatibility" in Figure 1 (left). In practice, difficulty annotations are often provided in ordinal levels, rather than as direct pairwise relationships. For example,

a question might be annotated as high-school or university level rather than being directly compared to another question (Figure 1, right). As shown in (a), eight items are annotated using three levels: Levels 1, 2, and 3. These levels, abbreviated as "Lv," indicate increasing difficulty with higher numbers. Levels 1, 2, and 3 have three, two, and three items, respectively. The entire ordinal structure can be converted into a directed graph as shown in (b). The directed edges signify pairwise relationships, and an edge from node *i* to node *j* indicates that *i* is easier than *j*. In this manner, without loss of generality, any finite set of ordinal annotations can be converted into a mathematically equivalent directed graph.

As illustrated in Figure 1, the difficulty annotations can be converted into a set of pairwise constraints. To ensure generality, we define a set of pairwise comparison constraints, in which each constraint indicates that one embedding vector is annotated more easily than the other. We define the set of pairwise order constraints as $C = \{(i_1, j_1), \ldots, (i_K, j_K)\}$, where K denotes the number of constraints. The k-th pair (i_k, j_k) represents a single constraint.
Here, $i_k \in \{1, \ldots, N\}$ and $j_k \in \{1, \ldots, N\}$ denote indices of N embedding vectors, each of which represents the meaning of each item. The annotation (i_k, j_k) indicates that x_{ik} is easier than x_{jk}. For simplicity, we omit the subscript k and denote the easier vector as x_i and the more difficult vector as x_j. We then project these vectors onto the direction defined by $\boldsymbol{w}$ to model the ordering. Let θ_i and θ_j denote the angle between x_i and the direction $\boldsymbol{w}$ and the angle between x_j and $\boldsymbol{w}$, respectively. Our goal is to adjust $\boldsymbol{w}$ such that the pairwise constraints are satisfied: $||\boldsymbol{x}_i|| \cos\theta_i < ||\boldsymbol{x}_j|| \cos\theta_j \Leftrightarrow ||\boldsymbol{w}||||\boldsymbol{x}_i|| \cos\theta_i < ||\boldsymbol{w}||||\boldsymbol{x}_j|| \cos\theta_j \Leftrightarrow \boldsymbol{w}^\top\boldsymbol{x}_i < \boldsymbol{w}^\top\boldsymbol{x}_j \Leftrightarrow \boldsymbol{w}^\top(\boldsymbol{x}_i - \boldsymbol{x}_j) < 0$. If K pairwise constraints exist, they must all hold simultaneously. However, for practical datasets with many items, there may not exist a $\boldsymbol{w}$ that meets all K constraints. To this end, we introduce a slack variable ξ_k to rewrite the inequality constraint into an equality constraint: $\mathbf{w}\top(\boldsymbol{x}_{ik} - \boldsymbol{x}_{jk}) + \xi_k = 0$. Intuitively, ξ_k represents the degree to which the constraint is maintained. A larger value of ξ_k means that a greater margin is preserved, whereas a smaller value indicates a minimal margin enforcement. While $\xi_k \geq 0$, we remove this constraint to relax the problem:

$$\text{maximize}_{\xi,\boldsymbol{w}} \sum_{k=1}^{K} \xi_k \quad \text{s.t.} \quad \forall k \in \{1, \ldots, K\}; \boldsymbol{w}\top\left(\boldsymbol{x}_{ik} - \boldsymbol{x}_{jk}\right) + \xi_k = 0, \left|\left|\boldsymbol{w}\right|\right|^2 = 1. \quad (1)$$

In Equation (1), we also impose a norm constraint $||\boldsymbol{w}||^2=1$ to obtain a fixed-form solution. First, by noting that $\xi_k = -\boldsymbol{w}\top(\boldsymbol{x}_{ik} - \boldsymbol{x}_{jk})$ and introducing a Lagrange multiplier λ to enforce the equality constraint $1 - ||\mathbf{w}||^2$, we obtain the following unconstrained optimization problem:

$$\text{maximize}_{\,\boldsymbol{w}} \sum_{k=1}^{K} -\boldsymbol{w}\top\left(\boldsymbol{x}_{ik} - \boldsymbol{x}_{jk}\right) + \lambda\left(1 - \left|\left|\boldsymbol{w}\right|\right|^2\right). \quad (2)$$

Differentiating Equation (2) with respect to $\boldsymbol{w}$, we obtain $\sum_{k=1}^{K} -\left(\boldsymbol{x}_{ik} - \boldsymbol{x}_{jk}\right) - 2\lambda\boldsymbol{w} = 0$, from which we obtain $\boldsymbol{w} \propto \sum_{k=1}^{K}\left(\boldsymbol{x}_{jk} - \boldsymbol{x}_{ik}\right)$. This confirms that the optimal direction is simply the mean of the difference vectors normalized to the unit length. This property enables us to determine the optimal vector efficiently without solving the optimization problem each time.

3. Experiments

3.1 Consistency Experiment Using Word Embeddings and Fine-Grained Annotations

Consistency experiments were conducted using word embeddings. The dataset used was the CEFR-J Vocabulary Profile, which contains manually annotated word difficulty levels based on the Common European Framework of Reference for Languages (CEFR) (https://github.com/openlanguageprofiles/olp-en-cefrj?tab=readme-ov-file). FastText (Bhattacharjee, 2018) was used as the word embedding model. We employed the Support Vector Machine (SVM) as the baseline method. The regularization parameter C was tuned using the validation data, and the optimal value was selected from {0.1, 1.0, 10.0}. To verify the consistency, we followed a procedure that categorizes words into four levels of difficulty. A total of 100 words were randomly selected from the second-easiest category. For each word, pairwise constraints were formulated by treating pairs in the easiest category as training data

and pairs in the third-easiest category as test data. Thus, the proposed convex optimization problem was solved. Both the proposed method and SVM achieved 100% accuracy.

In this approach, each word can be regarded as containing 100 different subset datasets. Another dataset, SVL, provides a more fine-grained 12-level annotation of word difficulty (https://eow.alc.co.jp/svl_level12.html). Using these 100 datasets, we assessed the annotation consistency based on the objective function value of the proposed convex optimization problem. As the annotation consistency is measured with respect to the simplest word category, words with higher difficulty levels in the 12-level SVL dataset were expected to exhibit greater consistency. Spearman's rank correlation analysis confirmed this expectation and yielded a statistically significant correlation ($p < 0.01$). These results demonstrate that the proposed method effectively detects consistency in word annotations.

Table 1. *Compatibility scores calculated using training data. "Model" represents each individual model, and each column corresponds to a pair of CEFR levels. The compatibility score indicates the degree of fit between the level annotations and each embedding. The first two models are 384-dimensional and the latter two models are 1,024-dimensional.*

Model	(A1, A2)	(A1, B1)	(A1,B2)	(A2,B1)	(A2,B2)	(B1,B2)
all-MiniLM-L6-v2	0.3227	0.5741	0.9308	0.4183	0.8599	0.7986
multilingual-e5-small	0.0835	0.1456	0.2301	0.0835	0.1961	0.1782
bge-m3	0.1354	0.2317	0.3930	0.1410	0.3516	0.3336
multilingual-e5-large	0.0826	0.1435	0.2256	0.0844	0.1952	0.1756

3.2 Consistency Experiment Using Sentence Embeddings

The proposed method is applicable to both word and sentence embeddings. To examine its utility further, we evaluated its ability to predict annotation inconsistencies in sentence embeddings. For this experiment, we used the CEFR-SP dataset (Arase et al., 2022), which consists of English sentences annotated with four difficulty levels by two annotators. Following the same approach as that in the word-level experiment, we applied the proposed convex optimization problem to the dataset of one annotator. Sentence embeddings were generated using 'multilingual-e5-small' 3. Owing to the norm constraint in the proposed convex optimization, the solution vector lengths were approximately 1.0, enabling a direct comparison of the objective function values across sentences. As the value of ξ can be interpreted as a margin, a higher value indicates greater consistency. Next, we compared the results with those of the second annotator and calculated the Spearman's rank correlation coefficient between the annotation agreement and objective function value. The correlation coefficient was 0.262, indicating a weak correlation but without statistical significance. This suggests that although the proposed method captures annotation inconsistencies to a certain extent, further investigation is required.

Using the CEFR-SP dataset, we conducted experiments to identify the most compatible sentence embedding model. Each difficulty level pair was treated as a separate dataset and the compatibility score was computed using the proposed method. Table 1 presents the compatibility scores obtained from the sentence embedding experiments. Each column corresponds to a CEFR difficulty level pair, and Table 1 summarizes the embedding models used. Higher scores indicate a better fit between the embedding space.

3.3 Prediction Experiment Using SVM

Finally, we conducted a prediction experiment using SVM with sentence embeddings. We utilized (A1, B1) as the training datasets and (B1, B2) as the test datasets. The four embedding models summarized in Table 1 achieved prediction accuracies of 0.26, 0.23, 0.63, and 0.37, respectively. Notably, the first two embeddings had 384 dimensions, whereas the latter two had 1,024 dimensions. We observed that when the dimensionality was the same, embeddings with higher compatibility scores tended to yield superior predictive performance. This finding suggests that, given the same dimensionality, the compatibility score can reliably estimate the predictive performance of SVM without requiring access to the test data. Consequently, our approach enables an efficient assessment of the embedding quality for text difficulty prediction, without the need for computationally expensive model training on each embedding.

4. Related Work

As an early study on the relationship between contextualized embedding vectors and difficulty, Ehara (2022) expressed the difficulty of word examples by the frequency of nearby examples in the embedding vector space. Recently, Ehara (2025) proposed a method for controlling the difficulty of educational items by combining linear interpolation of sentence embeddings with different difficulty levels in the embedding space and a technique called Inverse Embedding, which generates sentences from embedding vector coordinates. Several recent studies have investigated the interpretability of embedding vectors (Vasilyev et al., 2024; Li & Li, 2024; Chen et al., 2024). Vasilyev et al. (2024) examined multilingual embeddings using linear transformations under the assumption that embeddings are orthogonally aligned across languages. Li and Li (2024) analyzed the relationships between contextualized embeddings from large-scale language models and static embeddings such as sentence embeddings. Chen et al. (2024) proposed methods for constructing finer-grained embeddings for thematic sentence representations. Building on this research, our study focuses on the consistency of ordinal annotations and presents a method for quantifying this consistency.

5. Conclusion

This paper has introduced the Educational Cone Model, which posits that easier items exhibit lower diversity than more difficult items in embedding vector spaces owing to their inherent complexity. This principle aligns with the findings of previous studies on educational classifications. Leveraging this insight, we proposed a computationally efficient method for determining the directional orientation of educational items within an embedding space. As a convex optimization approach, our method guarantees convergence to a global optimum, making it robust to the initial conditions and free from approximation errors. This contrasts sharply with recent neural-network-based methods, which are typically sensitive to initialization and only achieve locally optimal solutions.

The empirical evaluation demonstrated a statistically significant correlation between our method and fine-grained annotation consistency in word difficulty datasets. We further conducted experiments using the proposed method with recent sentence embedding. The proposed objective function accurately predicted the performance of embeddings as predictive features without using test data. This result substantiates our method's ability to capture the "cone" structure of learning material difficulty within an embedding space effectively.

Future research directions include extending the approach to a broader range of educational tasks, including STEM-related questions. Notably, the interpretability of our method allows the identification of subject-specific difficulty dimensions, such as those unique to physics or chemistry, by treating difficulty as a directional property within the embedding space.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 22K12287 and by JST, PRESTO Grant Number JPMJPR2363.

References

Arase, Y., Uchida, S., & Kajiwara, T. (2022). CEFR-based sentence difficulty annotation and assessment. arXiv:2210.11766. https://doi.org/10.48550/arXiv.2210.11766

Bhattacharjee, J. (2018). fastText quick start guide: Get started with Facebook's library for text representation and classification. Packt Publishing.

Chen, S., Zhang, H., Chen, T., Zhou, B., Yu, W., Yu, D., Peng, B., Wang, H., Roth, D., & Yu, D. (2024). Sub-sentence encoder: Contrastive learning of propositional semantic representations. In K. Duh, H. Gomez, & S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1, Long Papers) (pp. 1596-1609). Association for Computational Linguistics.

Ehara, Y. (2022). An Intelligent Interactive Support System for Word Usage Learning in Second Languages. In: Rodrigo, M.M., Matsuda, N., Cristea, A.I., Dimitrova, V. (eds) Artificial Intelligence in Education. AIED 2022. Lecture Notes in Computer Science, vol 13355. Springer, Cham. https://doi.org/10.1007/978-3-031-11644-5_37

Ehara, Y. (2025). Generating Diverse Difficulty Examples in Embedding Vector Spaces via Inverse Embedding. In: Cristea, A.I., Walker, E., Lu, Y., Santos, O.C., Isotani, S. (eds) Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium, Blue Sky, and WideAIED. AIED 2025. Communications in Computer and Information Science, vol 2592. Springer, Cham. https://doi.org/10.1007/978-3-031-99267-4_9

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2020). Measuring massive multitask language understanding. arXiv:2009.03300. https://doi.org/10.48550/arXiv.2009.03300

Li, X., & Li, J. (2024). BeLLM: Backward dependency enhanced large language model for sentence embeddings. In K. Duh, H. Gomez, & S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Vol. 1, Long Papers) (pp. 792-804). Association for Computational Linguistics.

Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2022). MTEB: Massive text embedding benchmark. arXiv:2210.07316. https://doi.org/10.48550/arXiv.2210.07316

Piaget, J. (1952). The origins of intelligence in children (M. Cook, Trans.). International Universities Press.

Vasilyev, O., Isono, F., & Bohannon, J. (2024). Linear cross-lingual mapping of sentence embeddings. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024 (pp. 8163-8171). Association for Computational Linguistics.

Generative AI and Multimodal Profiling in Programming Education: A Systematic Review of Tools, Pedagogical Impacts, and Future Competencies

Aboul Hassane CISSE[a*] & Shoya ISHIMARU[b]
[a,b]*Graduate School of Informatics, Osaka Metropolitan University, Japan*
*sn22869p@st.omu.ac.jp

Abstract: This systematic review investigates the integration of generative artificial intelligence (GenAI) and multimodal learning analytics (MMLA) in programming education, with a focus on empirical implementations and their pedagogical implications. Drawing from 59 peer-reviewed studies published between 2015 and 2025, this review categorizes how GenAI tools such as ChatGPT, GitHub Copilot, and domain-specific tutors are being used to support real-time feedback, code generation, and learner scaffolding. While these tools enhance productivity and interaction, they also raise new challenges around critical thinking, overreliance, and the redefinition of assessment practices. The findings highlight a lack of longitudinal and cross-cultural studies and limited integration with multimodal sensing, such as eye tracking and emotion detection. We propose future directions for designing adaptive, ethically grounded learning environments that leverage GenAI alongside real-time multimodal feedback. This includes fostering AI literacy, ensuring learner agency, and equipping educators with tools for reflective and inclusive AI pedagogy.

Keywords: Generative Artificial Intelligence (GenAI), Programming Education, Multimodal Learning Analytics, Large Language Models (LLMs), AI-Assisted Learning

1. Introduction

Generative Artificial Intelligence (GenAI) is reshaping programming education by enabling real-time code generation, personalized feedback, and natural language explanations through tools such as ChatGPT and GitHub Copilot (Prather et al., 2025; Liu et al., 2024). These technologies promote a shift toward student-centered learning that emphasizes interaction and scaffolding (Zapata et al., 2025; Chan & Li, 2025).

Despite these benefits, concerns have emerged about overreliance, reduced critical thinking, and cognitive offloading (Bauer et al., 2025; Gerlich, 2025). At the same time, research on combining GenAI with multimodal learning analytics (MMLA) remains fragmented. MMLA leverages signals such as eye movements, keystrokes, and physiological responses to provide insights into learners' cognitive and emotional states (Bhatt et al., 2024; Morita et al., 2025). Yet few studies explore how GenAI and MMLA can be jointly used to enhance programming instruction.

This review addresses this gap through three objectives: (a) investigate GenAI-supported instructional practices; (b) categorize multimodal signals used; and (c) analyze their impact on engagement, self-efficacy, and skill acquisition. By consolidating recent findings, the study aims to inform the design of next-generation programming environments that are intelligent, adaptive, and pedagogically grounded.

2. Methodology

This study employed a systematic literature review (SLR) guided by PRISMA 2020 and software engineering guidelines, analyzing 59 empirical studies published between 2015–2025. Five databases (ACM, IEEE Xplore, Scopus, ResearchGate, arXiv) were searched using Boolean queries (e.g., “generative AI” AND “programming education”; “eye-tracking” AND “AI tutoring”). Inclusion criteria required empirical studies in English that used GenAI tools (e.g., ChatGPT, GitHub Copilot) within programming education and provided full-text access, excluding purely conceptual or technical papers. From 127 records, 59 met all criteria after PRISMA-documented screening (see repository: https://github.com/pachaboul/autoreviewx/tree/e774edc57c0e8d108455443292e7a300357eb819/data). Data were extracted with AutoReviewX, coding for GenAI tools, educational level, pedagogical design, modality, outcomes, and methods. Quality assessment (CASP, PICO, TAPUPAS) showed strengths in research clarity and transparency, with weaker reporting on limitations and accessibility.

3. Results

RQ1: How is GenAI integrated in programming education?
GenAI tools are mainly applied in code completion, prompt refinement, and AI tutoring, functioning as cognitive and instructional partners.

- ChatGPT supports explanations and debugging.
- GitHub Copilot generates context-aware multi-line suggestions.
- CodeHelp and CS50 Chatbot provide structured feedback and course-specific guidance (Prather et al., 2025; Chan & Li, 2025).

RQ2: What pedagogical changes does GenAI introduce?
Adoption shifts emphasis from coding to AI collaboration and critical evaluation. Emerging practices include:

- Prompt literacy: teaching students to design effective queries.
- Analytical reasoning: evaluating AI-generated outputs.
- Reflective practice: using GenAI as a tool for metacognitive growth (Chan & Li, 2025; Bauer et al., 2025).

RQ3: How does MMLA enhance GenAI use?
Though still limited, MMLA extends adaptivity by incorporating learner signals:

- Attention/engagement via eye-tracking and keystrokes (Bhatt et al., 2024).
- Affective states via facial, posture, or heart-rate cues (Melo et al., 2025).
- Personalization through fusion of text prompts and multimodal signals (Santhosh et al., 2024).

Challenges remain around privacy, cost, and ethics, but early prototypes suggest promise for hybrid, learner-centered ecosystems.

4. Discussion

This review shows that GenAI tools are becoming central in programming education, primarily supporting code generation, debugging, and tutoring. Their use shifts pedagogy from coding as a manual task toward collaborative engagement with AI, requiring students to evaluate, adapt, and reflect on outputs. Limited but promising studies demonstrate that multimodal extensions (e.g., gaze or emotion signals) can create more adaptive and inclusive learning environments.

Compared with traditional rule-based tutors, GenAI systems enable flexible, conversational learning, but also pose risks of overreliance, logic errors, and reduced critical thinking. This demands stronger metacognitive skills and AI literacy from learners.

Emerging trends highlight:

- Prompt literacy as a new programming skill.
- Multimodal integration to personalize support.
- Hybrid ecosystems combining GenAI with flipped or project-based learning.
- Innovative assessments (e.g., reflective tasks, oral defenses) to evaluate reasoning.

Theoretically, GenAI aligns with socio-constructivist and cognitive apprenticeship models, acting as a cognitive partner within the learner's zone of proximal development. Practically, educators should focus on AI literacy, reflective assessments, and explainability, while developers should prioritize personalization, ethics, and multimodal integration.

5. Conclusion

This review is limited by potential database and publication bias, as only five databases were searched and positive findings may be overrepresented. Most included studies were short-term pilots in well-resourced contexts, limiting generalizability to underrepresented or global settings.

Despite these limitations, the analysis of 59 studies shows that GenAI tools are increasingly integrated into programming education, supporting coding, debugging, and feedback. However, multimodal extensions (e.g., gaze, emotion, physiological signals) remain rare, leaving much of GenAI's adaptive potential unrealized.

Future research should:

- Examine long-term learning impacts beyond task performance.
- Combine GenAI with multimodal signals (eye-tracking, EEG, motion).
- Develop ethical, inclusive, and scalable practices.
- Provide educator training for effective and responsible adoption.

In sum, GenAI holds promise not as a replacement for human effort, but as a partner that supports deeper learning, reflection, and empowerment when combined with multimodal insights and sound pedagogy.

References

Bauer, E., Greiff, S., Graesser, A.C. et al. Looking Beyond the Hype: Understanding the Effects of AI on Learning. Educ Psychol Rev 37, 45 (2025). https://doi.org/10.1007/s10648-025-10020-8

Bhatt, A., Watanabe, K., Dengel, A., & Ishimaru, S. (2024). Appearance-based gaze estimation with deep neural networks: From data collection to evaluation. International Journal of Activity and Behavior Computing, 2024(1), 1-15.

Chan, J., & Li, Y. (2025). Enhancing Higher Education with Generative AI: A Multimodal Approach for Personalised Learning. arXiv preprint arXiv:2502.07401.

Gerlich, M. (2025). AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies*, *15*(1), 6. https://doi.org/10.3390/soc15010006

Melo, J., Fernandez, L., & Ishimaru, S. (2025). Automatic Classification of Difficulty of Texts from Eye Gaze and Physiological Measures of L2 English Speakers. IEEE Access.

Morita, R., Watanabe, K., Zhou, J., Dengel, A., & Ishimaru, S. (2025). GenAIReading: Augmenting human cognition with interactive digital textbooks using large language models and image generation models. *arXiv preprint arXiv:2503.07463*. https://arxiv.org/abs/2503.07463

Prather, J., Leinonen, J., Kiesler, N., Gorson Benario, J., Lau, S., MacNeil, S., ... & Zingaro, D. (2025). Beyond the Hype: A Comprehensive Review of Current Trends in Generative AI Research, Teaching Practices, and Tools. 2024 Working Group Reports on Innovation and Technology in Computer Science Education, 300-338.

Santhosh, J., Dengel, A., & Ishimaru, S. (2024). Gaze-Driven Adaptive Learning System with ChatGPT-Generated Summaries. IEEE Access, 12, 173714-173733.

Zapata, G. C., Cope, B., Kalantzis, M., Tzirides, A. O. (Olnancy), Saini, A. K., Searsmith, D., … Abrantes da Silva, R. (2025). AI and peer reviews in higher education: students' multimodal views on benefits, differences and limitations. Technology, Pedagogy and Education, 1–19. https://doi.org/10.1080/1475939X.2025.2480807

Designing an AI Reading Assistant to Support Home–School Engagement

Jhih-Kai TANG[a*], Yi-Cheng TSAI[b] , Jia-Yan LEE[c], Chang-Yen LIAO[d]& Tak-Wai CHAN [e]
[abcde]*Graduate Institute of Network Learning Technology, NationalCentral University, Taiwan*
*kentang0127@gmail.com

Abstract: This study proposes an AI reading assistant system grounded in the Interest-Driven Creator (IDC) theory to enhance parental engagement in literacy-oriented education. The system integrates students' reading records, grade-level data, and a thematic book map, enabling parents to receive personalized, real-time responses via natural language queries. It consists of three modules: (1) a thematic book map covering three domains and twelve subcategories, each with multidimensional tags for semantic matching; (2) a reading data integration mechanism that consolidates parent-entered book titles, dates, reflections, and optional media into personalized reading histories; and (3) an AI recommendation engine that analyzes reading history, preferences, and curriculum topics to generate relevant, age-appropriate suggestions. A preliminary test with six parents found most valued the assistant's timely, clear feedback for understanding their child's reading habits, while one suggested more specific follow-up recommendations. A six-week trial at an experimental elementary school in Taoyuan, Taiwan, will involve 30–40 families with students in Grades 2–6 to evaluate system usability, its impact on parents' thematic awareness, and its potential to foster richer parent–child reading interactions, informing optimization of the recommendation and thematic extension features.

Keywords: AI Reading Assistant, Thematic Reading, Family Engagement, Semantic Response System

1. Introduction

"Reading behavior" refers to the set of activities and strategies that students display when engaging with a text, including book selection, reading methods, comprehension processes, reflection, and extended application. Ideally, reading behavior should be closely connected with curriculum knowledge so that students can transform the information gained from reading into subject knowledge and interdisciplinary understanding. However, in the current educational context, there is often a disconnection between students' reading behavior and the curriculum. Many reading activities lack integration with subject knowledge, which not only limits the potential of reading as a learning strategy but also diminishes the sustainability of students' reading motivation(Schmoker, 2018). On the other hand, while parents generally value their children's reading development, the lack of timely curriculum information and personalized book recommendations often makes it difficult for them to effectively participate in home-based reading interactions, thereby reducing the potential role of the family in learning support.

Recent studies have emphasized the importance of parental involvement in primary school students' reading development. Çalışkan and Ulaş (2022), through a quasi-experimental design, found that parental involvement in reading activities significantly improves students' reading comprehension, reading motivation, and overall attitudes toward reading, including both leisure reading and academic reading. Bonanati and Rubach (2022)further pointed out that reciprocal involvement from both home and school creates a positive feedback loop for students' reading performance. Villiger et al. (2012) demonstrated that a school–home-based reading intervention could sustain improvements in students' reading motivation several months after the intervention, highlighting the medium- to long-term impact of family involvement on reading learning.

The design of the present system is grounded in the Interest-Driven Creator Theory(Chan et al., 2018), which emphasizes three interrelated processes: (1) linking thematic book tags with learning contexts to stimulate students' reading interest (*Interest*); (2) facilitating students' linguistic and cognitive expression and creation (*Creation*) through storytelling records and interactions with the AI assistant; and (3) guiding students to develop stable reading habits (*Habit*) through daily reading feedback and thematic exploration, where parents or teachers pose conversational questions after a student reads a book. In this study, "reading feedback" refers to direct responses provided by parents or teachers after a student engages in reading activities at home or school, or to the interactive feedback generated during "book talk" sessions between teachers and students. These inputs form an essential basis for the AI assistant's analysis and recommendations.

To address the above challenges and align with the theoretical foundation, this study proposes an AI reading assistant system designed to help parents actively pose questions during daily home interactions and receive personalized book recommendations or thematic summaries aligned with the student's reading status and curriculum topics. The system integrates students' reading records and grade-level information, constructs a thematic book map designed by the research team, and applies semantic tagging, semantic matching, and thematic expansion techniques to generate response content. Through this intelligent support framework, parents can participate more effectively in guiding and conversing about the reading process, thus strengthening the family's role and potential in literacy-oriented reading education.

2. System introdution

The AI reading assistant system developed in this study is designed to help parents promptly understand their child's reading status in daily contexts and provide personalized responses relevant to the reading content, thereby promoting literacy-oriented parent–child reading. The system architecture consists of three core modules: the Thematic Book Map, the Reading Data Integration Mechanism, and the AI Recommendation Engine.

2.1 Thematic Book Map Design

The thematic book map was designed by the research team based on book classification methods, covering three major domains and subdivided into twelve subcategories. Each book is labeled with 4–6 multidimensional tags—such as thematic attributes, reading difficulty, learning stage, interest domain, and interdisciplinary connections—to support precise searches and semantic matching. This design helps students extend reading themes, build intertextual connections, and provides the semantic foundation for the AI assistant's summaries and responses.

2.2 Reading Data Integration Mechanism

The system integrates students' reading records and grade-level information to build personalized reading histories and map them to the book categories for accurate thematic matching. Parents can input the book title, reading date, and brief observational notes, and may also upload photos of reading activities or provide feedback. The AI assistant analyzes these data to generate reading summaries, assess thematic coverage and reading frequency, and produce an initial score to support subsequent recommendations and learning assistance. This integration also provides teachers with a reference for adjusting instructional strategies.

2.3 *AI Recommendation Engine*

Parents can use natural language queries (e.g., "Please recommend books related to my child's recent reading topics"), and the system will analyze reading history, preferences,

and curriculum topics to recommend books with high semantic relevance and appropriate difficulty, along with reasons for the recommendation (e.g., “Matches past reading themes,” “Expands reading perspective”). Unlike commercial systems that operate primarily based on sales data (e.g., Amazon), this system combines students’ reading records, parent/teacher feedback, and the book map to perform semantic analysis and generate recommendations. In the future, a thematic extension mechanism will be added to explore co-occurrence relationships and semantic distances between tags, enabling the discovery of new topics and books to broaden the breadth and depth of students’ reading.

3. Preliminary Evaluation (Planned)

This study is in the system design and early development phase. A six-week trial will be conducted at an experimental elementary school in Taoyuan, Taiwan, with 30–40 families of students in Grades 2–6. Parents will input their child’s reading records (e.g., book title, date, brief reflection) into the platform, and teachers will provide curriculum topics at the start of the semester. The AI assistant will then analyze the data to generate reading summaries and thematic associations in response to parents’ natural language queries.

The preliminary trial focused on the Reading Data Integration Mechanism, testing whether the system could effectively consolidate parent-entered reading records and produce accurate, useful summaries. Six parents participated in a small-scale usability test; most reported that the AI assistant provided timely and clear feedback that helped them understand their child’s reading habits, while one suggested adding more specific follow-up recommendations.

The full evaluation will examine: (1) System usability and interaction quality – ease of use, response accuracy, and perceived value; (2) Enhancement of thematic awareness – whether summaries improve parents’ understanding of reading themes and trends; (3) Promotion of parent–child interaction – whether AI responses encourage more frequent reading discussions at home. The findings will guide system optimization and recommendation feature development.

Acknowledgements

This study was funded by the National Science and Technology Council of Taiwan (112-2410-H-008-020-MY3) and by Research Center for Science and Technology for Learning, National Central University, Taiwan.

References

Bonanati, S., & Rubach, C. (2022). Reciprocal Relationship between Parents’ School- and Home-Based Involvement and Children’s Reading Achievement during the First Year of Elementary School. *Societies*, *12*(2). https://doi.org/10.3390/soc12020063

Çalışkan, E. F., & Ulaş, A. H. (2022). The Effect of Parent-Involved Reading Activities On Primary School Students Reading Comprehension Skills, Reading Motivation, and Attitudes Towards Reading. *International Electronic Journal of Elementary Education*. https://doi.org/10.26822/iejee.2022.260

Chan, T.-W., Looi, C.-K., Chen, W., Wong, L.-H., Chang, B., Liao, C. C. Y., Cheng, H., Chen, Z.-H., Liu, C.-C., Kong, S.-C., Jeong, H., Mason, J., So, H.-J., Murthy, S., Yu, F.-Y., Wong, S. L., King, R. B., Gu, X., Wang, M.,…Ogata, H. (2018). Interest-driven creator theory: towards a theory of learning design for Asia in the twenty-first century. *Journal of Computers in Education*, *5*(4), 435-461. https://doi.org/10.1007/s40692-018-0122-0

Schmoker, M. (2018). *Focus: Elevating the Essentials to Radically Improve Student Learning*. ASCD.

Villiger, C., Niggli, A., Wandeler, C., & Kutzelmann, S. (2012). Does family make a difference? Mid-term effects of a school/home-based intervention program to enhance reading motivation. *Learning and Instruction*, *22*(2), 79-91. https://doi.org/10.1016/j.learninstruc.2011.07.001

Alternative Story Construction Support System for Narrative Therapy through Fact Recall Assistance

Rei Muguruma [a], Shogo Yano [a], Tomoko Kojiri [b]
[a] *Graduate School of Science and Engineering, Kansai University, Japan*
[b] *Faculty of Engineering Science, Kansai University, Japan*
k686879@kansai-u.ac.jp

Abstract: Students who suffer from psychological stress tend to perceive their experience negatively and estimate themselves low. In order to support such students, narrative therapy is important. The narrative therapy is the psychotherapy by changing dominant stories, which consist of the negative aspects of the facts, to the alternative stories, which consist of the positive aspects of them. Such therapy is achieved by the skilled therapists, but it is difficult for the schools to prepare enough number of therapists for the stressed students. This study proposes a system that substitutes for human therapists by supporting the construction of alternative stories.

Keywords: Narrative therapy, self-evaluation, alternative story, dialogue system, fact recall, psychological support

1. Introduction

In these years, schools are required to provide students with emotional support. Students who suffer from psychological stress tend to perceive their experience negatively and estimate themselves low. Narrative therapy is a form of psychotherapy aimed at improving self-esteem by reframing dominant stories which consist of the negative aspects of the facts, to the alternative stories, which consist of the positive aspects of them (White, M., et al., 1990). This approach is achieved by the skilled therapists, but it is difficult for the schools to prepare enough number of therapists for the stressed students. Based on this, our research aims to construct a system that can serve as a substitute for a narrative therapist.

In research targeting therapy support, chatbots such as Woebot (Fitzpatrick, K. K., et al., 2017) have emerged. Woebot detects negative thought patterns through daily conversations and applies techniques from Cognitive Behavioral Therapy (CBT) to guide users toward more realistic, positive interpretations. This agent focuses on alleviating emotional distress and correcting cognitive distortions. However, as they do not model how self-evaluation is formed, they may only offer temporary relief.

This study models human self-evaluation process. Based on this model, it proposes a system that can help recall positive facts to construct alternative stories.

2. Self-Evaluation Model

Self-evaluation is the process through which individuals determine whether they perceive themselves positively or negatively based on their own facts.

Figure 1 illustrates the self-evaluation model expressed in these three levels. Fact Evaluation is the evaluation result of individual fact from a specific perspective. Perspective Evaluation indicates the evaluation result of perspective, which is derived by

the fact evaluations of associated facts. Overall Evaluation is the self-evaluation derived from all perspective evaluations. It is calculated by adding perspective evaluation by considering each perspective's weight.

Based on this model, to improve overall evaluation, either the weight of a perspective or its evaluation must be changed. Since perspective weights represent personal values, they are not easily altered. In contrast, perspective evaluations can be changed by modifying the set of fact evaluations. This can be done by either recalling new relevant facts or reevaluating known facts from different perspectives.

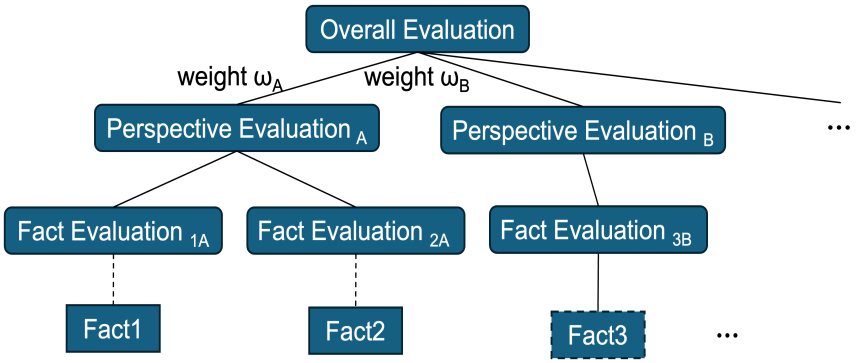


Figure 1. Self-Evaluation Model

This study proposes a therapist system that engages in dialogue to encourage the recall of new facts and the reevaluation of existing ones, thereby generating new fact evaluations.

3. Narrative Therapist System

3.1 System Interaction

Figure 2 illustrates the Interaction of the narrative therapist system. The system behaves like a narrative therapist and support a user to construct the alternative story based on his dominant story. The system does not create the alternative story, but it supports users to derive new facts and their evaluation perspectives that help users to create the alternative story by themselves.

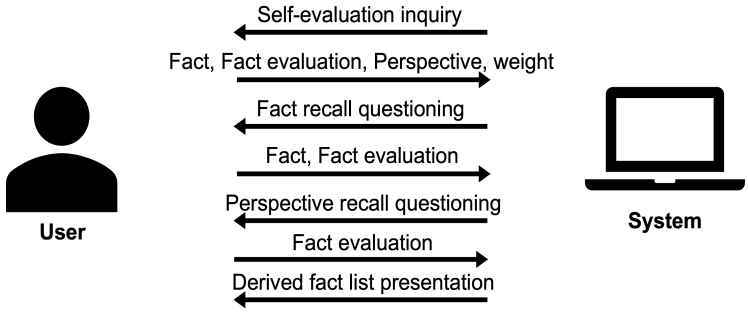


Figure 2. The Interaction of The Narrative Therapist System

3.2 Fact Recall Questioning and Perspective Recall Questioning

.

Fact recall questioning to the given facts and perspectives helps the user to recall new facts evaluated from a particular perspective. It fills in the template of each question using the facts and perspective given as a dominant story. (Type of action) (characters) (place), and (time) are the components of the given fact. (Perspective) means given the perspective. Followings are the templates: *"Do you have another episode where you (type of action) in a different place, related to being a (perspective) person?"* , *"Do you have another episode where you (types of action) (place) at a different time, related to being a (perspective) person?"*

Perspective recall questioning asks the user to evaluate known facts from new perspectives. For all perspectives that have not been applied to a fact, the system generates questions using the template: *"Does the fact that you (type of action) (place) (time) relate to being a (perspectives that are not applied to the fact) person?"*

3.3 Prototype System

The interface of the prototype system is shown in Figure 3. The interface is composed of a dialogue display area for conversation between the system and the user, and a response input area for answering questions. In the dialogue area, the system's messages appear in grey on the left, and the user's responses appear in blue on the right. The input area changes according to the given questions by the system so as to make the user fill in the required elements easily. Questions asking for perspectives are answered in free text, questions asking for perspective evaluation and fact evaluation use a five-point scale input, and questions asking for entering facts use a structured input in the format "I (type of action) (place) (time)". For organizing facts chronologically, the input is a selection of Fact 1 and Fact 2 in the form "after Fact 1, before Fact 2". A "None" button is also available in case the user cannot think of an answer.

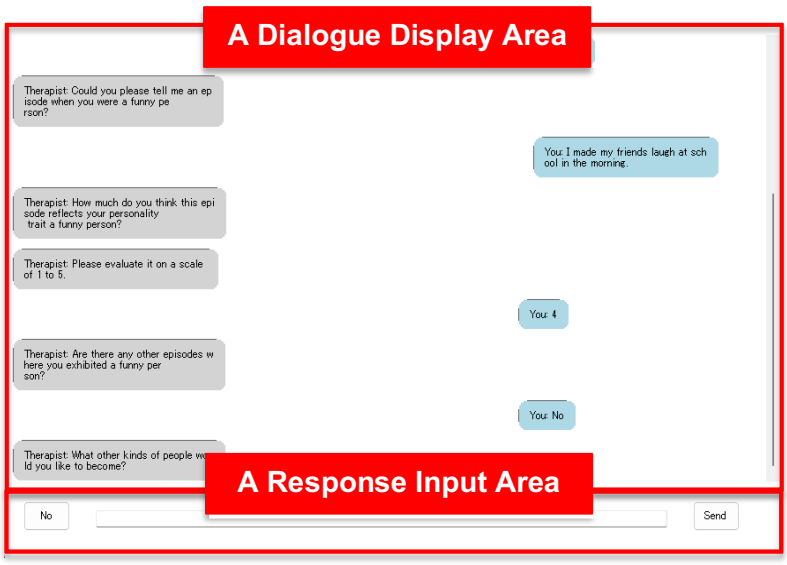


Figure 3. The System Interface

4. Conclusion

In this study, we proposed a self-evaluation model consisting of fact evaluation, perspective evaluation, and overall evaluation. Based on this model, we developed a system that simulates the role of a therapist and guides users in deriving elements for constructing alternative stories. The system generates questions based on the user's self-evaluation to support them recall new facts associated with the same perspective, or to reevaluate known facts from different perspectives. As our next step, we will conduct experiments to evaluate the validity of the activity model and the effectiveness of the proposed system.

References

White, M., & Epston, D. (1990). *Narrative means to therapeutic ends*. W. W. Norton & Company.

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health, 4*(2), e19. https://doi.org/10.2196/mental.7785

Fostering Creative Thinking in Animation Scriptwriting through AI-Driven Analysis and Socratic Questioning

Panadda Jaiboonlue[a*], Wasan Na Chai[b] & Nattapol Kritsuthikul[b]

[a]*Department of Animation and Digital Media, Faculty of Science and Technology Bansomdejchaopraya Rajabhat University, Bangkok, Thailand*
[b]*Language and Semantic Techonolgy Laboratory, National Electronics and Computer Technology Center, Pathum Thani, Thailand*
*panadda.ja@bsru.ac.th

Abstract: This paper proposes an AI-powered framework to foster creative thinking in animation scriptwriting by integrating narrative analysis, Retrieval-Augmented Generation (RAG), and Socratic questioning. The system enhances existing scriptwriting tools through modules that analyze narrative structure, assess creativity, verify logical consistency, and generate theory-informed feedback. Its core innovation is the Feedback & Socratic Questioning Module, which provides personalized, context-aware prompts supported by RAG-based retrieval of relevant storytelling theories and examples. Delivered via an interactive chatbot interface, this feedback engages students in real-time reflective dialogue and guides them through iterative script development. By combining computational analysis with pedagogically grounded questioning, the framework demonstrates how AI can support higher-order thinking and narrative creativity in animation education.

Keywords: Creative Thinking, Animation Scriptwriting, Socratic Questioning, AI-Driven

1. Introduction

In the field of animation education, scriptwriting is a foundational skill that enables students to construct meaningful narratives and convey emotion through visual storytelling. However, many students face challenges in transforming abstract ideas into coherent plots, developing characters with depth, and applying narrative theories such as the three-act structure or character arcs. While instructors provide essential feedback, the increasing number of students and limited contact hours often hinder the depth and frequency of guidance provided during the creative writing process. To address this issue, previous work introduced a learning support tool aimed at enhancing students' awareness of narrative structure and creative thinking in scriptwriting. Building upon this foundation, the current research proposes an extended framework that integrates Artificial Intelligence (AI) (Holmes, Bialik, & Fadel, 2019) with Retrieval-Augmented Generation (RAG). This system is designed to analyze both screenplays and storyboards and stimulate reflective thinking by emulating Socratic questioning (Paul & Elder, 2006) an approach in which guided questions help learners examine their ideas, clarify intentions, and consider narrative alternatives.

The paper proposes a framework for an intelligent tutoring system that supports both creativity and structured thinking in animation scriptwriting. It integrates Vision-Language Models for multimodal analysis, computational metrics for creativity assessment, and Retrieval-Augmented Generation for grounding in narrative knowledge. Unlike existing tools, it emphasizes real-time, context-aware guidance that nurtures creative reasoning. This framework establishes a foundation for next-generation AI tutoring systems in creative education, with a focus on animation storytelling.

2. Animation script writing supporting tool

A prior study (Jaiboonlue et al., 2023) introduced a structured learning support tool grounded in narrative theory to enhance students' awareness and understanding of animation scriptwriting by guiding them through essential components such as setting, theme, character arcs, and conflict resolution, while enabling decomposition of ideas, logical structuring, and iterative instructor feedback. Experimental results showed significant improvements in script completeness, coherence, and adherence to storytelling conventions, with higher evaluation scores on revised scripts and a marked reduction in missing mandatory elements (e.g., 34% of students corrected earlier omissions). This work established the value of theory-informed, structured, and feedback-driven scriptwriting instruction, providing a foundation that the current research extends by integrating AI and RAG to enrich feedback and foster creative, reflective thinking.

3. Proposed Framework System

The proposed framework introduces an intelligent support system for animation scriptwriting that integrates AI-driven analysis, Retrieval-Augmented Generation (RAG), and Socratic questioning techniques. The system is designed to work alongside an animation script writing supporting tool that students use to draft and revise scripts and storyboards under instructor guidance. An overview of the proposed system is illustrated in Figure 1.

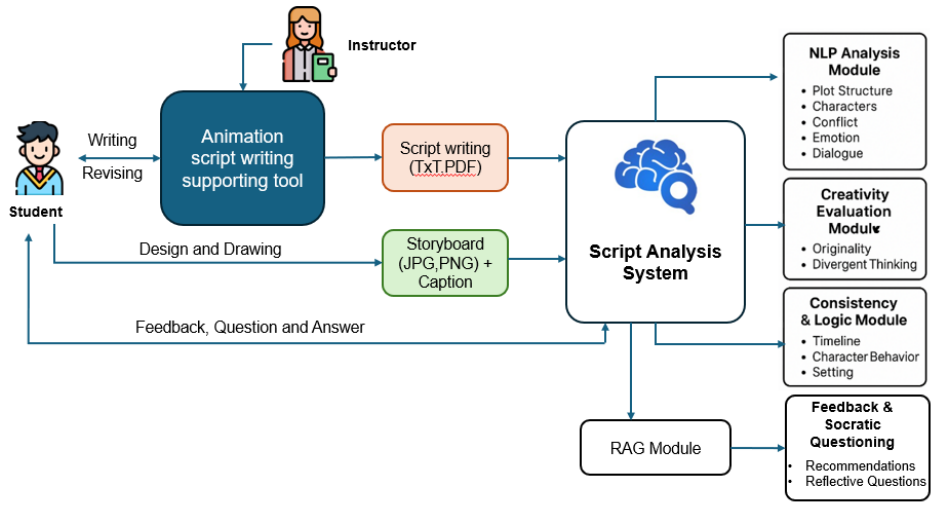


Figure 1. An overview of animation scriptwriting analysis system

The process begins with students composing their scripts using the supporting tool, with instructional guidance to ensure alignment with narrative theories such as the three-act structure, character development, and conflict resolution. Once the initial script has been reviewed and refined, it is exported in .TXT or .PDF format. Students then proceed to design a corresponding storyboard, complete with visual panels and descriptive captions, which represent the narrative flow of their animated film. Both the script and storyboard are submitted to the Script Analysis System, which connects directly to the authoring tool. This system is composed of 4 modules:

- **NLP Analysis Module:** This module uses a pre-trained Vision-Language Model (VLM) like CLIP (OpenAI, 2021) to verify alignment between storyboards and scripts. Storyboard panels are encoded into visual embeddings, while corresponding script captions are encoded into text embeddings within the same latent space. Alignment is measured via cosine similarity: scores near 1.0 indicate strong match, while lower scores signal mismatches for student reflection. The method relies on contrastive learning, which optimizes the model to link correct image-text pairs and separate incorrect ones.

- **Creativity Evaluation Module:** This module quantifies creativity by evaluating originality and divergent thinking in the narrative using computational linguistics techniques, turning abstract creative concepts into measurable metrics.
- **Consistency & Logic Module:** verifies logical coherence in terms of timeline, character behavior, and setting continuity.
- **Feedback & Socratic Questioning Module**: This module acts as the pedagogical core, combining outputs from all other modules into an interactive dialogue. An LLM, prompted with the student's script, analysis results (e.g., originality or alignment scores), and knowledge from the RAG module, generates targeted feedback that goes beyond generic comments for instance, suggesting revisions grounded in both analysis and screenwriting theory. This process ensures real-time, context-aware guidance and establishes a clear link from technical analysis to meaningful pedagogical support. Future work will explore optimal timing of Socratic questioning through A/B testing across different stages of the writing process.

Analytical results from the modules are sent to the RAG module(Bhatia, Vats, & Aggarwal, 2021), which retrieves relevant knowledge from sources like narrative theory, exemplary scripts, and rubric-based criteria. This context allows the system to provide targeted, theory-informed feedback and reflective prompts. All outputs are then consolidated in the Feedback & Socratic Questioning Module, delivering tailored guidance through a chatbot embedded in the scriptwriting tool. This real-time, interactive platform supports iterative refinement, combining technical analysis with reflective questioning to improve learning outcomes.

The primary pedagogical objective of the proposed system is to foster students' creative thinking through iterative reflection, guided exploration, and theory-informed questioning. By decomposing scripts into analyzable components and providing tailored feedback enhanced with Socratic questioning and external knowledge via RAG, the system encourages learners to move beyond surface-level storytelling. It supports the development of higher-order thinking skills such as divergent thinking, originality, synthesis of ideas, and evaluative judgment. Ultimately, the system aims to transform the scriptwriting process into a reflective, creative, and cognitively engaging learning experience.

4. Conclusion

This paper proposed a technically-grounded framework for an intelligent animation scriptwriting support system that integrates AI-driven analysis, RAG-based retrieval, and Socratic questioning to enhance students' creative thinking. By specifying the implementation strategies for each module—leveraging Vision-Language Models for multimodal alignment, computational metrics for creativity, and a RAG architecture for knowledge grounding—this paper provides a clear and feasible blueprint for development. The framework builds upon an existing scriptwriting support tool and extends its capabilities by providing structured, theory-informed, and cognitively engaging feedback.

References

Bhatia, A., Vats, A., & Aggarwal, S. (2021). A survey on retrieval-augmented generation. International Journal of Computer Applications, 183(21), 10–14. https://doi.org/10.5120/ijca2021921509

Holmes, W., Bialik, M., & Fadel, C. (2019). Artificial intelligence in education: Promises and implications for teaching and learning. Center for Curriculum Redesign.

Jaiboonlue, P., Chai, W. N., Ruangrajitpakorn, T., & Supnithi, T. (2023, December). Improving thinking awareness in animation scriptwriting through learning supporting tool. In Proceedings of the 31st International Conference on Computers in Education.

Paul, R., & Elder, L. (2006). The thinker's guide to the art of Socratic questioning. Foundation for Critical Thinking.

OpenAI. (2021). *CLIP: Contrastive Language-Image Pre-training*. GitHub. Retrieved August 16, 2025, from https://github.com/openai/CLIP

C2
Computer-Supported Collaborative Learning and Learning Sciences (CSCL/LS)

Enhancing Learning Achievement through GenAI Assisted Collaborative Drama-Based Role-Play Learning Metaverse in Higher Education

Su-Hang YANG[a], Mahesh LIYANAWATTA[b], Tzu-Hsien CHUANG[b],Hao TANG[b], Jen-Hang WANG[b], Gwo-Dong CHEN[b*]

[a]*Department of Hospitality Management, Chien Hsin University of Science and Technology, Taiwan*

[b]*Department of Computer Science and Information Engineering, National Central University, Taiwan*

* gwodong@gmail.com

Abstract: Cultivating real-world problem-solving, decision-making, communication, professional identity (PI), and workplace-related skills in higher education is crucial for preparing students for career readiness, yet conventional project-based learning (PBL) often lacks authentic, immersive learning contexts that foster deep engagement and learning achievement development. Addressing this gap, we designed a novel Learning Metaverse environment enhanced with assistive Generative AI (GenAI) agents, grounded in Self-Determination Theory (SDT) and Identity-Based Motivation (IBM) Theory. This approach enables students to create personalized 3D avatars through large language models (LLMs) by communicating verbally and non-verbally, using the Ready Player Me platform to map their facial features through 2D images and customize their PI, while engaging in collaborative, drama-based role-play learning activities based on a drama script related to the textbook content, assisted by real-time GenAI feedback. In a quasi-experimental study within a Taiwanese university's hospitality management program, two cohorts participated in drama-based role-play simulating Japanese restaurant workplace culture: the experimental group (n=34) used the GenAI assisted learning metaverse with avatar customization for better PI, while the control group (n=32) used a conventional metaverse with limited predesigned selective avatar options and no GenAI feedback. Results demonstrated that the GenAI-assisted learning metaverse significantly enhanced learning achievement, engagement, and key dimensions of PI, including responsibility, self-identity, accountability, self-efficacy, and belongingness. These findings underscore the efficacy of integrating metaverse and GenAI technologies to create immersive, contextually rich educational environments. Implications include providing students with authentic opportunities for PI construction and skill development, offering educators actionable frameworks for technology-enhanced learning design, and informing policymakers on the value of investing in robust, theory-driven digital learning infrastructures to improve graduate employability and lifelong learning outcomes.

Keywords: Collaborative learning, Metaverse, Interactive learning environments, Drama-based learning, GenAI-supported scaffolding

1. Introduction

The rapid advancement of digital technologies is reshaping higher education by creating new ways to develop essential career skills like problem-solving, decision-making, communication, and PI. Among these, the metaverse, a 3D environment enabling real-time interaction and digital identity, stands out as a promising educational innovation (Barry &

Kanematsu, 2025; Gim et al., 2023). The metaverse allows learners to transcend the limitations of physical classrooms, enabling authentic, multisensory, and context-rich experiences that closely mirror real-world professional environments (Zhang et al., 2022). While PBL (Han & Bhattacharya, 2002) provides authentic learning, metaverse environments enhance this by providing embodied, immersive experiences that simulate professional scenarios. (Zhang et al., 2022). However, a critical research gap remains on how learning metaverse combined with advanced Artificial Intelligence (AI) and grounded in motivation and identity theories can effectively promote deep engagement, learning achievement, and PI development.

Self-Determination Theory (SDT) posits that the satisfaction of students' basic psychological needs for autonomy, competence, and relatedness is fundamental for promoting intrinsic motivation, sustained engagement, and effective learning (Deci & Ryan, 2000). Complementing this, Identity-Based Motivation (IBM) theory highlights that students are more likely to engage in and persist with learning activities when these activities are perceived as congruent with their salient identities, and when they view challenges as meaningful instead of impossible to overcome. (Oyserman & Destin, 2010). In higher education, the development of a strong PI is increasingly recognized as a key factor in students' confidence, self-efficacy, goal setting, and employability, as well as their ability to navigate complex workplace environments (Tan et al., 2017; Tomlinson & Jackson, 2019).

Concurrently, the integration of AI technologies such as Chatbots or GenAI agents, intelligent tutoring systems, and real-time analytics has shown significant potential to personalize learning, provide adaptive feedback, and support students' cognitive and affective needs(Vázquez-Cano et al., 2021). Yet, empirical research into the combined impact of these technologies within immersive metaverse environments, particularly on students' PI creation and engagement in learning, remains limited. Addressing this gap, this study presents a novel Learning Metaverse environment enhanced with assistive GenAI agents. This environment empowers students to create and embody personalized 3D avatars by mapping their facial features through 2D images of their own face, selecting a gender, and customizing their digital identity, while participating in collaborative, drama-based role-play learning activities that simulate authentic workplace scenarios. Real-time GenAI feedback further supports students' autonomy, competence, and relatedness, while fostering identity-congruent experiences that underpin deep engagement and PI development.

This research is guided by two primary objectives: first, to examine whether the GenAI-assisted learning metaverse enhances students' learning achievements in collaborative classroom settings; and second, to verify its impact on student engagement and key dimensions of PI, including responsibility, self-identity, accountability, self-efficacy, and belongingness. (1) Does the GenAI-assisted learning metaverse enhance students' learning achievement compared to a conventional type of learning metaverse for collaborative drama-based role-play? (2) To what extent do avatar construction and real-time GenAI feedback mediate participant engagement, PI, and learning achievements? By addressing these research questions, this study aims to advance the understanding of how SDT and IBM theories in technology-enhanced PBL environments can provide authentic learning opportunities for PI construction and skill development, both of which are crucial for preparing students for career readiness, while offering actionable insights for educators, researchers, system developers, and policymakers seeking to leverage the metaverse and GenAI for more effective, engaging, and future-oriented education.

2. Literature Review

2.1 Theoretical Foundation

This study integrates SDT and IBM theories to explain how metaverse learning environments, enriched with embodied interactions for problem-solving, decision-making, communication, PI, and workplace-related skills development through collaborative, drama-

based role-play learning activities with GenAI feedback, can enhance career readiness in higher education. SDT emphasizes that satisfying the psychological needs for autonomy (volitional action), competence (skill mastery), and relatedness (social connection) promotes intrinsic motivation and sustained engagement (Deci & Ryan, 2000). In metaverse learning environments, these needs are addressed through avatar creation (autonomy), real-time GenAI feedback (competence), and collaborative drama-based role-play (relatedness). IBM Theory complements this by emphasizing that motivation is driven by identity congruence: learners persist when actions align with their PI and interpret challenges meaningfully (Oyserman, 2007; Oyserman & Sorensen, 2009). Metaverse environments enable students to build and express their PI by designing 3D avatars resembling themselves or aspirational roles, encouraging authentic behavior within realistic simulations (Tan et al., 2017; Zhang et al., 2022). Avatar construction enhances embodiment, agency, and identity ownership, strengthening PI (Repetto & Riva, 2023). GenAI supports this process by providing adaptive feedback aligned with learners' identity development, exemplified by AI role-play companions that guide skill-building and identity enactment in hospitality simulations (Chiu et al., 2023; Min & Cai, 2022). Together, these elements improve students' ability to apply problem-solving, communication, and responsibility skills in real-world contexts (Dahri et al., 2024).

2.2 Metaverse in Education

The metaverse, defined as a 3D digital space blending virtual and real-world elements, provides immersive, interactive platforms increasingly studied for educational potential (Park & Kim, 2022). Despite advances in platforms like Roblox, ZEPETO, and Second Life that offer seamless multisensory interactions, a significant gap remains in systematically integrating metaverse environments into higher education curricula to enhance authentic, collaborative, and career-relevant learning (Hwang & Chien, 2022; Rospigliosi, 2022; Zhang et al., 2022). Early virtual environments focused mostly on entertainment or informal learning rather than structured, curriculum-based outcomes (Davis et al., 2009; Haycock & Kemp, 2007; Livingstone & Kemp, 2006). Recent initiatives at universities such as INSEAD and the University of Pennsylvania demonstrate growing use of metaverse technologies for experiential learning, remote collaboration, and digital identity construction (Lee, 2022). However, empirical research is still limited on how these environments effectively foster PI, workplace skills, and sustained engagement in higher education. The broad technological infrastructure requirements, such as high-speed internet, cloud computing, AI analytics, and 3D modeling, present both opportunities and challenges for scalable and equitable adoption. (Gsaxner et al., 2023; Yang et al., 2022). Immersive metaverse applications have shown promise in disciplines like medicine, engineering, and business, where realistic simulations and real-time feedback benefit students (Gsaxner et al., 2023; Salloum et al., 2023; Siyaev & Jo, 2021). Yet, comprehensive empirical studies are necessary to thoroughly assess the technology's capabilities, instructional approaches, and learning outcomes. Addressing this research gap is critical for optimizing metaverse integration to enhance student engagement, professional development, and career readiness in a digitally connected world.

3. Methodology

3.1 Learning Model

Figure 1 illustrates the study's learning model, based on Morrison's Unit method of teaching and learning (Morrison, 1926), chosen for its clear structure in organizing collaborative activities. Although originally developed before immersive and AI technologies, this model was enhanced by integrating technological affordances such as real-time feedback, personalized avatar creation, and multisensory experiences. Additionally, contemporary instructional models like TPACK (Technological Pedagogical Content Knowledge) and SAMR (Substitution, Augmentation, Modification, and Redefinition), which explicitly address technology integration, could inform future research to maximize the potential of learning metaverse environments.

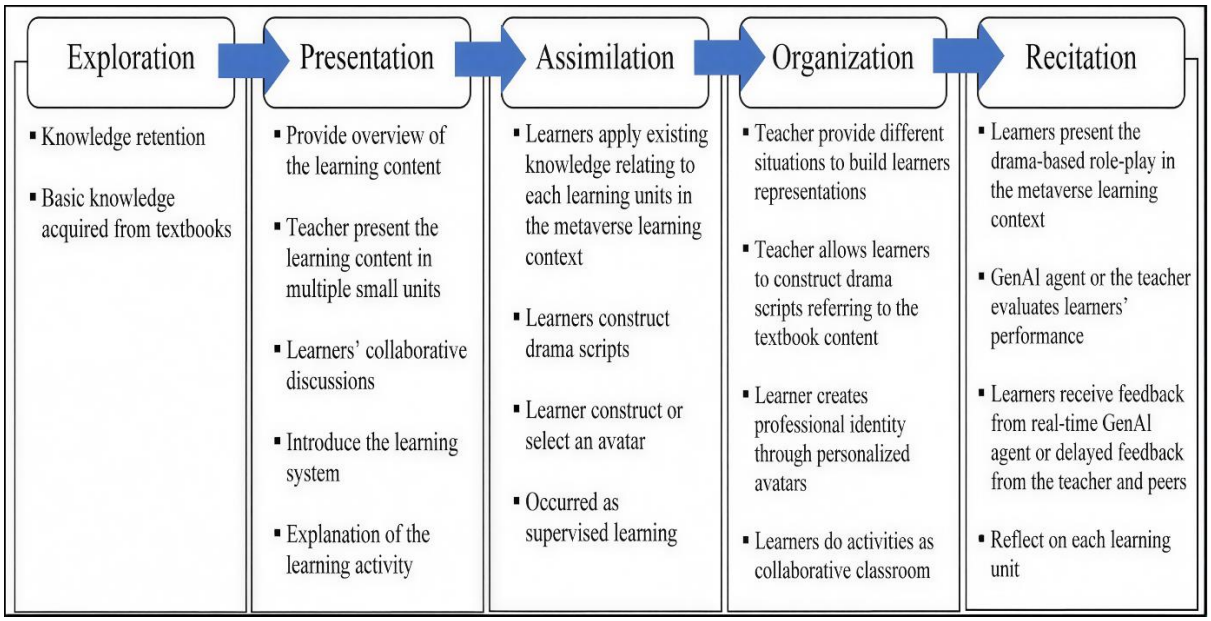


Figure 1. Learning Model

3.2 Learning Context

Figure 2 illustrates the learning context. In this context, students' avatars acted as extensions of their PI, enriching their personalized experiences inside the metaverse. Moreover, students engaged in collaborative drama-based role-play learning activities with peers while real-time GenAI agents provided guidance and feedback based on their virtual representations, facilitating meaningful learning and helping them understand and refine their mistakes immediately. Ultimately, this approach enabled students to experience personalized and immersive learning, improving their learning achievements and sustaining their motivation and engagement to learn. Figure 3 shows the classroom arrangement for the collaborative drama-based role-play learning activity in which student performers use their embodied interactions in front of the Zed2i depth camera and computer monitor, while other classmates, as the audience, observe the performance through a projection screen. The student performers have the freedom to construct their avatars and control them according to the drama script they created, and live inside the learning metaverse to complete the task.



Figure 2. Learning Context

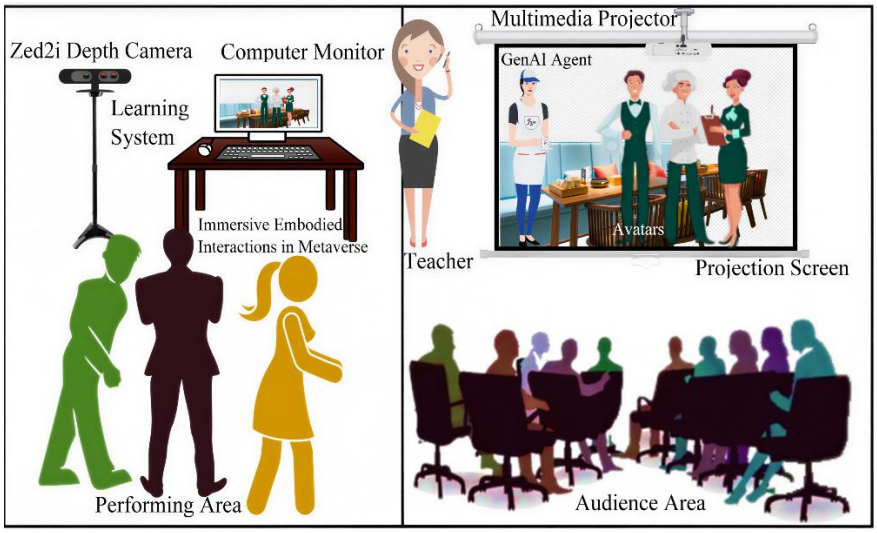


Figure 3. Classroom Arrangement for Learning Metaverse

3.3 Learning System

The learning metaverse was built on the Unity platform and uses natural user interface technologies. This learning system created an immersive, simulated environment to perform collaborative drama-based role-play learning activities in a classroom setting. This system facilitated building learner PI inside the immersive virtual simulation by allowing them to construct their personalized virtual representations as avatars. Learners designed these avatars through LLM by communicating verbally and non-verbally to the Ready Player Me platform, modifying facial features using 2D images of their faces, and also customizing gender type, skin tone, hairstyle, and clothing for highly personalized Avatar construction for better PI. The Zed2i depth camera captured learners' physical actions, behaviors, and object movements, processing them within the system to enhance immersion by synchronizing them with the learning metaverse. On the other hand, instructors could convert teaching materials into scripts, including character settings and scenes, allowing them to give feedback after the activity, oversee learning, and correct errors, thereby improving educational outcomes. Teachers could also use a rehearsal control app to evaluate learners' progress, configure scenes, and monitor learning status in real-time. By enabling the feature of GenAI agents, student performers can receive real-time feedback, oversee the learning process, and correct errors, thereby improving learners' educational outcomes. These GenAI agents provided multiple forms of feedback, such as corrective feedback (notifying students of language or role-play errors and suggesting improvements during real-time participation), reflective prompts (encouraging students to reflect on their choices and behaviors during the drama), and supportive feedback (reinforcing positive contributions, collaboration, and alignment with PI roles). Feedback was delivered in real-time via audio prompts, text overlays on avatars, and post-session summaries generated by the GenAI system, tailored to each student's avatar actions and dialogue.

3.4 Participants

This study involved 66 hospitality management majoring students (average age 18) from a university in Taiwan, working with a Japanese language instructor. The students were randomly divided into an experimental group (34 students: 11 males, 23 females) and a control group (32 students: 13 males, 19 females). Both groups completed the experiment independently to avoid any interference. Participants' identities were anonymized for confidentiality. While students had differing familiarity with GenAI tools like ChatGPT, none had used GenAI agents for learning activities before. To optimize resources, the instructor organized both groups into smaller teams of five to six students, balancing gender, communication skills, and GenAI experience.

3.5 Research Design

The study began with a knowledge retention activity on Japanese restaurant service, followed by a pre-test, unit-wise instruction, and system setup. Students engaged in the learning metaverse through drama script and avatar construction, rehearsals, and completed post-tests and questionnaires. The research examined how a GenAI-assisted learning metaverse with personalized avatars and real-time feedback for scaffolding enhances students' learning achievements and career readiness. Ethical standards, including informed consent, data anonymization, and IRB approval (NCUEREC-111-017), were maintained.

3.6 Measurements and Data Collection

A professional Japanese language instructor specializing in hospitality management designed the course materials based on the syllabus and students' proficiency, including basic vocabulary and conversational phrases relevant to the industry, which were developed into drama scripts. Pre-test and post-test, graded on a 0-100 scale and assessing cultural practices (60 points) and conversation skills (40 points), were used for both groups. ANCOVA analyzes the students' learning achievements in distinct learning approaches applied to experimental and control groups. A questionnaire, using a five-point Likert scale (1 = strongly disagree to 5 = strongly agree), was designed to evaluate learning motivation and engagement (assessed via Intrinsic Motivation and Task Value, adapted from the Motivated Strategies for Learning Questionnaire (Pintrich et al., 1991)) and psychological ownership for better PI (measured across five dimensions based on these references (Buchem, 2012) and (Van Dyne & Pierce, 2004)) with reliability confirmed by Cronbach's alpha.

3.7 Experimental Procedure

The eight-week experiment (see Figure 4) involved both groups taught by the same instructor to ensure consistent methods. The key difference was in learning approaches: the experimental group used a GenAI-assisted learning metaverse with personalized avatars and real-time feedback, while the control group used the learning metaverse without GenAI and received delayed instructor feedback. Students were grouped using an S-curve strategy based on pre-test results. Post-tests and questionnaires assessed motivation, engagement, and psychological ownership for improved PI.

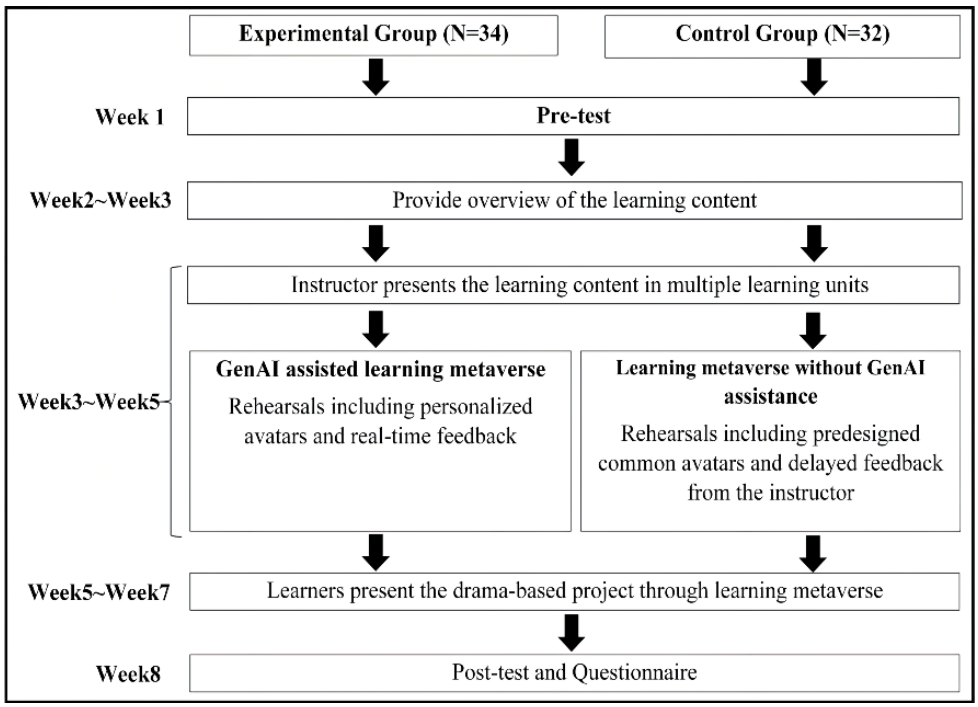


Figure 4. Experimental Procedure

4. Results

4.1 Analysis of the Students' Learning Achievement

Table *1. ANCOVA for Students' Learning Achievement*

Group	N	Mean	*SD*	Adjusted Mean	*SE*	*F*	Sig.	$\eta 2$
Experimental Group	34	84.94	8.01	84.41	1.72	4.52	0.04	0.07
Control Group	32	78.56	12.21	79.13	1.77			

This study analyzed the pre-test and post-test results of the experimental group (GenAI-assisted learning metaverse with real-time feedback) and the control group (Learning metaverse without GenAI assistance and delayed feedback) to assess students' learning achievement in the collaborative drama-based role-play learning activity. An independent samples t-test showed no significant difference in initial abilities between the groups ($p = 0.13$), confirming balanced assignment. Shapiro-Wilk analysis confirmed that both pre- and post-test scores were normally distributed ($p > 0.05$). A regression homogeneity test indicated no interaction between group and pre-test scores ($F = 0.23$, $p = 0.63$), and Levene's test validated the homogeneity of variance assumption for post-test scores ($p = 0.06$).

As shown in Table 1, a one-way analysis of covariance (ANCOVA), controlling for pre-test scores, showed significantly higher post-test scores in the experimental group compared to the control group ($F = 4.52$, $p = 0.04$). The effect size ($\eta 2$ $p = 0.07$) indicates a medium effect, suggesting that the GenAI-assisted learning metaverse with real-time feedback significantly improved student learning outcomes. The effect size for the ANCOVA ($\eta^2 = 0.07$) is considered medium based on Cohen's conventions: small ($\eta^2 = 0.01$), medium ($\eta^2 = 0.06$), large ($\eta^2 = 0.14$) (Cohen, 1988). Thus, the observed effect aligns with the medium benchmark for educational intervention research.

4.2 Analysis of the Questionnaire Results

This study used a highly reliable questionnaire (Cronbach's $\alpha = 0.89$) to analyze students' perceptions of the metaverse learning approaches, focusing on psychological ownership and engagement. In terms of psychological ownership, the experimental group significantly outperformed the control group ($p < 0.05$) across the sub-dimensions of responsibility, self-identity, self-efficacy, and belongingness.

Regarding responsibility, the experimental group averaged 4.38 compared to 3.91 for the control group ($p = 0.00$), indicating a stronger sense of responsibility among experimental students. For self-identity, the experimental group averaged 4.35 versus 3.97 for the control group ($p = 0.03$), showing a stronger connection to their learning outcomes. In self-efficacy, the experimental group averaged 4.29, while the control group scored 3.75 ($p = 0.01$), suggesting greater confidence in completing tasks. Finally, for belongingness, the experimental group averaged 4.38 compared to the control group's 4.00 ($p = 0.03$), indicating a greater sense of belonging in collaborative learning tasks (teamwork).

While the accountability dimension didn't show a significant difference ($p > 0.05$). Therefore, it cannot be concluded that the treatment increased accountability; any mean difference may be attributable to random variation or measurement error rather than the intervention itself. Accordingly, the experimental group scored higher (4.21 versus 3.94), suggesting a stronger sense of accountability. Overall, the findings suggest that the GenAI-assisted learning metaverse with a real-time feedback approach significantly enhances various aspects of students' psychological ownership and engagement with higher PI and may positively influence learning motivation.

Table 2. *Questionnaire Results Analysis*

Facets	Group	Mean	*SD*	Sig.
Sense of Responsibility	Experimental Group	4.38	0.65	0.00
	Control Group	3.91	0.78	
Sense of Self-identity	Experimental Group	4.35	0.65	0.03
	Control Group	3.97	0.78	
Sense of Accountability	Experimental Group	4.21	0.85	0.20
	Control Group	3.94	0.84	
Sense of Self-efficacy	Experimental Group	4.29	0.72	0.01
	Control Group	3.75	0.80	
Sense of Belongingness	Experimental Group	4.38	0.60	0.03
	Control Group	4.00	0.76	

5. Discussion and Conclusions

This study investigated the potential of collaborative drama-based role-play learning activities within an authentic classroom context by employing a GenAI-assisted learning metaverse with real-time feedback to enhance students' learning achievement. This approach enabled learners to engage interactively through personalized digital avatars and develop their understanding within digital realities while strengthening their PI. Students utilized tools such as Ready Player Me and Blender to design their avatars, Mixamo for real-time motion capture with a Zed2i depth camera, and integrated physical and virtual elements using the YOLO algorithm combined with OpenCV libraries and Polycam technologies. These technologies collectively fostered active learning by promoting creativity, increasing engagement, and enhancing intrinsic motivation and psychological ownership through student control over the design and construction processes. Such findings are consistent with prior research demonstrating that avatar customization and embodied interactions in metaverse environments support autonomy and identity development, which are critical for motivation and engagement (Hwang & Chien, 2022; Oyserman, 2007; Park & Kim, 2022).

Statistical analyses revealed that the experimental group exhibited significantly higher intrinsic motivation and task value compared to the control group. Improvements were also observed in key psychological ownership dimensions, including responsibility, self-identity, self-efficacy, and belongingness. These outcomes align with SDT, which emphasizes the importance of satisfying autonomy, competence, and relatedness needs to foster motivation and engagement (Chiu et al., 2023; Ryan & Deci, 2002). However, no significant difference was found in accountability between the groups, suggesting that accountability may be influenced more by external expectations and institutional norms than by internal motivational factors or identity congruence. This finding echoes previous studies indicating that accountability often requires clearly defined external standards and social reinforcement beyond the scope of autonomous learning environments (Tan et al., 2017).

The project-based, collaborative learning environment focused on constructing virtual scenarios and physical prototypes reflects the integration of SDT and IBM theories. LLMs played a critical role in scaffolding task design and construction, enabling students to generate creative drama scripts, 3D environments, and prototype sketches. This facilitated self-regulated learning and iterative refinement, demonstrating how accessible technologies can lower barriers to complex 3D modeling and scenario creation, thereby supporting broader adoption in educational settings. These results support previous findings that AI-powered tools can enhance personalized learning experiences and promote deeper cognitive engagement (Salloum et al., 2023). While technological challenges persist, targeted teacher training and

the development of user-friendly resources are effective strategies to overcome these obstacles and support educators in implementing metaverse-based learning (Kshetri et al., 2022; Siyaev & Jo, 2021).

In conclusion, the findings of this study provide strong evidence that combining learning metaverse platforms with GenAI assistance significantly enhances students' psychological ownership, engagement, and learning outcomes. This supports the growing body of literature advocating for immersive, student-centered learning environments that integrate advanced technologies with robust pedagogical frameworks (Park & Kim, 2022; Ryan & Deci, 2020). Future research should focus on strategies to strengthen students' sense of accountability within these environments, investigate the long-term effects of immersive, metaverse experiences on learning and PI development, and refine teacher support models to facilitate scalable implementation in formal education. By advancing the integration of LLMs, metaverse technologies, and motivational theories, educators and researchers can continue to innovate technology-supported learning and foster authentic, engaging, and career-relevant educational experiences.

Acknowledgements

The authors thank the teacher and students at Chien Hsin University of Science and Technology, Taiwan. This work was supported by the National Science and Technology Council (NSTC) of Taiwan under Grant Agreement Numbers MOST 110-2511-H-008-004-MY3, MOST 111-2410-H-008-012-MY3, and NSTC 113-2811-H-008-010.

References

Barry, D. M., & Kanematsu, H. (2025). Introduction to Metaverse with a Focus on Second Life (SL) and Japan's Island There. In D. M. Barry & H. Kanematsu (Eds.), *Applications of Metaverse and Virtual Reality to Creative Education and Industry* (pp. 45-55). Springer Nature Singapore. https://doi.org/10.1007/978-981-96-3341-8_3

Buchem, I. (2012). Psychological Ownership and Personal Learning Environments: Do sense of ownership and control really matter. PLE Conference Proceedings,

Chiu, T. K. F., Moorhouse, B., Chai, C., & Ismailov, M. (2023). Teacher support and student motivation to learn with Artificial Intelligence (AI) based chatbot. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2023.2172044

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2 ed.). Routledge. https://doi.org/10.4324/9780203771587

Dahri, N., Yahaya, N., & Al-Rahmi, W. (2024). Exploring the influence of ChatGPT on student academic success and career readiness. *Education and Information Technologies, 30*, 8877-8921. https://doi.org/10.1007/s10639-024-13148-2

Davis, A., Murphy, J., Owens, D., Khazanchi, D., & Zigurs, I. (2009). Avatars, people, and virtual worlds: Foundations for research in metaverses. *Journal of the association for information systems, 10*(2), 1. https://doi.org/10.17705/1jais.00183

Deci, E. L., & Ryan, R. M. (2000). The" what" and" why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological inquiry, 11*(4), 227-268. https://doi.org/10.1207/S15327965PLI1104_01

Gim, G., Bae, H. K., & Kang, S. A. (2023). The Effect of Self-Determination and Quality of VR-Based Education in the Metaverse on Learner Satisfaction. In R. Lee (Ed.), *Emotional Artificial Intelligence and Metaverse* (pp. 41-54). Springer International Publishing. https://doi.org/10.1007/978-3-031-16485-9_4

Gsaxner, C., Li, J., Pepe, A., Jin, Y., Kleesiek, J., Schmalstieg, D., & Egger, J. (2023). The HoloLens in medicine: A systematic review and taxonomy. *Medical Image Analysis, 85*, 102757. https://doi.org/10.1016/j.media.2023.102757

Han, S., & Bhattacharya, K. (2002). Constructionism, learning by design, and project-based learning. In M. Orey (Ed.), *Emerging perspectives on learning, teaching, and technology*. University of Georgia.

Haycock, K., & Kemp, J. (2007). Immersive Learning Environments in Parallel Universes: Learning through Second Life. *School Libraries Worldwide, 14*, 89-97. https://doi.org/10.29173/slw6782

Hwang, G.-J., & Chien, S.-Y. (2022). Definition, roles, and potential research issues of the metaverse in education: An artificial intelligence perspective. *Computers and Education: Artificial Intelligence, 3*, 100082. https://doi.org/10.1016/j.caeai.2022.100082
Kshetri, N., Rojas-Torres, D., & Grambo, M. (2022). The Metaverse and Higher Education Institutions. *IT Professional, 24*(06), 69-73. https://doi.org/10.1109/MITP.2022.3222711
Lee, I. (2022). Penn students use digital platform Gather to imitate in-person office hours. Retrieved 2025/5/12, from https://www.thedp.com/article/2020/10/gather-town-penn-cis-virtual-office-hours
Livingstone, D., & Kemp, J. (2006). Massively multi-learner: Recent advances in 3D social environments. *Computing and Information Systems Journal*.
Min, T., & Cai, W. (2022). Portrait of decentralized application users: an overview based on large-scale Ethereum data. *CCF Transactions on Pervasive Computing and Interaction, 4*(2), 124-141. https://doi.org/10.1007/s42486-022-00094-6
Morrison, H. C. (1926). *The Practice of Teaching in the Secondary School* (Vol. 10). University of Chicago Press. https://doi.org/10.1177/019263652601001201
Oyserman, D. (2007). Social Identity and Self-Regulation. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles* (2 ed., pp. 432–453). The Guilford Press.
Oyserman, D., & Destin, M. (2010). Identity-based motivation: Implications for intervention. *Couns Psychol, 38*(7), 1001-1043. https://doi.org/10.1177/0011000010374775
Oyserman, D., & Sorensen, N. (2009). Understanding cultural syndrome effects on what and how we think: A situated cognition model. In *Understanding culture* (1 ed., pp. 25-52). Psychology Press.
Park, S. M., & Kim, Y. G. (2022). A Metaverse: Taxonomy, Components, Applications, and Open Challenges. *IEEE Access, 10*, 4209-4251. https://doi.org/10.1109/ACCESS.2021.3140175
Pintrich, P. R., Smith, D. A., García, T., & McKeachie, W. J. (1991). The motivated strategies for learning questionnaire (MSLQ). *Ann Arbor, MI: NCRIPTAL, The University of Michigan*.
Repetto, C., & Riva, G. (2023). The neuroscience of body memory: Recent findings and conceptual advances. *EXCLI Journal, 22*, 191–206. https://doi.org/10.17179/excli2023-5877
Rospigliosi, P. a. (2022). Metaverse or Simulacra? Roblox, Minecraft, Meta and the turn to virtual reality for education, socialisation and work. *Interactive Learning Environments, 30*(1), 1-3. https://doi.org/10.1080/10494820.2022.2022899
Ryan, R. M., & Deci, E. L. (2002). Overview of self-determination theory: An organismic dialectical perspective. *Handbook of self-determination research, 2*(3-33), 36.
Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary Educational Psychology, 61*, 101860. https://doi.org/10.1016/j.cedpsych.2020.101860
Salloum, S., Al Marzouqi, A., Alderbashi, K. Y., Shwedeh, F., Aburayya, A., Al Saidat, M. R., & Al-Maroof, R. S. (2023). Sustainability Model for the Continuous Intention to Use Metaverse Technology in Higher Education: A Case Study from Oman. *Sustainability, 15*(6), 5257. https://doi.org/10.3390/su15065257
Siyaev, A., & Jo, G. S. (2021). Neuro-Symbolic Speech Understanding in Aircraft Maintenance Metaverse. *IEEE Access, 9*, 154484-154499. https://doi.org/10.1109/ACCESS.2021.3128616
Tan, C. P., Van der Molen, H. T., & Schmidt, H. G. (2017). A measure of professional identity development for professional education. *STUDIES IN HIGHER EDUCATION, 42*(8), 1504-1519. https://doi.org/10.1080/03075079.2015.1111322
Tomlinson, M., & Jackson, D. (2019). Professional identity formation in contemporary higher education students. *Studies in Higher Education 46*. https://doi.org/10.1080/03075079.2019.1659763
Van Dyne, L., & Pierce, J. L. (2004). Psychological ownership and feelings of possession: three field studies predicting employee attitudes and organizational citizenship behavior. *Journal of Organizational Behavior, 25*(4), 439-459. https://doi.org/https://doi.org/10.1002/job.249
Vázquez-Cano, E., Mengual-Andrés, S., & López-Meneses, E. (2021). Chatbot to improve learning punctuation in Spanish and to enhance open and flexible learning environments. *International Journal of Educational Technology in Higher Education, 18*(1), 33. https://doi.org/10.1186/s41239-021-00269-8
Yang, Q., Zhao, Y., Huang, H., Xiong, Z., Kang, J., & Zheng, Z. (2022). Fusing Blockchain and AI With Metaverse: A Survey. *IEEE Open Journal of the Computer Society, 3*, 122-136. https://doi.org/10.1109/OJCS.2022.3188249
Zhang, X., Chen, Y., Hu, L., & Wang, Y. (2022). The metaverse in education: Definition, framework, features, potential applications, challenges, and future research topics. *Frontiers in Psychology, 13*. https://doi.org/10.3389/fpsyg.2022.1016300

How Student-Led Collaborative Metaverse World Creation Fosters Information Literacy and Intercultural Competence: A Quantitative Analysis of a Synchronous International Virtual Exchange

Masako HAYASHI[a*], Takehiro SUZUKI[b], Hirokazu HAYAMA[a] & Keisuke GOTO[a]
[a]*Tohoku University, Japan,* [b]*Kyoto University, Japan*
*masako.hayashi.c5@tohoku.ac.jp

Abstract: This study investigates how student-led Metaverse World creation, conducted as part of a synchronous international virtual exchange (VE), fosters information literacy (IL) and intercultural competence (IC) among university students. Eighteen Japanese students collaborated with international and overseas students in multinational teams to design and present culturally themed Metaverse environments using Unity, VRChat, and a Metaverse World Creation Package. Quantitative analyses of pre- and post-program surveys revealed significant gains in digital fluency, immersive communication, and cultural learning strategies. Notably, participants enhanced their confidence and sensitivity in intercultural interactions while acquiring XR design skills and operating head-mounted displays (HMDs) within 15 weeks, despite minimal prior experience. Survey responses also indicated high motivation and positive perceptions of the technology's educational value. These findings highlight the potential of immersive, co-creative environments to promote 21st century skills. By shifting learners from passive recipients to active creators, the program demonstrates a scalable model for integrating XR into international education. This study contributes to the fields of educational technology and Computer-Supported Collaborative Learning (CSCL) by evidencing the pedagogical benefits of student agency and intercultural engagement within the Metaverse.

Keywords: Metaverse in Education, International Virtual Exchange (VE), Information Literacy, Intercultural Competence, XR-Based Learning, Computer-Supported Collaborative Learning (CSCL)

1. Introduction

In the context of global education, both information literacy (IL) and intercultural competence (IC) are recognized as essential twenty-first century skills (Partnership for 21st Century Skills, 2009). Telecollaborative projects provide students with opportunities to develop IC by interacting with individuals from diverse cultural backgrounds in a digital learning environment (Çiftçi, 2016).

IL involves not only the ability to locate and evaluate information, but also the ability to use digital tools for creative production and effective communication. IC encompasses the knowledge, attitudes, and behaviors that enable meaningful and respectful engagement with culturally diverse peers. As global interconnectedness deepens, the ability to navigate digital and intercultural spaces simultaneously is becoming increasingly crucial in higher education.

In response to these demands, the Cabinet Secretariat's Council for Creating the Future of Education launched the J-MIRAI initiative to promote the internationalization of education through digital means, aiming to expand online international exchange programs nationwide (Cabinet Secretariat, Council for Creating the Future of Education, 2023). National strategies also emphasize the educational use of Metaverse technologies for future economic development. In line with these policies, our project investigates the potential of

XR-based virtual exchange programs as a scalable and innovative approach to fostering both IL and IC in global education (Hayashi, 2024).

Recent advancements in Extended Reality (XR) and Metaverse technologies have introduced new opportunities for immersive, participatory learning that transcends physical and cultural boundaries. While previous studies, such as Shadiev et al. (2023), have explored the impact of asynchronous VR-based bilateral cultural exchanges, few have examined real-time, multinational, student-generated Metaverse learning environments. To address this gap, the present study investigates a synchronous international collaboration project involving Japanese university students, international students, and overseas participants, designed to promote experiential and co-creative learning.

Through the co-creation of culturally themed Metaverse Worlds using Unity and a custom-developed Metaverse World Creation Package (MWCP), students engaged in real-time intercultural collaboration, presentation, and reflection. This study aims to explore how such immersive, student-led activities foster the development of 21st-century skills, with particular attention to IL, IC, and learners' perceptions of immersive technologies. The following research questions guide this study:

1. To what extent did the learning activity promote the development of Japanese students' information literacy (IL), particularly in terms of Metaverse World creation, head-mounted display (HMD) operation, and immersive communication?
2. To what extent did the activity facilitate the development of Japanese students' intercultural competence (IC), specifically in terms of their knowledge and sensitivity?
3. After participating, what were Japanese students' perceptions of Metaverse World creation technology and its role in international collaborative learning?

2. Literature Review

2.1 Information Literacy in XR-Enhanced Learning

IL, a foundational 21st Century Skills encompassing abilities to access, evaluate, create, and communicate information with digital tools (Shadiev et al., 2023), expands its definition in education with the prominence of XR technologies to include immersive communication and creative production in digital environments (Rauschnabel et al., 2022). While recent studies explore XR's potential in developing technological fluency, design thinking, and storytelling (Bower et al., 2020), research on student-led Metaverse World creation using advanced tools like Unity or VRChat remains limited.

In this study, IL is conceptualized, following Shadiev et al. (2023, p. 2510), as primarily "the individual's ability to locate, evaluate and use information effectively" and to "effectively communicate information through different media on various digital platforms." This is further expanded to include "learners' ability to use digital tools when collaborating with others to share information, to create media products, and to perform and complete assigned tasks utilizing digital tools effectively."

Accordingly, we interpret IL as a multidimensional competency of critical evaluation and creative expression via digital technologies. This perspective is particularly relevant in student-led Metaverse World production, where learners must combine technical, collaborative, and communicative skills.

2.2 Intercultural Competence

IC refers to the ability to interact effectively across cultures, encompassing cognitive (knowledge), affective (sensitivity), and behavioral dimensions. In this study, we adopt the conceptual framework of IC as defined by Shadiev et al. (2023), drawing on Chen and Starosta (1997) and Deardorff (2006). IC is understood as "the ability of individuals to function effectively across cultures, to think and act appropriately, and to communicate and work with other individuals from different cultural backgrounds based on their intercultural knowledge and sensitivity" (Shadiev et al., 2023, p. 2510).

According to this framework, intercultural knowledge refers to cultural information acquired through communication and exchange with individuals from other cultures, while intercultural sensitivity is defined as "an individual's ability to develop a positive emotion towards understanding and appreciating cultural differences that promotes appropriate and effective behavior in intercultural communication" (Chen & Starosta, 1997, p. 5).

This two-dimensional model provides a foundation for assessing students' development in cross-cultural competence, as applied in our study through pre- and post-surveys adapted from Shadiev et al. (2023).

2.3 Educational Use of the Metaverse in International Collaborative Contexts

In higher education, Virtual Exchange (VE) and Collaborative Online International Learning (COIL) offer practical alternatives to physical relocation for fostering intercultural skills. O'Dowd (2018) defines VE as sustained, educator-facilitated online intercultural collaboration between students in different cultural contexts.

Prior research (e.g., Shadiev et al., 2023) has shown that VR-based cultural exchanges can enhance both IL and IC. In addition, recent reviews highlight the Metaverse as an emerging learning ecology with transformative potential for education (Mystakidis, 2022). While metaverse platforms in education have traditionally been implemented through instructor- or institution-led initiatives, an increasing number of projects are now being co-created by students. These cases illustrate the growing pedagogical potential of metaverse platforms, but most serve primarily as programmatic showcases or isolated experiential learning activities without systematic assessment.

Unlike teacher-led programs where instructors design the content and guide interactions, this project adopted a student-led approach in which learners assumed responsibility for the design, content creation, and presentation of culturally themed Metaverse Worlds. This shift placed students in active agentive roles, promoting ownership of learning, intercultural negotiation, and digital creativity throughout the process.

Few educational practices exist in which students independently create Metaverse Worlds and deliver real-time, multinational presentations aimed at enhancing their peers' IC. Furthermore, empirical research measuring the learning outcomes of such student-led, immersive, synchronous intercultural activities remains limited. This study aims to fill this research gap by providing evidence of the educational impact of student-generated Metaverse environments in global collaborative learning contexts.

3. Methods

3.1 Participants and Group Composition

The program engaged 34 students from 10 countries across Asia, Africa, Europe, and North America. The participants included 18 Japanese students, 8 international students on-site in Japan, and 8 overseas students from partner institutions who joined remotely via videoconferencing tools and the Metaverse. Based on their language preferences (English or Japanese), students were divided into four groups, each with diverse national and linguistic backgrounds. Each group included students from 4 to 5 different countries and was further divided into three smaller teams. Every team was structured to include at least one Japanese student together with one international or overseas student to ensure cross-cultural collaboration. Many participants had little to no prior XR experience.

3.2 Learning Activity Overview and Instructional Support and Tools

Over a 15-week semester, students were organized into multinational teams to collaboratively create culturally themed Metaverse Worlds. Each team selected a subtopic under a shared group theme and worked together to prepare their presentation. Within their groups, teams then rehearsed their presentations and engaged in post-presentation

discussions. The process concluded with a whole class sharing session. This structure clarified the distinct roles of teams and groups, fostering intercultural understanding, mutual engagement, and reduced isolation while offering more opportunities for active participation.

Real-time collaboration, discussions and presentations were conducted in VRChat (a social VR platform featuring user-generated 3D worlds; content is built in Unity and experienced on desktop or in VR). Discord served as both a real-time and asynchronous communication hub, enabling students to exchange messages, voice chat, and share files across different time zones. The creation process was supported by Unity, VRChat Creator Companion (the official tool for creating Unity projects for VRChat), the beginner-friendly MWCP tool (Hayashi et al., 2025), and GenAI.

Instructional sessions, tutorial slides, and teaching assistant (TA) support were provided for both VRChat and Unity. MWCP enabled students with no prior experience to build interactive environments through drag-and-drop functions, making Metaverse world creation accessible. While most participants on-site used HMDs during final presentations for increased immersion, some chose desktop interfaces due to motion sickness.

3.3 Data Collection and Analysis

To assess learning outcomes, we adapted a structured questionnaire framework from Shadiev et al. (2023) aligned with the course's learning goals. Key competencies—IL, IC—were sorted into specific micro-skills, which were assessed via pre- and post-surveys using a five-point Likert scale. Quantitative data was analyzed using paired samples t-tests to examine changes in students' responses before and after the course. Internal consistency was evaluated using Cronbach's α, with all domains exceeding 0.7, indicating acceptable reliability overall. Ethical approval (Tohoku University) and informed consent were obtained.

To ensure a coherent and statistically robust analysis, this study focuses on the 18 Japanese students. Comparing them with the smaller group of international students, who had significant prior study-abroad experience, was deemed problematic due to different starting points and insufficient sample sizes for subgroup analysis. This focus aligns with Japanese educational policy priorities such as the J-MIRAI initiative and allows for a meaningful contribution to fostering global competencies among this specific cohort.

4. Results

4.1 Information Literacy Development (Results for RQ1)

To address Research Question 1 (RQ1), a pre- and post-survey was conducted among Japanese university students (N = 18) to evaluate changes in their IL across three domains: (1) use of Metaverse Worlds, (2) creation of Metaverse Worlds, and (3) operation of HMDs. Sixteen students used HMDs, while those who experienced motion sickness opted not to use them. Therefore, only Table 3 reports results based on N = 16. As the number of valid responses met the general threshold suggested by Creswell (2018), statistical analysis was performed using paired samples t-tests.

All domains showed statistically significant gains. As shown in Table 1, the average score for Metaverse World use (Items 1–16) increased from 2.01 (SD = 0.51) to 4.67 (SD = 0.10). Table 2 presents the results for world creation using Unity and VCC (Items 17–28), where the average score rose from 1.62 (SD = 0.13) to 4.38 (SD = 0.23). As shown in Table 3, the average score for HMD operation (Items 29–39) improved from 1.31 (SD = 0.12) to 4.47 (SD = 0.17). These results suggest that the activity effectively enhanced students' operational and creative digital literacy.

Table 1. *Pre- and Post-Test Results for Basic Metaverse Skills*

#	Item	Before		After		Results	
		M	*SD*	*M*	*SD*	*t*	*p*
1	I know how to run Metaverse application on PC.	1.61	1.037	4.78	0.428	12.234	< .001
2	I know how to enter a Metaverse World.	1.39	0.850	4.78	0.428	14.693	< .001
3	I know how to send Friend Requests.	1.44	0.922	4.83	0.383	15.689	< .001
4	I know how to enter a World that my Friends are in.	1.33	0.840	4.83	0.383	17.317	< .001
5	I know how to teleport to another room within the same World.	1.50	0.924	4.78	0.428	15.542	< .001
6	I know how to move around within the Metaverse World.	1.89	1.231	4.83	0.383	9.923	< .001
7	I know how to turn on my PC microphone and talk within the Metaverse World.	3.11	1.491	4.78	0.428	4.610	< .001
8	I know how to identify who is speaking within the Metaverse World.	1.67	1.138	4.78	0.428	11.662	< .001
9	I know how to react using the Metaverse World Reaction feature.	1.72	1.179	4.78	0.428	11.159	< .001
10	I know how to sit in a chair within the Metaverse World.	1.61	1.037	4.72	0.461	11.662	< .001
11	I know how to use assets that amplifies my voice to present within the Metaverse World. (holding a microphone for the metaverse, standing on a podium, etc.)	1.39	0.850	4.67	0.485	14.512	< .001
12	I know how to move to another World through a Portal placed in the World I am currently in.	2.17	1.581	4.78	0.428	7.197	< .001
13	I know how to operate gimmicks within the Metaverse World.	1.83	1.043	4.61	0.608	10.576	< .001
14	I know how to enter a World via a Metaverse application website link.	2.22	1.437	4.56	0.616	7.000	< .001
15	I know how to operate presentation slides within the Metaverse World.	1.39	0.850	4.50	0.707	12.237	< .001
16	I know how to stream videos within the Metaverse World.	1.39	0.850	4.00	1.188	7.583	< .001
	Average (1-16)	1.73	0.813	4.69	0.390	14.655	< .001

Table 2. *Pre- and Post-Test Results for Metaverse World Creation Skills.*

#	Item	Before		After		Results	
		M	*SD*	*M*	*SD*	*t*	*p*
17	I know how to import a Metaverse world package.	1.67	1.085	4.33	0.767	9.522	< .001
18	I know the basics of operating a Metaverse World creation application and the main user interface (UI). (Unity)	1.61	0.916	4.39	0.778	11.115	< .001
19	I know how to import objects.	1.67	0.907	4.56	0.616	10.829	< .001
20	I know how to change the size and rotation of an object.	1.61	1.037	4.44	0.784	9.284	< .001
21	I know how to move objects to my desired location.	1.61	1.037	4.61	0.608	9.918	< .001
22	I know how to duplicate or delete objects.	1.61	1.037	4.61	0.608	9.918	< .001
23	I know how to export the world I have created as a package so I can send it to my teachers and classmates.	1.56	0.922	4.22	0.808	9.945	< .001
24	I know how to place presentation slides, images, and other files in the world.	1.39	0.850	4.28	0.826	10.363	< .001
25	I know how to view the contents of a World file (Scene) created by a classmate using a Metaverse creation application.	1.44	0.856	4.17	0.985	9.410	< .001
26	I know how to upload the world I have created to a Metaverse application.	1.39	0.850	4.22	0.808	10.449	< .001
27	I know how to change the design of floors and walls.	1.50	0.924	4.06	1.056	8.693	< .001
28	I know how to set up gimmicks.	1.39	0.850	3.72	1.364	6.153	< .001
	Average (17-28)	1.54	0.858	4.30	0.712	11.035	< .001

Table 3. *Pre- and Post-Test Results for HMD Usage Skills*

#	Item	Before		After		Results	
		M	*SD*	*M*	*SD*	*t*	*p*
29	I know how to run Metaverse application on the HMD.	1.13	0.342	4.38	0.885	12.211	< .001
30	I know how to enter a Metaverse World while wearing the HMD.	1.13	0.342	4.50	0.632	18.781	< .001
31	I know how to move around within the Metaverse World while wearing the HMD.	1.13	0.342	4.50	0.516	21.804	< .001
32	I know how to turn on my microphone on the HMD and talk within the Metaverse World.	1.25	0.577	4.38	0.885	11.491	< .001
33	I know how to mute my microphone on the HMD.	1.38	0.719	4.50	0.816	11.491	< .001
34	I know how to identify who is speaking within the Metaverse World while wearing the HMD.	1.38	0.806	4.56	0.814	11.500	< .001
35	I know how to react using the Metaverse World Reaction feature while wearing the HMD.	1.38	0.806	4.44	0.629	13.190	< .001
36	I know how to sit in a chair within the Metaverse World while wearing the HMD.	1.19	0.403	4.50	0.632	15.174	< .001
37	I know how to move to another World through a Portal placed in the World I am currently in while wearing the HMD.	1.19	0.403	4.63	0.500	18.902	< .001
38	I know how to operate gimmicks within the Metaverse World while wearing the HMD.	1.19	0.403	4.56	0.629	16.745	< .001
39	I know how to use my hand-held controllers or head to react nonverbally (clapping, nodding, etc.) without using the Reaction feature.	1.50	0.894	4.50	0.816	9.909	< .001
	Average (29-39)	1.26	0.423	4.49	0.572	17.334	< .001

4.2 Intercultural Competence Development (Results for RQ2)

To address Research Question 2 (RQ2), a pre- and post-survey based on Shadiev et al. (2023) was administered to assess changes in students' intercultural knowledge and sensitivity. Paired samples t-tests were conducted for each item.

Intercultural knowledge improved from an average of 3.39 (*SD* = 0.75) to 4.38 (*SD* = 0.51), with all 10 items showing statistically significant increases. The most prominent gains were in learning strategies and cultural stress management. Intercultural sensitivity increased from 3.90 (*SD* = 0.41) to 4.05 (*SD* = 0.41). Subdomains, such as "interaction confidence" and "interaction engagement" showed strong gains, while other areas, such as "respect for cultural differences" and "intercultural attentiveness" remained stable. The combined average across all 34 items rose from 3.75 (*SD* = 0.48) to 4.16 (*SD* = 0.36), confirming positive development in students' IC ($p < .001$).

Table 4. *Pre- and Post-Test Results for Average Intercultural Competence*

#	Domain	Before		After		Results	
		M	*SD*	*M*	*SD*	*t*	*p*
1	Intercultural competence (Average:2-3)	3.75	0.475	4.16	0.395	5.176	< .001
2	Intercultural knowledge	3.39	0.751	4.38	0.511	6.013	< .001
3	Intercultural sensitivity (Average:4-8)	3.90	0.410	4.05	0.410	2.456	0.025
4	Interaction engagement	4.24	0.413	4.43	0.462	1.930	0.070
5	Respect for cultural differences	4.68	0.316	4.61	0.253	-0.875	0.394
6	Interaction confindence	2.91	0.749	3.32	0.777	3.297	0.004
7	Interaction enjoyment	3.28	0.909	3.33	0.792	0.419	0.681
8	Interaction attentiveness	3.85	0.698	4.07	0.479	1.799	0.090

Note: All survey items were adapted from Shadiev et al. (2023). Due to space limitations, only the average scores for each category are presented here.

4.3 Participants' Perceptions of Metaverse World Technology (Results for RQ3)

To address Research Question 3 (RQ3), a post-survey (Items 40–51) assessed Japanese students' perceptions of the Metaverse World production technology. As these items were administered only after the course, descriptive statistics were reported without comparison.

Students responded positively overall, with an average score of 4.13 (*SD* = 0.32). High satisfaction was reported for enjoyment and motivation (e.g., Item 46, M = 4.61, *SD* = 0.50). Students also expressed confidence in the technology's educational value (Item 48, M = 4.06, *SD* = 1.00) and immersive quality (Item 50, M = 4.28, *SD* = 0.83). These results indicate favorable reception of the technology and potential to support future VE initiatives.

Table 5. *Post-Survey Results for Perceptions of Metaverse World Creation Technology*

#	Item	M	SD
40	I want to continue using Metaverse World production technology.	4.11	0.900
41	I prefer Metaverse World production technology to other tools.	3.67	0.970
42	If I can, I will continue to use Metaverse World production technology.	4.11	0.832
	Continuance intention (40-42)	3.96	0.831
43	I am happy with the decision to use Metaverse World presentation.	4.33	0.686
44	I chose to use this technology wisely.	4.17	0.786
45	I am happy with my choice to use Metaverse World presentation.	4.39	0.502
46	I am very pleased with the experience of using this technology.	4.61	0.502
	Satisfaction (43-46)	4.38	0.570
47	Using Metaverse World presentation to conduct cross-cultural learning meets my expectation.	4.39	0.608
48	This technology has given me all the information and tools I need to learn across cultures.	4.06	0.998
49	My experience of cross-cultural learning through the use of Metaverse World did not meet my expectations.	4.28	0.461
50	The sense of presence provided by Metaverse World presentation meets my expectations.	4.28	0.826
51	I think Metaverse World presentation achieves the level of functionality I want.	3.89	0.963
	Confrmation (47-51)	4.18	0.609

5. Discussion

5.1 Information Literacy (Discussion for RQ1)

Japanese students demonstrated statistically significant improvements across all three IL domains; Metaverse world usage, Metaverse world creation and HMD operation. Many began with little to no XR experience and participated passively at the outset. Through structured tutorials, hands-on activities, and collaborative design tasks, students developed both technical and communicative skills. These included navigating virtual environments, interacting with 3D objects, communicating effectively in VR, and ultimately building their own Metaverse worlds.

This learning trajectory is well explained by Kolb's experiential learning theory (1984), which emphasizes the cyclical nature of learning through four stages: concrete experience, reflective observation, abstract conceptualization, and active experimentation. The design and presentation of Metaverse content allowed students to move through each phase as

they transformed individual experience into shared understanding and applied insights into their world construction process.

In addition, many students acquired XR skills despite having no background in coding or 3D design. Their progress supports Bandura's self-efficacy theory (1977), which highlights the importance of mastery experiences in building confidence. The course structure—including peer collaboration, TA support, and low-pressure environments—appears to have enhanced students' confidence and motivation.

These two frameworks—experiential learning and self-efficacy—are widely recognized as foundational theories in educational psychology for explaining how learners acquire and apply complex skills (Schunk, 2012).

Importantly, this growth aligns with the concept of participatory digital literacy, which involves not just consuming, but actively creating digital content through meaningful interaction with tools and communities (Rheingold, 2012; Bawden, 2008). In this program, students did not remain passive receivers of XR media, but became co-constructors of immersive, culturally informed digital environments.

Together, these findings indicate that experiential, collaborative design activities within XR environments can foster information literacy even among novice users, especially when supported by structured instruction and community scaffolding.

5.2 Intercultural Competence and Virtual Exchange (Discussion for RQ2)

Statistically significant improvements were observed in students' IC, particularly in the domains of Intercultural Knowledge and Intercultural Sensitivity, thus addressing RQ2. Among the five subcomponents of Sensitivity, the most substantial gain was found in Interaction Confidence, notably in the item "I can communicate with people from different cultures very easily."

These improvements can be attributed to a sequence of immersive, collaborative learning experiences in the XR-based Metaverse environment. Notably, the largest pre-post increase among Intercultural Knowledge items was found in "I know some techniques to help me learn the target culture." This suggests that the course supported students in acquiring learning strategies essential for navigating unfamiliar cultural contexts.

This gain was facilitated by two key learning activities; 1) Cultural comparison and conceptualization: Participants selected a topic and engaged in critical comparison between their home and target cultures, guided by intercultural collaboration and 2) Collaborative world creation for peer learning: In mixed-nationality teams, students designed and built culturally themed Metaverse Worlds to provide classmates with new knowledge and stimulate deeper reflection and discussion. These activities were enhanced by authentic cultural materials, made accessible through real-time input from overseas participants. Such materials offered culturally grounded perspectives that strengthened content relevance and engagement.

Furthermore, learners explored the target culture through immersive embodied experiences using HMDs, interacting with grabbable content within the Metaverse. Through these dynamic encounters, they engaged in repeated cycles of intercultural interaction, real-time feedback, and reflection, which promoted deepened awareness and adaptability.

The marked improvement in Interactional Confidence is closely linked to the unique affordances of the XR environment. Unlike asynchronous exchanges, this setting provided synchronous, real-time interaction and responsive engagement, helping students gain communicative fluency and intercultural responsiveness. In contrast to traditional video conferencing, the Metaverse enables embodied communication through avatar gestures, spatial positioning, and shared virtual spaces.

This social presence fostered emotional connectedness and reduced psychological barriers, leading to greater confidence and increased willingness to engage. This outcome aligns with the notion of the affective filter (Krashen, 1982), in which emotional safety and motivation lower anxiety and facilitate language and cultural learning.

These gains are well explained by Kolb's (1984) experiential learning theory, in which learners progress through concrete experience, reflective observation, abstract

conceptualization, and active experimentation. The iterative process of designing and presenting Metaverse Worlds aligns with this cycle, offering students authentic opportunities to internalize and apply intercultural knowledge and strategies.

Moreover, the observed improvement in confidence and engagement reflects the development of self-efficacy—a learner's belief in their own capabilities. While Bandura's (1977) theory emphasizes the role of mastery experiences, Schunk (2012) expands on this by situating self-efficacy within self-regulated learning, motivation, and performance in complex educational settings.

Finally, the observed improvements in affective and behavioral aspects—such as confidence, engagement, and attentiveness—underscore the value of immersive and synchronous Metaverse interactions. These findings align with Slater & Wilbur (1997) and Chen & Starosta (1997), who highlight social presence and mutual attentiveness as key to building trust and openness in intercultural communication.

5.3 Perceptions of Metaverse World Production Technology (Discussion for RQ3)

The post-activity survey results addressed RQ3 and revealed strong acceptance and satisfaction among Japanese participants with Metaverse World production technology. Most students rated the experience as enjoyable and valuable, citing increased motivation and perceived educational relevance (Table 5). These results suggest that students' engagement was deeply connected to the autonomy and creativity afforded by the learning task. Deci and Ryan's (1985) Self-Determination Theory explains that when learners feel competent, autonomous, and connected, their intrinsic motivation increases. The process of designing, collaborating in teams, and publicly presenting their own creations likely fulfilled these three basic psychological needs—autonomy, competence, and relatedness—thereby fostering intrinsic motivation.

High ratings for presence (Item 50) and intention to continue using the technology (Items 42–43) indicate strong engagement. This aligns with Slater and Wilbur's (1997) concept of presence as psychological immersion. The observed coherence between instructional design and perceived value also supports Biggs' (1996) theory of constructive alignment. Davis' (1989) Technology Acceptance Model (TAM) further explains that users are more likely to adopt technology when they perceive it to be useful and easy to use. Students' continued willingness to engage with Metaverse technology supports this model.

However, our findings also suggest that these two factors—usefulness and ease of use—are not sufficient on their own in educational XR contexts. Despite encountering some usability challenges, students overcame technical barriers with support and still expressed strong intention for future use. Rather, task designs that allow learners to exercise agency and creativity (e.g., student-led world creation) play a critical role in enhancing both satisfaction and long-term adoption intentions. This suggests the need to expand the TAM framework to include motivational dynamics such as intrinsic motivation.

Taken together, the results of RQ3 indicate that creating Metaverse Worlds in international learning contexts fostered emotional engagement, technological confidence, and self-directed learning. As shown in Table 5, students reported high levels of enjoyment (Item 46), meaningful learning experiences (Item 48), and motivation to use the technology again (Items 42–43). These findings underscore the importance of integrating learner autonomy and creative control into XR-based instructional design, especially in intercultural and collaborative learning environments where sustained engagement is crucial.

5.4 Limitations and Future Directions

This study demonstrates that student-led creation of Metaverse Worlds in XR environments effectively fosters both IL and IC, offering new perspectives on participatory learning by engaging students as content creators rather than passive users. Based on these findings, we recommend: (1) aligning clear learning objectives with structured technical support, (2) designing open-ended, collaborative tasks to promote student agency and creativity, and (3) incorporating reflective activities to deepen intercultural understanding. For XR platform

developers, usability, real-time collaboration, and culturally diverse digital assets, such as virtual objects and environments, are important considerations. While teaching assistants supported both technical and intercultural learning, future work should explore AI-driven guidance and feedback to enhance scalability and learner autonomy.

However, this study is subject to several limitations. The small number of international participants limited cross-cultural comparisons and reliance on self-reported data may introduce bias. The one-semester duration was not sufficient to evaluate long-term outcomes and qualitative data collected were not analyzed here. Future research should include broader international cohorts, longer study durations, and comparisons of different XR platforms and activity types. In addition, exploring AI tutors, peer support systems, and learning analytics could deepen our understanding of learning processes. Analysis of the qualitative data collected will further clarify how XR-based, student-led environments contribute to IL and IC development in global education.

The findings of this study suggest potential applicability beyond the current higher education setting, particularly in the internationalization of primary and secondary education—as emphasized in the third pillar of the J-MIRAI initiative—and in line with OECD's strategic priorities for effective and inclusive digitalization and the role of higher education in upskilling and reskilling (OECD, n.d.). Although technical and institutional challenges remain for implementing XR and Metaverse-based approaches in these wider educational contexts, we believe that such barriers can be gradually addressed through the accumulated pedagogical expertise developed in this project and our ongoing efforts to integrate generative AI into educational support systems.

6. Conclusion

This study demonstrates that student-led, collaborative Metaverse world creation in synchronous VE enhances IL and IC among students. Quantitative results revealed significant gains in technical, communicative, and cultural competencies, confirming the effectiveness of this approach. In addition to measurable improvements, students reported high enjoyment, motivation, and positive perceptions of the immersive learning environment, emphasizing the intrinsic value of learner agency in XR-based education.

Beyond these findings, this research makes a novel contribution to the fields of educational technology and CSCL by showing how student-led creation in XR environments can concurrently foster the development of 21st-century skills. It advances ongoing scholarly discussions on how immersive, collaborative environments can integrate skill development with intercultural learning.

From a practical perspective, the successful implementation of MWCP underscores the importance of accessible and adaptable tools in making advanced XR content creation achievable for learners with varying technical backgrounds. As global education seeks scalable and engaging internationalization strategies—such as Japan's J-MIRAI initiative—our findings provide a strong, evidence-based model for incorporating immersive XR technologies into curricula. Furthermore, these implications resonate with OECD's policy priorities for inclusive digitalization and upskilling in higher education, reinforcing the relevance of this study in the broader context of international education policy.

Taken together, this study validates the educational value of student-created Metaverse environments and provides a robust foundation for advancing future research on learning in dynamic virtual spaces.

Acknowledgements

This research was partially supported by grants from the Japan Society for the Promotion of Science (JSPS, JP23KK0037, JP24K06075) and The Open University, as well as through a joint research project between the Research Institute of Electrical Communication at Tohoku University. I would also like to extend my gratitude to VRChat, Inc. (https://hello.vrchat.com/),

the study participants, and to the teaching assistants, including Abe Manato, Suzuki Koshiro, Alisa Yoshida Belotti, Umemura Yoshiki, Saito Kairu, and Caleb Paice. Special thanks are also due to everyone who contributed to the preparation of this paper.

References

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191–215.

Bawden, D. (2008). Origins and concepts of digital literacy. *Computer Science, Education*.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, *32*(3), 347–364.

Bower, M., DeWitt, D., & Lai, J. W. M. (2020). Reasons associated with preservice teachers' intention to use immersive virtual reality in education. *British Journal of Educational Technology*, *51*(6), 2214–2232.

Cabinet Secretariat, Council for Creating the Future of Education. (2023, April 27). *J-MIRAI: Japan–Mobility and Internationalisation: Re-engaging and Accelerating Initiative for future generations (2nd proposal)*. Cabinet Secretariat.

Chen, G. M., & Starosta, W. J. (1997). A review of the concept of intercultural sensitivity. *Human Communication*, *1*, 1–16.

Çiftçi, E. Y. (2016). A review of research on intercultural learning through computer-based digital technologies. *Journal of Educational Technology & Society*, *19*(2), 313–327.

Creswell, J. W., & Guetterman, T. C. (2018). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (6th ed.). Pearson.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340.

Deardorff, D. K. (2006). Identification and assessment of intercultural competence as a student outcome of internationalization. *Journal of Studies in International Education*, *10*(3), 241–266.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic Motivation and Self-Determination in Human Behavior*. Springer.

Hayashi, M. (2024). The challenges of connecting the world through VR and the metaverse in international collaborative learning. In M. Hayashi (Ed.), *The Metaverse and XR Technology – Educational Applications and International Collaboration*, Tohoku University Press. 「VR・メタバースで世界をつなぐ国際協働学修の挑創」『メタバース・XR技術の教育利用と国際協創』

Hayashi, M., Abe, M., & Hayama, H. (2025). The benefits and impact of introducing a metaverse world creation package in international virtual exchange. In *Proceedings of the 11th International Conference of the Immersive Learning Research Network (iLRN 2025)*.

Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development*. Prentice-Hall.

Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Pergamon Press.

Mystakidis, S. (2022). Metaverse. *Encyclopedia*, *2*(1), 486–497.

O'Dowd, R. (2018). From telecollaboration to virtual exchange: State-of-the-art and the role of UNICollaboration in moving forward. *Journal of Virtual Exchange*, *1*, 1–23.

OECD. (n.d.). *Higher Education Policy*. Retrieved August 15, 2025, from https://www.oecd.org/en/about/programmes/higher-education-policy.html

Rauschnabel, P. A., Felix, R., Hinsch, C., Shahab, H., & Alt, F. (2022). What is XR? Towards a framework for augmented and virtual reality. *Computers in Human Behavior*, *133*, Article 107289.

Rheingold, H. (2012). *Net smart: How to thrive online*. MIT Press.

Schunk, D. H. (2012). *Learning theories: An educational perspective* (6th ed.). Pearson.

Shadiev, R., Dang, C., Sintawati, W., Yi, S., & Huang, Y. M. (2023). Facilitating information literacy and intercultural competence development through the VR tour production learning activity. *Educational Technology Research and Development*, *71*(6), 2507–2537.

Slater, M., & Wilbur, S. (1997). A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, *6*(6), 603–616.

The Partnership for 21st Century Skills. (2009). *P21 framework definitions*. https://files.eric.ed.gov/fulltext/ED519462.pdf

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Understanding the Triggers and Dynamics of Socially Shared Metacognitive Regulation in Teams with Divergent Outcomes

Vishwas BADHE[a*], Ramkumar RAJENDRAN[a] & Chandan DASGUPTA[b]
[a]*Centre for Educational Technology (CET) IIT Bombay, India*
[b]*Department of Learning, Data Analytics, and Technology, University of Twente,* Netherlands
*vishwasbadhe@iitb.ac.in

Abstract: Collaborative problem-solving (CPS) offers opportunities for team members to apply their acquired skills and knowledge to shared tasks, particularly when addressing ill-structured problems that require collaboration among individuals with diverse perspectives and competencies. During CPS, learners may encounter various challenging situations referred to as trigger events that serve as markers or catalysts for strategic regulation. Previous studies on socially shared metacognitive regulation (SSMR) have primarily focused on the occurrence of SSMR during collaborative learning, with limited attention to its triggering factors. This study investigates the triggers and dynamics of SSMR in teams with divergent outcomes (i.e. more successful outcome teams (MSOT) and less successful outcome teams (LSOT)). Conducted over 12 weeks, the research focused on ill-structured tasks in a project-based Human-Computer Interaction course. We analyzed 35 hours of video data using thematic analysis, examining teams' verbalized interactions to identify SSMR triggers, followed by coding the identified SSMR episodes to determine their focus and function. Findings revealed differences between MSOT and LSOT in terms of the variety of triggers, focus, function, and the shifts in SSMR strategies. MSOT demonstrated a fundamental focus that enabled teams to grasp essential aspects needed to solve the task, whereas LSOT exhibited surface-level focuses, often involving non-essential components. MSOT applied SSMR strategies optimally and at appropriate times during CPS, while LSOT displayed suboptimal and untimely applications. Moreover, strategy formulation and execution in MSOT were consistent and goal-directed, whereas in LSOT, they were interruptive and momentary.

Keywords: Socially shared metacognitive regulation (SSMR), Collaborative problem-solving (CPS), Project-based learning (PBL), ill-structured problem-solving, Triggers of SSMR, Focus of SSMR, Function of SSMR.

1. Introduction

Collaborative Problem Solving (CPS) in Computer-Supported Collaborative Learning (CSCL) environments provides team members the opportunity to apply their existing skills and knowledge toward accomplishing a shared task. As learners often come from diverse socio-cultural backgrounds, they contribute differing goals, perspectives, attitudes, and experiences; factors that become critical and dynamic during CPS. Managing this diversity effectively while making meaningful progress on a task requires various socially shared regulation strategies (Järvelä and Hadwin, 2024).

Within CPS, team members regulate their cognition, metacognition, motivation, emotions, and behaviors through shared metacognitive monitoring (Järvelä et al., 2013). Yet, cognitive and metacognitive challenges often arise from differing interpretations of the task or content. In such cases, shared metacognition is essential for collaboration, as it enhances awareness of challenges through monitoring and signals the need for regulation. Learners employ various shared regulation strategies, including monitoring, controlling, planning, and reflecting (Lobczowski et al., 2021). Socially Shared Metacognitive Regulation (SSMR)

specifically refers to processes where group members jointly monitor and guide collective cognitive efforts (Iiskala et al., 2015). SSMR is crucial in CPS for aligning group cognition through continuous monitoring and control. These processes include identifying goals and expectations (e.g., what needs to be done), planning (e.g., time allocation), tracking progress, adapting strategies, monitoring understanding (e.g., questioning reasoning), and evaluating the final output (Badhe et al., 2022; Iiskala et al., 2015; Kerrigan et al., 2021).

The selection of more and less successful outcome teams (MSOT and LSOT) in this study is grounded in the theoretical premise that socially shared metacognitive regulation (SSMR) contributes meaningfully to collaborative learning. Preliminary data analysis supports this premise. SSMR includes setting shared learning goals, mutually monitoring comprehension and progress, and reflecting collaboratively on both the process and outcomes (Järvelä et al., 2013). When paired with domain-specific knowledge, these processes positively impact individual learning gains (Chan, 2012). Though theoretical frameworks link SSMR to deep understanding and high-level outcomes (Iiskala et al., 2011, 2015), empirical evidence remains limited. Nonetheless, it is assumed that quality SSMR supports group progress by guiding cognitive control toward shared goals (Iiskala et al., 2021). Thus, this study examines MSOT and LSOT to explore how the dynamics of SSMR correlates with collaborative learning success. SSMR's role is especially critical when learners engage in complex, ill-structured tasks (Iiskala et al., 2021). Research (Iiskala et al., 2004) shows that metacognitive regulation tends to emerge more in such tasks than in well-structured ones. Additionally, task difficulty significantly influences metacognitive activation, as demanding tasks intensify metacognitive experiences (Iiskala et al., 2011; Efklides et al., 1998). However, prior studies mainly focused on how SSMR manifests, not what triggers it (Vauras, Volet, & Iiskala, 2021). To address this gap, this study investigates the triggers of SSMR during a semester-long, ill-structured project and explores its focus and function within CPS.

Trigger events are circumstances that disrupt the learning process, requiring regulatory action (Järvelä & Hadwin, 2024). They act as challenges that prompt learners to apply strategic responses. Studies identify three main SSMR triggers: metacognitive experiences (ME), socio-cognitive conflict (SCC), and error feedback (EF). Metacognitive experiences refer to internal thoughts or feelings that raise awareness about one's thinking or task approach (Efklides, 2006). These include metacognitive feelings (e.g., difficulty, confidence, satisfaction) and judgments or estimates. Both play a key role in monitoring and regulating task performance. For example, feeling that a task is difficult may prompt strategic intervention (Vauras, Volet, & Iiskala, 2021). Cognitive conflict occurs when current understanding is challenged by new or conflicting input. In group settings, this often appears as socio-cognitive conflict—discussions of differing views. Such interactions push learners to reconcile opposing ideas and promote deeper understanding. Lastly, error feedback, which signals incorrect responses, is another common trigger. Feedback prompts learners to reevaluate and adjust their strategies accordingly (Vauras, Volet, & Iiskala, 2021). Overall, trigger events serve as markers for regulation, prompting learners to adjust their motivational, cognitive, emotional, or behavioral strategies to enhance performance (Järvelä & Hadwin, 2024).

Hence, it is important to scrutinize the role of various triggers of SSMR and the corresponding regulatory responses of teams across different contexts. In this study, we investigated the types of triggers that occurred in more and less successful outcome teams, along with the differing responses of these teams to various trigger situations. This longitudinal study was conducted in a semester-long course involving an ill-structured design challenge. The research question (RQ) guiding this investigation is: How do the triggers and responses of SSMR differ between more and less successful outcome teams during ill-structured CPS?

2. Methodology

To investigate the triggers and dynamics of SSMR of teams, we chose an authentic semester-long course in Human-Computer Interaction for educational technology (HCI for ET). The course used a project-based learning pedagogy, which involved four milestones (see table 1) and many CPS tasks. Hence the study was conducted over 12 weeks in a

graduate-level, face-to-face HCI for ET course in a collaborative classroom setting during fall 2022. To achieve heterogeneous grouping, a total of sixteen learners participated in this study. The group consisted of five Ph.D. students, one Master's student, and ten Bachelor's-level students (Mean age = 23.4 years, SD = 4.09; 65% Male, 35% Female). None of the participants knew each other before the course. The learners were divided into four teams, each with four members. All teams included a mix of Ph.D. and Bachelor's-level learners, while the single Master's learner was assigned to Team 3. All sixteen learners were graduates from the engineering, architecture, or science domains, broadly aligning with the STEM field. The course followed a project-based learning approach in which the following ill-structured problem (design challenge) was given to all the teams - “Design an intervention that supports special needs education (formal/informal) for speech and hearing impaired (DHH: Deaf or Hard of Hearing) students”. All teams worked towards designing a solution for the given open-ended problem statement throughout the semester.

Table 1. *Set of Milestones and Subtasks given for Teams in Project-Based Learning HCI Course*

Milestones	Task Name
1	Understanding problem & user needs using concept mapping, literature review
	Data gathering using interviews (on-field task)
	Problem definition using a fishbone diagram
	Analysis of user needs using empathy maps and user persona
2	Ideation for the design solution
	Study of existing systems
	Finalizing one idea using a decision matrix
3	Developing low-fidelity prototype
	Mapping prototype with the problem statement and theories
	Checking adherence to learned design principles with prototype
4	Evaluation of prototype with testing matrix and heuristics
	Refinement of prototype based on evaluation

After basic orientation, the instructor announced the ill-structured design challenge in class. The semester-long course was divided into four major milestones (See table 3) leading to the final solution. Each task spanned approximately 3 weeks with predefined deliverables contributing to the final solution. For each week, learners were having two 1.5-hour-long in-class sessions. Each team was given the opportunity to collaborate and work on ill-structured problems for a total 8.5 hours in 7 weeks. Each session consisted of the following - (a) half an hour of instruction covering required concepts, tasks, deliverables, and resolving doubts, and (b) one hour for teamwork at the team’s dedicated collaborative space (round table). During teamwork, learners discussed the design challenge and task strategies face-to-face and simultaneously documented their progress using the ConceptboardTM platform - a collaborative whiteboard enabling distributed teams to work together - and shared Google Document which contained their design journal. The instructor and TAs visited the teams at their tables for collecting feedback and addressing clarification questions. The course readings corresponding to each week and task were shared with the learners a week prior to the instruction. Learners were briefed about the tasks, associated activities, and deliverables each week as per the weekly course plan.

At the start and end of each milestone, team members were asked to do collective planning and evaluation. This facilitated metacognitive regulation opportunities for the teams while working collaboratively in each milestone. At the end of each milestone, teams were asked to present their team progress to the entire class. They were instructed to log their progress in shared group journals asynchronously (reflecting groups’ status and individual contribution).

2.1 Data Collection

The data was collected for the four teams and prior consent was taken. The verbal interaction of collaborating team members was video recorded, and the milestone-wise deliverable (performance) was evaluated using the rubric. Learners also worked synchronously and asynchronously outside regular class times, but that part was not recorded. ConceptboardTM board activity screenshots and shared group journals for teams were also collected. However, the solution they have developed, their write ups in group journals, written responses in planning, and evaluation should have been factored in the data analysis.

2.2 Data Analysis

We evaluated the team performance associated with each task using a self-evaluation rubric which was shared with teams in the course orientation. All teams were first evaluated task-wise, and then the total score was calculated by summing up the task-wise scores. The tasks were grouped logically into different milestones. While doing the task-wise evaluations using a rubric, we have also considered the team's ConceptboardTM screenshots and their shared group journal to validate the work done. This rubric had been shared with all the teams ahead of time. Out of four teams, two teams were placed in a more successful outcome team (MSOT), and two were placed in a less successful outcome team (LSOT). The MSOT (Team 1 and Team 3) scored 9 and 8.14 out of 10, respectively, whereas the LSOT teams (Team 2 and 4) scored 5 and 5.28 out of 10 respectively. We then sampled video data of the first two milestones (first 7-8 weeks) of all four teams. In the first two milestones, various opportunities were given to all teams to decide problem statements and decide probable solution ideas, which were more challenging and involved substantial amounts of brainstorming and decision-making. The first two milestones allowed learners to put forward their thought processes more openly. To investigate the triggers and dynamics of SSMR, we analyzed the video data (of 35 hours) from a synchronous face-to-face classroom interaction.

Table 2. *Deductive Coding Scheme Followed while Analyzing SSMR and the Degree of Transactivity for Both Teams*

Particular	Subtype	Description
Triggers of SSMR (Vauras et al., 2021)	Metacognitive Experience (ME)	Metacognitive experiences are person's subjective cognitive or affective experiences that monitor and inform a person about a feature of cognitive processing in relation to the task at hand
	(Socio) Cognitive Conflict (SCC)	Based on Piagetian conceptualization of (dis)equilibrium, cognitive conflict refers to a discrepancy between a person's own view of the world and new information that conflicts with the person's existing view.
	Error Feedback (EF)	Students receive feedback from each other during the collaborative learning process. Error feedback may also affect students' metacognitive behavior.
Focus (Grau and Whitebread, 2012)	Fundamental	Refers to essential aspects discussed to solve the task. It is always related to the final goal of the task. it could include or not include discussions about knowledge.
	Surface	Refers to non-essential aspects of the task, such as time management, choice of resources, etc. They are relevant to complete the task; however, the way this is done does

		not have a great influence on the quality of the outcomes.
Function (Iiskala, 2011; De Backer, 2022)	Activate	Activating a new direction for ongoing interaction or a new way of thinking in line with and building upon previous activity
	Confirm	Confirm ongoing interaction, eliciting a continuation of previous activity in the same direction
	Change	Changing the flow of collaborative learning, implying ongoing interaction is challenged and current activities are questioned and rethought to the extent that an alternate direction is taken

All four teams have chosen the problem statements around the proposed themes. The thematic analysis of content analysis approach (Mayring, 2015, p. 95) was followed to analyze students' verbalized interactions during collaborative work (including teams' interactions with instructor or TAs). The verbal interactions during CPS were video recorded and transcribed for data analysis. The triggers of SSMR were coded using the description given in the literature (i.e. ME, SCC, and EF). The start points of conversational segments marked by shared metacognitive experiences were identified as trigger events. The endpoint was marked by the last conversational turn on the topic or the emergence of a new trigger (Iiskala et al., 2011). Segments were considered SSMR episodes if they included verbalizations of monitoring and controlling cognitive processes (De Backer et al., 2022). Each episode contained multiple conversational turns by team members. After identifying SSMR episodes from the video data, we coded the triggers, focus, and function of each SSMR episode using the coding scheme shown in table 2.

SSMR statements were coded by a trained educational technology researcher. The reliability of statement coding was assessed using Cohen's kappa (κ) as a measure of inter-rater reliability. In the first round, two independent educational technology researchers coded 20% of the SSMR statements (210 statements), followed by discussions to resolve disagreements and establish consensus in the second round. Both researchers were well-versed in the metacognition and CPS research area. Inter-rater reliability was assessed using Cohen's kappa (κ). The coding of SSMR skills, focus and function demonstrated a high level of agreement between coders (κ = 0.85). These values indicate reliable coding procedures across constructs, with overall kappa values suggesting substantial inter-rater reliability (Landis & Koch, 1977).

3. Findings

This descriptive statistics and findings for the RQ are presented in this section.

3.1 Descriptive statistics

Table 3 presents descriptive statistics for SSMR episodes across four teams, revealing diverse patterns in metacognitive regulation.

Table 3. *Descriptive Data Showing Information about Team-wise Total Episodes*

	Team 1	Team 2	Team 3	Team 4
Number of Episodes (100)	26	17	31	25
Total Episode Duration	52.43	49.35	77.22	77.28
Min	0.24	0.11	0.17	0.15
Max	8.56	8.7	11.1	10.27

Mean	1.85	2.94	2.30	3.06
SD	2.16	2.5	2.35	2.40

Note: All data is in Minutes (mm.ss format)

The 100 total episodes varied significantly among teams (17 to 31), with notable differences in total duration and average episode length. Episode lengths ranged from 0.11 to 11.1 minutes, with high standard deviations indicating substantial within-team variation. These findings highlight diverse SSMR engagement patterns in terms of frequency, duration, and consistency during collaborative problem-solving.

Table 4 gives team-wise information on triggers of SSMR along with different characteristics of the SSMR episodes (such as focus and function). Further descriptive statistics are shown in the following table 4, which shows the instances of SSMR episodes, triggers, focus, and the function of SSMR episodes. Table 4 groups teams 1 and 3 as more successful outcome teams (MSOT) and teams 2 and 4 as less successful outcome teams (LSOT).

Table 4. *Descriptive Data Analysis for Coding Team-wise Episodes (Showing Triggers, Focus, and Function of SSMR)*

		Team 1	Team 2	Team 3	Team 4	MSOT (Team 1 & 3)	LSOT (Team 2 & 4)
Number of Episodes		26	17	31	25	58	42
Triggers	Metacognitive Experience (ME)	23	14	25	16	NA	NA
	Socio-Cognitive Conflict (SCC)	2	0	4	4	NA	NA
	Error Feedback (EF)	1	3	2	5	NA	NA
Focus	Fundamental	19	8	27	19	46	27
	Surface	7	9	4	6	11	15
Function	Activate	7	6	9	18	18	24
	Confirm	6	8	7	0	13	8
	Change	13	3	15	7	28	11

3.2 Findings from thematic analysis

The triggers and the dynamics of the SSMR (in terms of focus and function) for MSOT and LSOT are presented in Table 4. Furthermore, the thematic analysis helps unpack the differences between the focus and function of SSMR. The following themes emerged from the thematic analysis.

Theme 1: Differentiating focus of SSMR in trigger situation

Differences in SSMR responses were observed between MSOT and LSOT when experiencing SSMR triggers during CPS. The findings highlighted differences in the focus of SSMR among the teams. It was observed that the fundamental focus acquired by MSOT enabled teams to understand the essential aspects required to solve the task, whereas LSOT exhibited surface-level focuses, which are not desirable, as team members engaged in pragmatic or non-essential components of the task. This finding aligns with Iiskala et al. (2011) and Grau & Whitebread (2012). Examples of differentiated SSMR responses and their corresponding focus for similar types of trigger situations are presented in Table 5.

Table 5. *For the SCC Trigger, the Difference in Type of SSMR Response and Focus is given for MSOT & LSOT*

Team 3 (MSOT)	Team 4 (LSOT)
(L1) M4- There is special education for concept learning.. (L2) M1- Special education is for reading, writing and .. (L3) M2- No, Special educators work separately. .. in early years of education ... (Explains that this is how special educators work and for this reason..) (L4) M4 & M1- Ok (Agrees by nodding)	(After looking at the updated problem-statement and realizing that this PS is an updated one and he was not aware about it). (L1) M2- When did we update the problem statement ? (L2) M1- (Smiles).. (L3) M3- What? (L4) M1- M2 is asking when we updated the problem statement. (L5) M3- (Smiles).. Simultaneously.. in the process.
SSMR skill: Monitoring Comprehension Focus: Fundamental Metacognitive strategy: content understanding through Highlighting a discrepancy	SSMR skill: Monitoring Comprehension Focus: Surface Metacognitive strategy: Content understanding through Social Questioning

The episodes presented in Table 5 illustrate how, for a similar type of trigger (i.e., SCC), the SSMR skill, focus, and metacognitive strategy differed between MSOT (team 3) and LSOT (team 4). In response to the SCC trigger, both teams began with monitoring comprehension. However, member M2 (at L3) from team 3 identified a discrepancy in understanding and actively attempted to regulate the content comprehension of other team members. In contrast, in team 4, M2 (at L1) recognized that he had missed the updated content knowledge related to the problem statement and was unaware of the revised version. He externalized his cognitive conflict through social questioning, but other team members did not acknowledge this opportunity to resolve the ambiguity in the shared problem statement or align everyone's understanding. As a result, team 4 ended with a surface-level focus, while team 3 achieved a fundamental focus. This demonstrates the differentiated regulative responses by team 3 and team 4. De Backer (2022) suggests that diverse responses to trigger conditions can lead to varying degrees of facilitation for SSMR, which supports this finding.

Theme 2: Timing and functioning of SSMR in triggering situations

The SSMR responses to the ME trigger also varied between team 1 (MSOT) and team 2 (LSOT). For team 2, the trigger was ME–feeling low satisfaction, whereas for team 1, it was ME–metacognitive judgments/estimates of learning, which stimulated the subsequent SSMR episodes. MSOT exhibited optimal application of SSMR strategies at the appropriate time during CPS, whereas LSOT demonstrated suboptimal application of SSMR strategies.

Table 6. *For the ME Trigger, the Difference in Type of Shared Regulation Response and Focus is given for More and Less Successful Outcome Teams*

Team 1 (MSOT)	Team 2 (LSOT)
(L1) M2- Do we detect and correct errors? (L2) M1- Did we detect something? (L3) M3- and what would be the errors ? (L4) M1- Yeah. (L5) M4- Error means it was just an improvement. (L6) M3- I mean we discuss individual thoughts on what we want to do and	(L1) M3- We also don't meet if we miss the class, and to catch up and update absent person.. (L2) M1- See.. you people (who miss class frequently) have to become more particular, just I have done my part on time, but you people were doing it very late night, I was observing that live on Document.. but you didn't inform.. (L3) M1- Just like, see, M2 have also not done respective parts.. now I can't fill their columns in this

based on discussion we decide. (L7) M2- In our group even if we had contradiction still (after making consensus) we build upon idea so we just don't just reject anyones' idea, (L8) M1- But what to mark here. Neutral..! we are not sure. (L8) All- Yeah. Neutral	sheet right.. (Unpleasant Feeling of (low) satisfaction). and I have prepared a whole structured sheet for us, and just thought let me take up that responsibility and do it. but you people have to at least put your ideas in it... (L4) M3 - Hmm (Yes) (L5) M1- If you want to meet then just schedule and fix the meeting, lets work. just don't say that we don't meet.. this doesn't work..
Major MRS: Process Evaluation Focus: Fundamental Metacognitive strategy: Evaluating Task performance Function: Facilitate- Confirm	Major MRS: - Monitoring Progress - Strategic planning Focus: Fundamental Metacognitive strategy: Planning Task Performance Function: Facilitate- Activate

The difference in the shared regulation responses is evident in the two episodes from both teams (Table 6). In team 1, at L6, the response from M3 indicates that all members contributed to the discussion, and individual thoughts were discussed collectively. The following response at L7 by M2 shows that the team had a shared understanding of how to handle contradictions or conflicts in opinions and how to align perspectives to achieve a common goal. Members of team 1 evaluated their collaboration effectively and confirmed their overall strategy for managing contradictions during teamwork. In contrast, in team 2, at L1, M3 expressed difficulty with group coordination, particularly when members missed class and did not meet outside of class. In response, M1 at L2 suggested that if everyone became more task-oriented, group coordination would improve. Members of team 2 monitored the team's overall progress and proposed a remedial strategy to address coordination issues after recognizing a shortfall in their performance. Here, we can observe that team 1 regulated task-related challenges by achieving shared understanding at the desired time during CPS, whereas team 2 addressed group coordination challenges by proposing remedial actions at an undesired time (i.e., at the end of Milestone 2). Although both teams ultimately achieved a fundamental focus, the distinction lies in the timing and optimal application of SSMR strategies during CPS. These findings align with De Backer (2022) and Vauras, Volet, & Iiskala (2021).

Theme 3: Nature of SSMR strategy

Differences in formulating and adapting task strategies were observed between MSOT and LSOT. In MSOT, strategy formulation and execution were consistent and goal-directed, whereas in LSOT, they were interruptive and momentary. In team 1 (MSOT), while discussing and refining the final goal, the conversation triggered an SSMR episode through a metacognitive judgment about the final goal. In one such episode, M2 identified a discrepancy in the team's goal, saying, *"M2 – I think in sign language the sentence is.. semantics.. but the challenge here would be contextualizing, no matter what the language is.."* This prompted M3 to experience cognitive conflict and regulate M2's interpretation of the perceived goal by responding, *"M3 – So, here we are saying that it's possible to match, but would it be better to say, given the understanding of sign language sentence construction, we need to teach them English language's sentence construction?.."* This exchange was seen as a goal-directed approach, after which the team began working toward the finalized goal. This indicates that team 1 (MSOT) adopted consistent and goal-directed SSMR during CPS. In contrast, team 2 (LSOT), while evaluating the task using the evaluation sheet, saw M1 state, *"M1 – I have updated everything in the task evaluation sheet (without discussing with all team members). If you want to update, then you can..."* The other members accepted this without objection, even though they were aware that the evaluation sheet was intended to help teams identify gaps in their task strategy and address them collaboratively. This indicates that team 2 (LSOT) adopted a momentary and unfocused SSMR approach during CPS. These results indicate that MSOT and LSOT exhibited diverse regulation responses to trigger conditions.

4. Discussion

The present study aimed at investigating various triggers conditions experienced by teams and their corresponding SSMR response for MSOT and LSOT. The limited research on the triggers of SSMR presents practical challenges, as it hinders the ability to effectively support activities that could foster SSMR during collaborative learning (Vauras, Volet, & Iiskala, 2021). The findings have highlighted occurrences of three types of triggers of SSMR (i.e. metacognitive experience (ME), socio-cognitive conflict (SCC), and error feedback (EF)) for MSOT and LSOT. The SSMR response in terms of focus and function showed the highlighting differences in those teams. Difference in the focus was observed in the data such as varied focus while creating understanding, performing operations, and producing outcomes. It is important to note that the level of cognitive effort and level (surface vs. fundamental) on which the group works and discusses during CPS may determine the magnitude of the effect of SSMR but it remains within the scope of learners' competence. The findings highlight meaningful differences in focus between MSOT and LSOT, and suggest the need for targeted support for teams in challenging situations.

Further, the evidence on function of SSMR indicates that MSOT and LSOT differ in terms of timing and adequacy of the SSMR during CPS. If the functions of SSMR are ill-timed or inadequate during regulation, they may significantly impact the effectiveness of SSMR, regardless of how frequently regulatory processes occur (Badhe et al., 2022; Vauras, Volet, & Iiskala, 2021). The differences with respect to the nature of the SSMR shows that in MSOT, strategy formulation and execution was consistent and goal-directed, whereas in LSOT, it was interruptive and momentary. The interruptions and abrupt shifts in the SSMR strategies may hamper collective discussions and outcomes. This study adds value to SSMR research by revealing different types of trigger situations experienced by teams with divergent outcomes, and by highlighting the nature of their varied SSMR responses during CPS. The findings are aligned with the prior but limited literature on triggers of SSMR. This study provides empirical evidence of occurrences of trigger situations for MSOT and LSOT and differences in their corresponding SSMR response. These findings have implications for developing targeted scaffolds for teams during CPS.

5. Limitation & Future Work

Although this study provides insights into shared metacognition in CPS, it has limitations. Some teams may have worked outside the classroom, with those interactions not captured. Future research should include asynchronous data and broader sources for comprehensive analysis. The small sample size and factors like motivation or task interest may have influenced SSMR. Larger, more diverse samples and replication across educational contexts would enhance generalizability and validate findings.

6. Conclusion

This study explored triggers and SSMR behavior during ill-structured CPS tasks among four teams with divergent outcomes in semester-long project-based learning HCI course, categorized as MSOT and LSOT based on their performance. We have investigated triggers of SSMR (i.e. ME, SCC, EF). The contrast between these types of teams enabled a deeper examination of the focus and function of their SSMR. It also revealed the dynamics of their SSMR, showing how team members negotiated, shared perceptions about collaboration, and exercised control over the tasks in different ways. The findings indicated notable differences between MSOT and LSOT in terms of the triggers, the focus, and function of SSMR, and how SSMR strategies evolved over time. MSOT maintained a fundamental focus, allowing teams to identify and engage with the core elements necessary for solving the task. In contrast, LSOT tended to operate with surface-level focuses, often concentrating on non-essential aspects. MSOT implemented SSMR strategies effectively and at timely moments during CPS, whereas LSOT's application of these strategies was less effective and poorly timed.

Additionally, while MSOT's approach to strategy formulation and execution was consistent and aligned with their goals, LSOT's efforts appeared sporadic and lacked continuity. These insights contribute to a more refined understanding of the triggers of SSMR along with the defining differences among MSOT and LSOT. By identifying triggers, focus, and functions of SSMR, the findings provide a foundation for designing targeted scaffold mechanisms to strengthen SSMR in CPS contexts.

Acknowledgements
We gratefully acknowledge the Ministry of Education (MoE), Govt. of India, for supporting this work with a PhD research grant.

Ethics approval
This research study was conducted in strict accordance with the guidelines of the Institutional Review Board (IRB) and Ethics Committee at the Indian Institute of Technology Bombay, India.

References

Badhe, V., Priyadarshini, R., & Dasgupta, C. (2022). Understanding SSRL strategies and its impact on group performance in a CSCL environment: a case study from a project-based learning course in India. In Proceedings of the 15th ICLS-CSCL 2022, pp. 187-194.

Chan, C. K. (2012). Co-regulation of learning in computer-supported collaborative learning environments:A discussion. Metacognition and Learning, 7, 63–73.

De Backer, L., Van Keer, H., & Valcke, M. (2022). The functions of shared metacognitive regulation and their differential relation with collaborative learners' understanding of the learning content. Learning and Instruction, 77, 101527.

Efklides, A. (2006). Metacognition and affect: What can metacognitive experiences tell us about the learning process? Educational Research Review, 1, 3–14.

Efklides, A., Papadaki, M., Papantoniou, G., & Kiosseoglou, G. (1998). Individual differences in feelings of difficulty: the case of school mathematics. European Journal of Psychology of Education

Grau, V., & Whitebread, D. (2012). Self and social regulation of learning during collaborative activities in the classroom: The interplay of individual and group cognition. Learning and Instruction, 22(6), 401-412.

Iiskala, T., Vauras, M., & Lehtinen, E. (2004). Socially-shared metacognition in peer learning. Hellenic Journal of Psychology, 1, 147–178. https://psycnet.apa.org/record/2005-14171-003

Iiskala, T., Vauras, M., Lehtinen, E., & Salonen, P. (2011). Socially shared metacognition of dyads of pupils in collaborative mathematical problem-solving processes. Learning and Instruction

Iiskala, T., Volet, S., Jones, C., Koretsky, M., & Vauras, M. (2021). Significance of forms and foci of metacognitive regulation in collaborative science learning of less and more successful outcome groups in diverse contexts. *Instructional Science*

Iiskala, T., Volet, S., Lehtinen, E., & Vauras, M. (2015). Socially shared metacognitive regulation in asynchronous CSCL in science: Functions, evolution and participation. Frontline Learning Research, 3(1), 78-111.

Järvelä, S., Järvenoja, H., Malmberg, J., & Hadwin, A. F. (2013). Exploring socially shared regulation in the context of collaboration. Journal of Cognitive Education and Psychology, 12(3), 267-286.

Järvelä, S., & Hadwin, A. (2024). Triggers for self-regulated learning: A conceptual framework for advancing multimodal research about SRL. Learning and Individual Differences, 115, 102526.

Kerrigan, S., Feng, S., Vuthaluru, R., Ifenthaler, D., Gibson, D. (2021). Network Analytics of Collaborative Problem-Solving. In: Ifenthaler, D., Sampson, D.G., Isaías, P. (eds) Balancing the Tension between Digital Technologies and Learning Sciences. Cognition and Exploratory Learning in the Digital Age. Springer, Cham.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159–174. https://doi.org/10.2307/2529310

Lobczowski, N. G., Lyons, K., Greene, J. A., & McLaughlin, J. E. (2021). Socially shared metacognition in a project-based learning environment: A comparative case study. *Learning, Culture and Social Interaction,* 30, 100543. https://doi.org/10.1016/j.lcsi.2021.100543

Mayring, P. (2015). Qualitative content analysis: Theoretical background and procedures. Approaches to qualitative research in mathematics education: Examples of methodology and methods, 365-380

Vauras, M., Volet, S., & Iiskala, T. (2021). Socially-shared metacognitive regulation in collaborative science learning. Trends and prospects in metacognition research across the life span: A tribute to Anastasia Efklides, 83-102.

Persistence in Electronic-Making Projects: A Socio-Cognitive Perspective

Nisumba SOODHANI*, Ramkumar RAJENDRAN & Sridhar IYER
Centre for Educational Technology, Indian Institute of Technology Bombay, India
*nisumba_soodhani@iitb.ac.in

Abstract: Persistence is crucial in maker-based learning, where challenges are inherent to the creative process and persistence is one of the important behavioral markers of motivation. This study examines how students respond to troubles during making activities, reflecting persistence, using the socio cognitive theory framework of triadic reciprocal causation. Our focus research question is how do the personal beliefs, and environmental factors that influence students' persistence during trouble in making activities? To answer this, we conducted a multi-level video analysis of two student teams engaged in electronic-making projects as case studies. At the macro level, we visualized 16.5 hours of making activities segmented into 1-minute intervals using making-process rugs. At the micro level, we analyzed the identified trouble and repair episodes using interaction analysis, focusing on verbal and embodied actions with peers, facilitators, and the environment. By integrating micro-level interaction data, we observed that persistence is not uniformly distributed among team members, with personal factors prominently shaping responses early in projects, while environmental influences become increasingly significant over time.

Keyword: Persistence, Motivation, Makerspace, Interaction analysis, making process rug

1. Introduction

Sustaining motivation in environments where failure is integral requires a nuanced understanding of motivational processes during making activities to scaffold or design intervention. Existing research has predominantly measured motivation using self-reported survey, interviews, and observation. Self-reports capture cross-sectional motivational states and interviews provide retrospective accounts, yet these methods frequently fail to capture the detailed, real-time processes through which motivation unfolds in situ (DiBenedetto & Schunk, 2022). Observational methods, particularly those involving video data, facilitate repeated and systematic micro-analysis, providing deeper insight into motivational processes as they are enacted (Derry et al., 2010). Unlike self-report surveys and retrospective interviews, video analysis captures immediate, fine-grained interactions, providing insights into how motivation fluctuates during making activities. The observable measures of motivation include task choice, effort, and persistence, which help in understanding how motivation unfolds (Schunk & DiBenedetto, 2021). Choice of task is not always a reliable measure of motivation, as external factors such as social influences or task constraints can shape decision-making (Patall, 2013). Effort is often associated with motivation, but its reliability as an indicator is limited by skill level—as proficiency increases, individuals may require less effort to achieve the same level of performance (Schunk et al., n.d.). Another behavioral index is persistence and from early on, researchers viewed persistence as an objective measure of one's underlying motivation to achieve. Persistence is defined as "*sustained engagement with a task despite encountering obstacles, challenges, or difficulties*" (Feather, 1962). Despite research on motivation in makerspaces (discussed in section 2 – related work), little is known about how persistence is enacted through real-time interactions while working on a project. *RQ: How do the personal beliefs, and environmental factors that influence students' persistence during trouble in making activities?*

2. Related Work

In makerspace research a few studies have attempted to understand persistence. According to Petrich, Wilkinson, and Bevan in the Learning Dimensions Framework, persistence is considered as an important learning outcome in making (Petrich et al., 2013). Hilppö and Stevens conducted a year-long video ethnography of a FUSE Studio makerspace, showing how explicit framing of failure as learning opportunities influenced students' persistence (Hilppö, 2020). They studied persistence through interviews, where students reported that they came to perceive failed attempts not as indications of inadequacy, but as integral steps toward eventual success, reinforced by teachers who celebrated persistence and experimentation. Brahms and Crowley (2016), through their analysis of MAKE magazine describe making as a cycle —marked by purposeful play, experimentation, and refinement—inherently demands persistence. Regalla (2016) identifies persistence as a key attribute of the maker mindset, emphasizing its importance in navigating the complexities of making activities. The maker mindset promotes a failure-positive approach, viewing setbacks as learning opportunities rather than endpoints (Vongkulluksn, Matewos & Sinatra, 2021). The literature reviews on research in makerspaces emphasizes persistence's role in developing problem-solving skills and resilience (Mersand ,2021; Nikou, 2024). Yuan and Murai note that designing context-relevant activity helps learners overcome frustration and develop persistence (Yulis San Juan & Murai, 2022). Given the limitations of previous research on motivation, that largely relies on retrospective accounts or static measures, this study uniquely employs real-time video-based interaction analysis to capture persistence dynamics as they naturally unfold during trouble and repair episodes in makerspaces. This will contribute to extending our understanding of persistence by providing empirical insights into the dynamics of persistence in natural settings.

3. Methods

3.1 Participants

The study was conducted as a part of the workshop designed to engage students in hands-on electronics projects, allowing for observation and analysis of their behavior and interaction. Initially, 33 ninth-grade students enrolled in the workshop, but 24 participants remained until the end, comprising 9 females and 15 males. All participants were approximately 14 to 15 years old and attended the same secondary school. They had limited prior experience in makerspace settings, especially electronic making. Participants formed teams autonomously (2 or 3 in a team), resulting in a total of eight teams, including one individual who chose to work independently. In this paper we focus on two teams – Ace (3 girls) and Atoms (3 boys). We selected these two teams because their projects represented contrasting conditions (imitation vs. creation), allowing deeper exploration of persistence dynamics under different task conditions and team compositions.

3.2 Data Collection

The entire project-making sessions during Week 3 were video-recorded, capturing approximately 11 to 14 hours of footage per group. Cameras were positioned to unobtrusively record both verbal and non-verbal interactions among participants and between participants and the facilitator. This study was approved by the Institute Review Board (IRB) and informed consent was obtained from all participants and their guardians, explaining the purpose of the recordings and ensuring confidentiality.

3.3 Teams' Description

The Ace group (team name given by the participants) consisted of three participants – P1, P2, P3. They undertook an imitation project, building a smoke detector prototype based on a

tutorial found on YouTube. This project involved replicating an existing design, with students following step-by-step instructions to construct the digital model in Tinker cad and assemble the physical components using Arduino. They extended YouTube project by creating a casing for the prototype and decorating it. The useful video data of this making process consisted of 8 hours and 55 minutes. The Atoms group consisted of three participants – P4, P5 and P6. Atoms worked on a creation-based project to develop a prototype of a cane equipped with multiple ultrasonic sensors for visually impaired users. They aimed to create a cane that would allow the user to sense objects at different height level from ground. This project involved conceptualizing a design, building a digital model in Tinker cad, coding for sensor functionality in Arduino, and creating a physical prototype. The useful video data of this creation process consisted of 7 hours and 23 minutes.

3.4 Analysis

The analysis employed a two-level approach (macro, micro). At the macro-level, we segmented videos into one-minute intervals, categorizing each interval based on observed activities aligned with the making design framework (Davies et al., 2024). At the micro-level, SCT guided detailed interaction analyses to identify how personal, behavioral, and environmental factors influenced persistence. Video data were annotated using ELAN software to systematically track verbal and embodied interactions. At the macro level, we utilized the making-process rug visualization method (Paavola et al., 2021) to map the sequence of making design activity over time. This involved segmenting 16.3 hours of video data into one-minute intervals and coding each segment based on the verbal and embodied process based on the making design process framework (Davies, Seitamaa-Hakkarainen, & Hakkarainen, 2024). Due to limited space, we show a few interesting segments of making process rug of team Ace (see Figure 1) where the participants move into off-task action after facing troubles. Time axis runs vertically from top to bottom, the rug on the left shows team Ace's complete making process rug (Figure 1a) and the figures on the right (Figure 1b, 1c) shows segments of team Ace's making process rug. Each cell represents a one-minute interval, color-coded to indicate different making processes. Px, where x = 1,2,3 represents the participants in team Ace. Px_V represents the verbal making process and Px_E represents the embodied making process. To maintain confidentiality, all participants were anonymized using participant codes (e.g., P1, P2), and personal identifiers were removed from video transcripts and analysis logs. In addition to the making process, we also noted down instances of troubles and facilitators or peers' involvement in making as shown in the Figure 1b and Figure 1c.

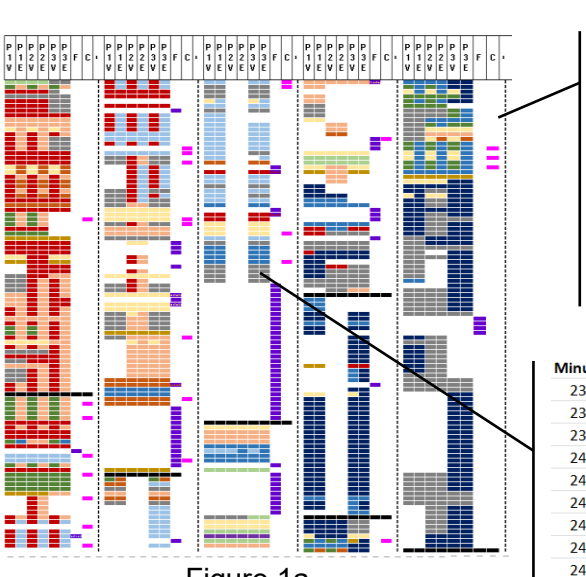


Figure 1a

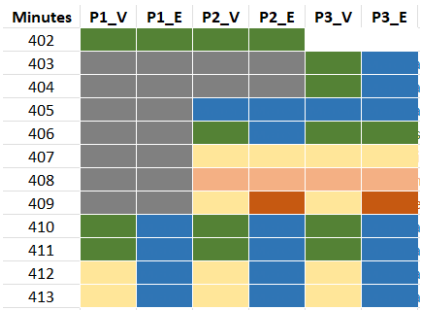


Figure 1b

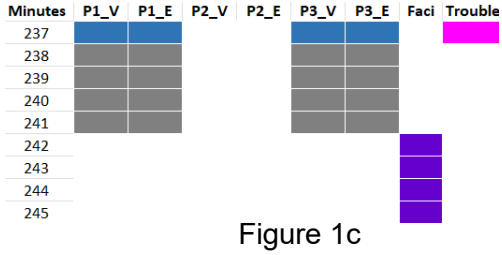


Figure 1c

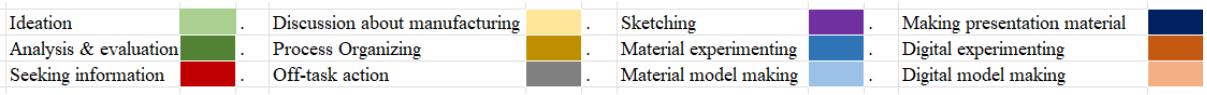


Figure 1a. Making-Process Rug of team Ace; *Figure 1b.* Peer role in sustaining motivation; *Figure 1c.* Facilitator's role in sustaining persistence.

From the macro-level data, we identified episodes where participants faced troubles. Operationally, a 'trouble episode' was identified by observable disruptions such as tinker cad, hardware circuit either partially working or not working, explicit requests for assistance, or clear verbal/non-verbal frustration signals. An episode start was defined as a sequence where a disruption or a perceived challenge occurred and end was defined as when the trouble was resolved, ignored or erased. At the micro level, we conducted a detailed interaction analysis focusing factors influencing persistence on the identified episodes of trouble and repair. The video recordings were repeatedly attended to by the authors and conducted the operations of interaction analysis (Jordan & Henderson, 1995). This multi-level approach provided a comprehensive understanding of how participants respond to troubles and the factors influencing their persistence.

4. Findings

4.1 Personal Factors that Influence Persistence

To address the research question, we conducted a micro-level interaction analysis of the identified trouble and repair episodes, deductively coding them through an SCT lens to understand the factors contributing to persistence. We present exemplar episodes to illustrate how persistence is influenced by different factors. The below transcript is of team Ace, during the initial phase of the making process. The team was working on creating a tinker cad digital circuit of smoke detector by following a YouTube video. The simulation was not working as intended, the buzzer did not beep when smoke was within the threshold value.

Time	Actor	Verbal and non-verbal
01:26:15	P1, P2, P3	*[P1 is simulating the tinkercad circuit, the buzzer doesn't ring. P2 and P3 are observing*]
01:26:21	P3	Teacher! [*P1 pauses, turns to P3]*
01:26:24	P1	Wait……wait, wait
01:26:26	P2	It is not working *[P2 and P3 overlap]*
01:26:26	P3	That's why
	P1	"Teacher" *[imitating P3*] You both are same, for everything you call "teacher! teacher!"
01:26:50	P1	What did the teacher tell? We have to think about logic. First, we'll try.

In this instance P3 and P2 want to seek help to overcome this trouble. But P1 exhibits the desire to solve the problem by themselves, without relying solely on the facilitator. The team then moves on to the analysis phase followed by digital modelling, where they try to debug the block-based program. Using YouTube video as reference they check their block-based codes and all three are convinced that code is correct, but there is some error in connections. At the end of this episode the team Ace was able to partially resolve the trouble – narrowed down the source of trouble. This instance brings out the self-efficacy of P1 in resolving the issues that arise while making (K et al., 2024). This verbal exchange externalized her personal attitude towards making.

In another episode of team Atoms, we noticed that after they encounter challenge P5 and P6 move to off-task action, but P4 persists. Team Atoms were in the material modelling and prototyping process, but the model was not working. As they had tried a few times to power two ultrasonic sensors with one Arduino and it was not working, P5 and P6 wanted to change their entire project (giving up). But P6 persuades them to continue the same project, but they disagree. After some discussion when they don't arrive at a conclusion, P5 and P6 get involved in off-task action. The idea of building a cane for visually challenged people was proposed by P4, and he expresses a sense of ownership in the discussion by trying to persuade them. This episode again brings out the personal processes of his attachment towards the idea which made P4 persist.

4.2 Environmental Determinants that Influence Persistence

In figure 1c, we can notice that P1 and P3 of Team Ace face a trouble and go into off-task action for nearly 4 minutes till the facilitator arrives. The facilitator now tries to debug the circuit along with the participants. After a few minutes of interaction, they find the error in circuit connection and rectify it. As they troubleshoot the error with the facilitator P1 exclaims "*Now I'm confident*". P1 and P3 proceed to material experimenting and presentation material. Here even though they had given up, interaction with facilitator brought them back on track and we can also witness P1 verbalizing her boost in confidence. Here the environmental determinant affected the personal determinant, after which a change in behaviour is noticed.

In the following episode we look at figure 1b in detail. Team Ace was nearing the completion of project. The prototype was working during previous session, but now it was malfunctioning. The buzzer beeps every 2 seconds without smoke. P1, P2 and 3 try to solve the issue.

Time	Actor	Verbal and non-verbal
06:38:14	P1	There is something wrong! Urrghh! *[expresses frustration]*
06:38:49	P2	What happened? *[looking at P1]*
06:38:53	P1	Let's change position *[P1 moves towards P3 ans asks her to change. P3 is reluctant to change, then P1 and P2 exchange places – change in participation structure]*
06:39:00	P3	Don't do like this repeatedly.
06:39:24	P1	*[off-task action – fiddling with paper]*
06:39:30	P3	*[Checks the circuit connections, reconnects a couple of wires]*
06:40:35	P2	Give it here *[Points to the circuit and connects the circuit to the CPU]*
	P2, P3	*[P2 and P3 continue to work on the prototype. 5 minutes of trial and error. Still the buzzer beeps every 2 seconds]*
06:45:45	P1	*[Notices the serial monitor]* Change the value to 200.

After trying to troubleshoot for some time, P1 gets frustrated and shifts the position. Initially, P1 was controlling the mouse, now P1 wants to move away and does not work on this prototype anymore. P2 is now in control of the computer system. P2 and P3 try to troubleshoot, they go through this analysis and material experimenting phase for seven minutes. They then try digital model making and testing during which P1 notices the serial monitor and joins them making process. This episode shows how the behaviour of P2 and P3 brought back P1, a part of the environment as peer, into the making process. It is also important to note here that, it was the same P1 who exhibited self-efficacy in episode 1, gave up in this episode. Episode 1 took place when the team had just started off their project and this episode occurred towards the end of their project, after facing multiple troubles. This serves as evidence that there is a varying effect of these personal and environmental determinants on persistence and that persistence is not solely influence by personal determinants.

5. Conclusion

Addressing the call in contemporary research to analyse motivation in real-time, we employed interaction analysis to reveal how persistence unfolds through trouble and repair episodes in authentic makerspace activities. We observed that persistence is not uniformly distributed among team members, with personal factors prominently shaping responses early in projects, while environmental influences become increasingly significant over time. Our findings highlight the value of combining visual representations (making-process rugs) and micro-level analysis to capture both immediate and longitudinal motivational responses. Also, the findings extend SCT by highlighting how the temporality of persistence, evident through fluctuating self-efficacy and environmental support, suggests a need to better theorize moment-to-moment motivational dynamics. However, the study is limited by its analysis of two teams and the context-specific nature of the projects. To strengthen the generalizability of these findings and to better understand the causal mechanisms underlying persistence, future research would benefit from employing controlled experimental designs that systematically manipulate prominent factors influencing persistence. Additionally, examining a broader range of participant characteristics could further enrich our understanding of how persistence develops and is sustained in makerspace environments.

References

Bandura, A. (1986). Social foundations of thought and action: A social cognitive theory. Prentice-Hall.

Brahms, L., & Crowley, K. (2016). Making sense of making: Defining learning practices in MAKE magazine. In K. Peppler, E. Halverson, & Y. B. Kafai (Eds.), Makeology: Makers as learners (Vol. 2, pp. 13-28). Routledge.

Davies, S., Seitamaa-Hakkarainen, P., & Hakkarainen, K. (2024). Knowledge creation through maker practices and the role of teacher and peer support in collaborative invention projects. International Journal of Computer-Supported Collaborative Learning, 19(3), 283-310.

DiBenedetto, M. K., & Schunk, D. H. (2022). Assessing academic self-efficacy. In M. K. DiBenedetto & D. H. Schunk (Eds.), Academic self-efficacy in education: Nature, assessment, and research (pp. 11-37). Springer.

Hilppö, J. (2020). “Failure is just another try” Re-framing failure in school through the FUSE studio approach. International Journal of Educational Research.

Jordan, B., & Henderson, A. (1995). Interaction analysis: Foundations and practice. The Journal of the Learning Sciences, 4(1), 39-103.

K, N. S., V, A., Saxena, A., & Rajendran, R. (2024). Unraveling Learner Self-Efficacy in the Electronic Making: Insights from Interaction Analysis, Self-Reports and Interview. Proceedings of the 18th International Conference of the Learning Sciences - ICLS 2024 (pp. 1454-1457). International Society of the Learning Sciences.

Mersand, S. (2021). The state of makerspace research: A review of the literature. TechTrends, 65(2), 174-186.

Nikou, S. A. (2024). Student motivation and engagement in maker activities under the lens of the Activity Theory: A case study in a primary school. Journal of Computers in Education, 11(2), 347-365.

Paavola, T., Kangas, K., Kokko, S., Riikonen, S., Mehto, V., Hakkarainen, K., & Seitamaa-Hakkarainen, P. (2021). Video data analysis for tracing emotional aspects of collaborative design and making processes. Techne Series: Research in Sloyd Education and Craft Science A, 28(3), 1-14.

Patall, E. A. (2013). Constructing motivation through choice, interest, and interestingness. Journal of Educational Psychology, 105(2), 522–534. https://doi.org/10.1037/a0030307

Regalla, L. (2016). Developing a maker mindset. In K. Peppler, E. Halverson, & Y. B. Kafai (Eds.), Makeology: Makers as learners (Vol. 2, pp. 257-272). Routledge.

Rusk, N. (2016). Motivation for making. In K. Peppler, E. Halverson, & Y. B. Kafai (Eds.), Makeology: Makerspaces as learning environments (Vol. 1, pp. 85-108). Routledge.

Schunk, D. H., Meece, J., & Pintrich, P. R. (2014). Motivation in education: Theory, research, and applications (4th ed.). Pearson.

Serholt, S., Pareto, L., Ekström, S., & Ljungblad, S. (2020). Trouble and Repair in Child–Robot Interaction: A Study of Complex Interactions With a Robot Tutee in a Primary School Classroom. Frontiers in Robotics and AI, 7, 46. https://doi.org/10.3389/frobt.2020.00046

Vongkulluksn, V. W., Matewos, A. M., & Sinatra, G. M. (2021). Growth mindset development in design-based makerspace: A longitudinal study. The Journal of Educational Research, 114(2), 139-154.

Yulis San Juan, A., & Murai, Y. (2022). Turning frustration into learning opportunities during maker activities: A review of literature. International Journal of Child-Computer Interaction, 33, 100519. https://doi.org/10.1016/j.ijcci.2022.100519

Comparing students' social knowledge construction in the Metaverse-based and chat-based collaborative learning

Suwandi SUWANDI[a], Darmawansah DARMAWANSAH[b*], Benni Ichsanda Rahman HZ[c] & Hsueh-Hua CHUANG[d]

[a,d]*International Graduate Program of Education and Human Development, National Sun Yat- sen University, Taiwan*

[b*]*Global Foreign Language Education Program, Providence University, Taiwan*

[c]*Tarbiyah and Teacher Training Faculty, Universitas Islam Negeri Sumatera Utara, Indonesia*

*darmawansahlavoe@gmail.com

Abstract: This study compares students' social knowledge construction (SKC) in an immersive Metaverse platform and an asynchronous chat platform within a cross-cultural collaborative task on the Sustainable Development Goals (SDGs). Guided by social constructivist theory and the Cognitive Affective Model of Immersive Learning (CAMIL), interaction data from 72 undergraduate students were analyzed using a revised Interaction Analysis Model (IAM) and Epistemic Network Analysis (ENA). Results show that Metaverse-based sessions promoted higher frequencies of information sharing (KC1) and socially oriented discourse (KC6), whereas chat-based interactions supported more negotiation of meaning (KC3) and application of knowledge (KC5). ENA visualizations revealed stronger epistemic connectivity among advanced discourse moves in chat-based environments. These findings suggest aligning platform affordances with pedagogical goals, leveraging immersive tools for engagement and rapport, and asynchronous tools for deeper elaboration, while reframing social talk as a relational bridge in cross-cultural collaboration.

Keywords: Collaborative learning, chat, immersive learning, social knowledge construction, metaverse,

1. Introduction

Immersive technologies such as the Metaverse enable real-time, embodied, and spatially rich interaction, offering unique affordances such as avatar-mediated presence, multimodal cues, and spatial navigation that can deepen engagement and support situated, constructivist learning (Dede, 2009; Mystakidis, 2022). While prior research highlights the Metaverse's potential for collaborative problem-solving (Radianti et al., 2020), little empirical work has examined *how* learners construct knowledge socially in such environments, and how these processes differ from those in asynchronous, text-based settings.

Social knowledge construction (SKC), grounded in social constructivist theory (Vygotsky, 1978; Stahl, 2006), involves negotiation, questioning, and mutual regulation to advance shared understanding, aligning with the principle of collective cognitive responsibility (Scardamalia & Bereiter, 2010). In immersive contexts, SKC combines verbal, spatial, and multimodal communication, yet its interactional patterns and underlying mechanisms remain underexplored, particularly the role of technological affordances in shaping epistemic engagement. Epistemic engagement here refers to sustained participation in higher-order processes such as questioning, elaborating, and synthesizing.

This study compares two deliberately sequenced environments within the same collaborative learning design: synchronous, voice-based sessions in the Metaverse

(Gather Town) for idea generation and embodied discussion, and asynchronous, text-based communication in LINE for refinement and elaboration. Beyond modality differences, the Metaverse affords spatial presence and avatar embodiment theorized to foster rapport and immediacy (Makransky & Petersen, 2021), whereas LINE's persistent textual format may encourage reflective contributions but lacks embodied cues. Guided by the Cognitive Affective Model of Immersive Learning (CAMIL) and affordance theory, we conceptualize technological features as mediators interacting with task purpose, temporal structure, and pedagogical intent to shape SKC outcomes.

This study addresses the following research questions:

(1) What are the structural characteristics of student interaction networks in Metaverse-based collaborative learning environments compared to the conventional interaction through real-time asynchronous chat?
(2) What patterns of social knowledge construction behaviors emerge among students during collaborative tasks in the Metaverse compared to the conventional interaction through real-time chat?

2. Literature Review

2.1. Social knowledge construction in online collaborative learning

Social knowledge construction (SKC) refers to learners' collaborative negotiation, conflict resolution, and joint meaning-making through discourse (Gunawardena et al., 1997). Grounded in Vygotsky's (1978) social constructivist theory, SKC frames learning as a dialogic process in which higher-order thinking first develops socially before becoming internalized (Stahl, 2006). The Interaction Analysis Model (IAM) (Gunawardena et al., 1997) categorizes SKC into five phases: information sharing (KC1), cognitive dissonance (KC2), negotiation of meaning (KC3), synthesis testing (KC4), and application of new knowledge (KC5).

Although IAM is widely applied in asynchronous text-based settings, its use in synchronous, multimodal, and immersive contexts remains limited. Existing work often focuses on cognitive or linguistic outcomes (Lee & Bonk, 2016) without examining how technological affordances such as modality, persistence, and spatial representation mediate both social dynamics and epistemic engagement. Given the shift toward multimodal, distributed learning environments, SKC analysis must integrate these affordance-related mechanisms, particularly in contexts where tools are intentionally assigned for distinct phases of collaboration.

2.2. Metaverse-based collaborative learning

The Metaverse is a digitally immersive environment enabling embodied, spatially aware, and interactive learning spaces (Mystakidis, 2022). In education, it supports enhanced presence, co-location, identity exploration, and multimodal expression (Dede, 2009). Its unique affordances, proximity-based audio, avatar-mediated interaction, and manipulable artifacts can foster immediacy, rapport, and co-construction of knowledge (Park & Kim, 2022).

However, most studies in this area focus on usability, engagement, or motivation (Radianti et al., 2020), leaving a gap in understanding the epistemic processes within immersive environments. Moreover, when the Metaverse is deliberately positioned for synchronous ideation and LINE for asynchronous refinement, as observed in this study, SKC differences may stem from the interaction of affordances, temporal structure, and pedagogical purpose, not from modality alone. This highlights the need to disentangle how real-time spatial dynamics and asynchronous persistence shape the mechanisms of SKC.

3. Methodology

3.1. Participants

The study involved 72 undergraduate students (aged 19 to 24) from two universities in Taiwan and Indonesia who participated in a Cross-Cultural Communication course. All had basic digital literacy but no prior structured intercultural online discussion experience. They were

placed into 15 mixed-nationality groups of 5 to 6 members, each led by a facilitator. Taiwanese students joined in-person on campus, while Indonesian students participated remotely from campus or home. Gather Town was used exclusively for synchronous, avatar-mediated sessions; LINE was used solely for asynchronous follow-up, ensuring differences reflected each platform's affordances and temporal structure.

3.2. The instructional design of the study

The three-week SDG-themed project was deliberately sequenced to match platform strengths. Gather Town supported synchronous video/audio and spatial interaction for idea generation, while LINE enabled asynchronous coordination, file sharing, and refinement. Following a social constructivist multi-layered framework, students first held intragroup discussions in Gather Town, then participated in intergroup feedback sessions, and finally co-created an infographic. The instructor provided ongoing scaffolding, technical support, and guided summarization. Figure 1 shows the task flow and platform transitions.

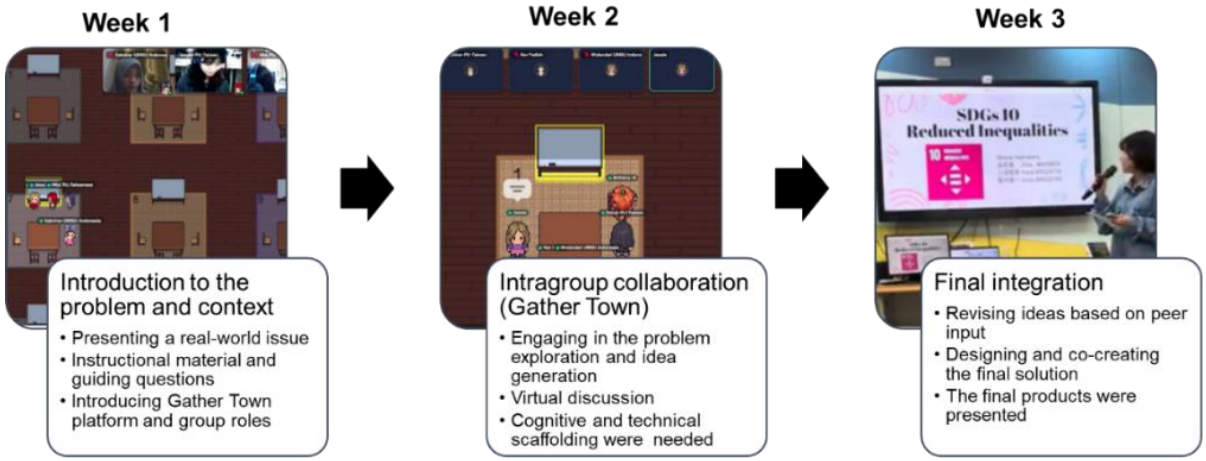


Figure 1. The instructional design of the study

3.3. Data collection and analysis

To investigate the dynamics of social knowledge construction in the Metaverse, this study analyzed social knowledge construction (SKC) using content analysis supported by Epistemic Network Analysis (ENA) to capture both the structural properties and epistemic quality of interactions. Data came from two sources: (1) 1,312 transcribed utterances from Gather Town (synchronous, embodied negotiation) and (2) 784 cleaned LINE messages (asynchronous coordination and refinement). All entries were time-stamped and sequenced for temporal modeling, ensuring consistency across platforms.

Discourse was coded using a modified Interaction Analysis Model (Gunawardena et al., 1997), retaining KC1–KC5 and adding KC6 for off-task/relational exchanges that support group cohesion in cross-cultural contexts. Two trained coders reached $\kappa = 0.81$ (Fleiss, 1981); disagreements were resolved through discussion to maintain theoretical alignment.

ENA (Shaffer et al., 2016) was applied via the ENA Web Tool (https://www.epistemicnetwork.org) to model co-occurrence and temporal proximity of coded moves. Gather Town and LINE datasets were chronologically merged so that temporally adjacent utterances, regardless of platform, were analyzed in sequence. Each unit was treated as a node, segmented by turn or message sender. Network graphs visualize co-occurrence strength (edge thickness) and epistemic focus (centroid position), allowing direct comparison of patterns across the two environments.

4. Findings and Discussion

4.1. What are the structural characteristics of student interaction networks in Metaverse-based collaborative learning environments compared to the conventional interaction through real-time asynchronous chat?

The two platforms were intentionally assigned distinct roles: Gather Town for synchronous, embodied discussion, and LINE for asynchronous, text-based refinement. Differences in discourse patterns thus reflect not only platform affordances but also task sequencing and temporal structure.

Off-topic or socially oriented talk (KC6) dominated both contexts, 63.26% in Gather Town and 58.63% in LINE. While not directly epistemic, KC6 served relational purposes, fostering rapport and easing intercultural interaction functions supported in prior work (Chamola et al., 2024). The slightly higher KC6 in Gather Town may relate to avatar presence and multimodal cues enhancing social bonding (Lin et al., 2022), whereas in LINE, KC6 more often transitioned to negotiation (KC3), suggesting reflective scaffolding. As social presence was not directly measured, these interpretations are inferential.

Information sharing (KC1) was higher in Gather Town (23.06%) than LINE (15.26%), likely due to immediacy of verbal exchanges and use of shared artifacts, but much remained surface-level, consistent with findings that immersive contexts require scaffolding for deeper reasoning (Parong & Mayer, 2018). Conversely, KC3 was notably higher in LINE (18.47% vs. 2.29%), aligning with literature on asynchronous environments enabling reflection and deliberate response (Lee & Bonk, 2016).

Both settings showed low synthesis (KC4) and application (KC5), with slightly higher KC5 in LINE (6.83% vs. 2.76%), indicating a common challenge in advancing toward integrative reasoning without explicit facilitation (Barana et al., 2023). Discovery of dissonance (KC2) was rare in both (8.04% in Gather Town, 0.80% in LINE), reflecting limited peer-challenging again consistent with studies showing students often avoid critique unless prompted (Yoo et al., 2024).

These findings indicate that platform affordances and design intentions jointly shaped interaction patterns: Gather Town promoted immediacy and rapport, while LINE supported more reflective negotiation and application.

Table 1. *The structural characteristics of students' interaction networks across the two tools*

	Metaverse-based interaction		Conventional real-time chat interaction		Total
	N	%	N	%	
KC1	301	23.06	38	15.26	339
KC2	105	8.04	2	0.80	107
KC3	30	2.29	46	18.47	76
KC4	7	0.54	0	0	7
KC5	36	2.76	17	6.83	53
KC6	826	63.26	146	58.63	972
Total	1305	100	249	100	1554

4.2. What patterns of social knowledge construction behaviors emerge among students during collaborative tasks in the Metaverse compared to the conventional interaction through real-time chat?

ENA visualizations (Figure 2) compare the Metaverse (red) and LINE (blue) networks. Although LINE enables near real-time messaging, replies were often delayed by hours; thus, it is treated as asynchronous. The maps show both the frequency of epistemic moves and the transitions between them.

In the Metaverse, the strongest link was between KC1 (Sharing/comparing information) and KC6 (Social/off-topic talk) at 0.87, indicating frequent alternation between task-related and informal exchanges. Affordances such as avatar presence, proximity-based audio, and spatial navigation likely supported this blending of social and task talk. However, transitions from KC6 to higher-order moves, namely KC5 (Application) at 0.13, KC2 (Dissonance discovery) at 0.28, KC3 (Co-construction), and KC4 (Synthesis), were minimal, echoing findings that immersive environments encourage engagement but require structured facilitation for deeper reasoning (Lin et al., 2022).

LINE networks showed stronger links between advanced moves, particularly between KC6 and KC3 (0.39) as well as between KC1 and KC6 (0.38). In these instances, social talk often served as a bridge to negotiation of meaning, benefiting from asynchronous conditions that allow planning, reflection, and elaboration, consistent with Gagliani et al. (2025).

Statistical results reinforce this: a two-sample t-test on ENA centroid X-axis values (co-construction weight) found Gather Town (M = -0.25, SD = 0.12, N = 15) significantly lower than LINE (M = 0.29, SD = 0.45, N = 13), t(13.40) = -4.18, p < .001, d = 1.69. This confirms that asynchronous chat supported more frequent and stronger KC3, while Gather Town emphasized KC1 and KC6, similar to patterns reported by Li et al. (2024).

KC4 and KC2 remained low in both contexts; the Metaverse showed slightly more idea divergence, while LINE had more KC5, indicating greater convergence and application. These patterns align with Barana et al. (2023), suggesting that without scaffolding, immersive tools may stall in exploratory phases, whereas asynchronous tools more effectively promote integrative reasoning.

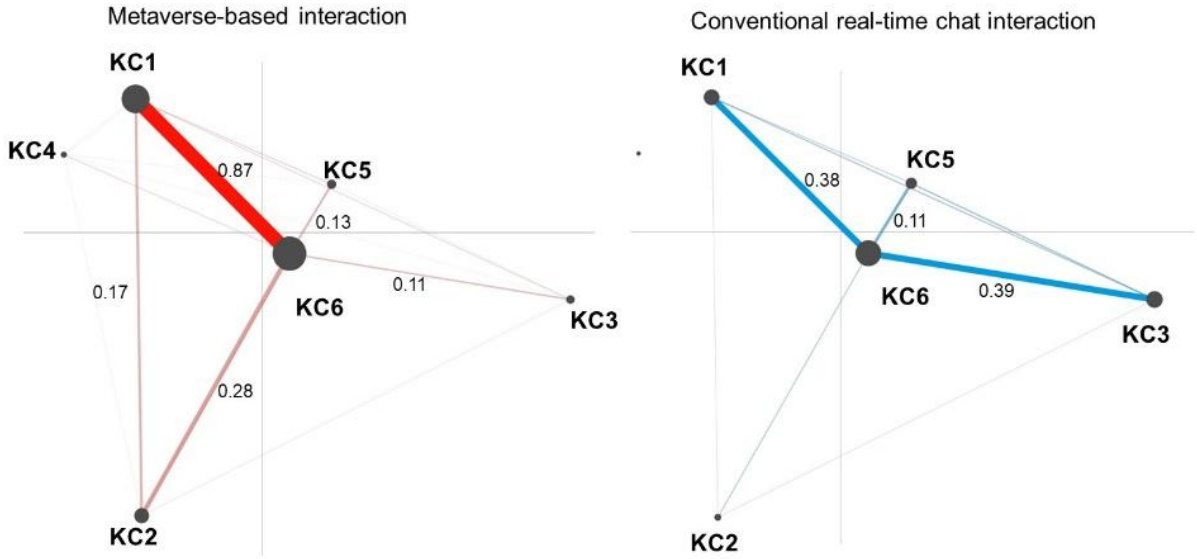


Figure 2. The social knowledge construction of students across two tools

5. Conclusion

This study compared social knowledge construction (SKC) in an immersive Metaverse environment (Gather Town) and an asynchronous chat platform (LINE) within the same cross-cultural collaborative design. Using content analysis and Epistemic Network Analysis (ENA), results showed that Gather Town interactions were dominated by information sharing (KC1) and socially oriented talk (KC6), whereas LINE more frequently supported negotiation of meaning (KC3) and application of knowledge (KC5). Interpreted in light of the intentional design synchronous embodied discussion in the Metaverse and asynchronous refinement in LINE these differences reflect the combined influence of platform affordances, task sequencing, and temporal structure. Immersive tools can enhance engagement and rapport but need structured scaffolding to promote higher-order reasoning; asynchronous tools better enable reflection and elaboration. Socially oriented talk should be recognized as a relational bridge in cross-cultural collaboration. Limitations include the modest sample size, short duration, and context-specific scope. Future research should involve more diverse cohorts, extended interventions, varied platforms, and targeted scaffolding strategies, including AI-assisted prompts. Overall, findings emphasize that epistemic engagement emerges from the interaction of technological affordances and socio-relational dynamics, and that channeling social exchanges into knowledge building is critical for effective collaborative learning.

Acknowledgements

This study is supported in part by the National Science and Technology Council of Taiwan under contract number NSTC 114-2410-H-126-002-MY2

References

Barana, A., Boetti, G., Marchisio, M., Perrotta, A., & Sacchet, M. (2023). Investigating the knowledge co-construction process in homogeneous ability groups during computational lab activities in financial mathematics. *Sustainability, 15*(18), 13466. https://doi.org/10.3390/su151813466

Chamola, V., Peelam, M. S., Mittal, U., Hassija, V., Singh, A., Pareek, R., ... & Brown, D. J. (2024). Metaverse for education: Developments, challenges, and future direction. *Computer Applications in Engineering Education*. https://doi.org/10.1002/cae.70018

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

Dede, C. (2009). Immersive interfaces for engagement and learning. *science*, *323*(5910), 66-69. https://doi.org/10.1126/science.1167311

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). John Wiley & Sons.

Gagliani Caputo, S., Branchetti, L., & Cusi, A. (2025). Intertwining students' social modes of co-construction and epistemic aspects of algebraic thinking in asynchronous mathematical discussions. Educational Studies in Mathematics. https://doi.org/10.1007/s10649-025-10395-z

Gunawardena, C. N., Lowe, C. A., & Anderson, T. (1997). Analysis of a global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing. *Journal of educational computing research*, *17*(4), 397-431. https://doi.org/10.2190/7MQV-X9UJ-C7Q3-NRAG

Lee, J., & Bonk, C. J. (2016). Social network analysis of peer relationships and online interactions in a blended class using blogs. *The Internet and Higher Education*, *28*, 35-44. https://doi.org/10.1016/j.iheduc.2015.09.001

Li, W., Chang, H.-Y., Bradford, A., Gerard, L., & Linn, M. C. (2024). Combining natural language processing with epistemic network analysis to investigate student knowledge integration within an AI dialog. *Journal of Science Education and Technology, 33*(1), 1–17. https://doi.org/10.1007/s10956-024-10176-y

Lin, H., Wan, S., Gan, W., Chen, J., & Chao, H.-C. (2022). Metaverse in education: Vision, opportunities, and challenges. *Computers and Electrical Engineering, 102*, 108171. https://doi.org/10.1016/j.compeleceng.2022.108171

Makransky, G., & Petersen, G. B. L. (2021). *The Cognitive Affective Model of Immersive Learning (CAMIL): A theoretical research-based model of learning in immersive virtual reality.* Educational Psychology Review, 33, 937–958. https://doi.org/10.1007/s10648-020-09586-2

Mystakidis, S. (2022). Metaverse. *Encyclopedia*, *2*(1), 486-497. https://doi.org/10.3390/encyclopedia2010031

Park, S.-M., and Kim, Y.-G. (2022). A metaverse: taxonomy, components, applications, and open challenges. IEEE Access 10, 4209–4251. https://doi.org/10.1109/access.2021.3140175

Radianti, J., Majchrzak, T. A., Fromm, J., & Wohlgenannt, I. (2020). A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Computers & education*, *147*, 103778. https://doi.org/10.1016/j.compedu.2019.103778

Scardamalia, M., & Bereiter, C. (2010). A brief history of knowledge building. *Canadian Journal of Learning and Technology/La revue canadienne de l'apprentissage et de la technologie*, *36*(1).

Shaffer, D. W., Collier, W., & Ruis, A. R. (2016). A tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. *Journal of learning analytics*, *3*(3), 9-45. https://doi.org/10.18608/jla.2016.33.3

Stahl, G. (2006). *Group cognition: Computer support for building collaborative knowledge (acting with technology)*. The MIT Press.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (Vol. 86). Harvard university press.

Yoo, H. J., Ryu, H., & Lee, H. (2024). Effectiveness of Metaverse-based collaborative learning in nursing education: A mixed-methods study. *Journal of Nursing Education, 63*(9), 585–593. https://doi.org/10.3928/01484834-20240514-04

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

The Use of Sensor Data for the Multimodal Analysis of Mathematical Proof Activity in a CSCL Environment

Masataka KANEKO[a*], Hironori EGI[b], Takuya KITAMOTO[c] & Takeo NODA[d]
[a]*Center for Higher Education Development, Toho University, Japan*
[b]*Graduate School of Informatics and Engineering, The University of Electro-Communications, Japan*
[c]*Faculty of Education, Yamaguchi University, Japan*
[d]*Faculty of Science, Toho University, Japan*
*masataka.kaneko@phar.toho-u.ac.jp

Abstract: Recent technological developments have greatly enhanced the power of multimodal learning analytics to help researchers better understand how students learn in a CSCL environment in a scalable manner. In this study, we try to use depth sensor data to identify the temporal transition of learners' behaviors and investigate the relationship between it and the progress of their mathematical thinking. Human-coded behavioral data are compared with the log data of learner movements derived from a depth sensor to evaluate the extent to which the sensor data can be used to infer the progress of learners' thinking during mathematical proof activity in a CSCL environment.

Keywords: Depth sensor data, mathematical proof activity, dynamic geometry, multimodal learning analytics

1. Introduction

Many previous studies in cognitive neuroscience have illustrated that mathematical thinking emerges from the interplay of the cerebral networks for language processing and those for visuospatial processing (Dehaene, et al., 1999). Moreover, mathematical cognition is grounded in the mental simulation of dynamic imagery supported by the interlocking network of the visual system and the motor system (Lakoff and Nunez, 2001). Regarding these characteristics of mathematical cognition, it is natural that many CSCL studies in mathematics have been conducted from a Multimodal Learning Analytics (MMLA) perspective. For instance, one of the pioneering studies used a process data stream derived from high school students' collaborative work for solving algebra and geometry problems and extracted features of their behavioral patterns to predict their expertise with automatic algorithms (Ochoa, et al., 2013). The data used in the study was the Math Data Corpus (Oviatt, et al., 2013) which is time-synchronized multimodal data composed of video recordings of learners' physical movements including their calculator use, audio data from their conversations, and digital pen stroke data derived from their writing. It also includes human-coded data on each group's problem-solving correctness and the representational content of each student's writing. While automatic algorithms can give good predictors in these cases when students are engaged in well-structured tasks, extra elaboration for combining hand-annotated data and system-collected data is needed to track and interpret learners' thinking during project-based tasks such as programming, electronics, and mathematical proof (Spikol, et al., 2016, Blikstein, & Worsley, 2016). One previous study (Spikol, et al., 2017) illustrated how sensor-based learning analytics systems can help us identify the MMLA features relevant for collaborative problem solving even in the case of complex engineering tasks, by the regression analysis using human-evaluated scores of learners' collaborative problem solving as output and data from learners' hand movements and face directions collected through low-cost sensing

technologies as input, together with the machine-learning based estimation of those regression methods.

In this study, we investigate the verbal and nonverbal behaviors of several pairs of university students as they were engaged in a mathematical proof activity. Using HTML-based dynamic content collaboratively, they were asked to prove the addition theorem for trigonometric functions when the related angles were not necessarily acute. The analysis was conducted by aligning the behavioral data derived from video recordings, the sound data of their conversations derived from audio recordings, and the log data of body movements derived from a depth sensor camera on a single timeline. Based on the theory of mathematical cognition, it is assumed that the progress of learners' mathematical thinking might be linked to the appropriate pattern of temporal transition involving their linguistic, visual, and motor functions. Moreover, a recent CSCL study investigating the quality of collaborative group engagement in designing a floor plan of an office while using a 3D modeling tool has demonstrated that the quality rated by using learners' verbal communication data is highly correlated to their nonverbal behaviors (Paneth, et al., 2023). This knowledge gave the authors the idea that the qualitative analysis of verbal communication data can be complemented by an analysis of the change in behavioral patterns observed from the depth sensor data. As a preliminary step to examine this idea, we pose the following two RESEARCH QUESTIONS in this study: (1) Is the progress of learners' thinking identified in their verbal communication data interrelated to any change in their behavioral pattern observed in videotaped images? (2) Can the change in learners' behavioral patterns be portrayed on the log data derived from the depth sensor?

2. Methods

2.1 The Content Used and Task Design

The learning task is to prove the addition theorem for the trigonometric function

$$\sin(\alpha + \beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta$$

when the relevant angles α, β, and $\alpha + \beta$ are not necessarily acute. In these general cases, the positional relationship between geometric elements becomes complex compared to the rudimentary case where the relevant angles are acute. Figure 1 shows the HTML-based dynamic geometry content used in this study. The left and right sides display the geometric figure and the mathematical expression involved in the proof for the rudimentary case respectively. The central part displays the dynamic geometry screen on which learners can move the angles α and β by manipulating points A and B.

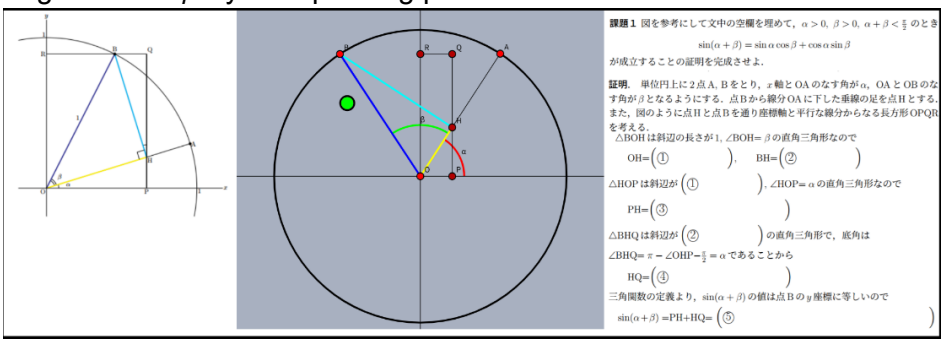


Figure 1. The HTML-based content used in this study

Before the collaborative learning session, participants were asked to review the proof in the rudimentary case individually using printed material including the left and right parts in Figure 1. After the instructor confirmed the completion of this preliminary step, they were asked to work on the proof activity in pairs using an iPad onto which HTML content had been implemented. While manipulating the points A and B, learners can discuss why the same formula holds although the positional relationship between geometric elements changes accordingly. The focal point of discussion is the signs of the trigonometric ratios for

supplementary and complementary angles. Since identifying those signs in specific situations was assumed to burden heavy cognitive load on the participants (Sweller, et al., 2011), they were allowed to use a worksheet to write down their ideas and discuss them with their partners.

2.2 The Flow of Experiment and Data Analysis

Participants were first-year students in a Japanese university majoring in engineering or information technology. In high school, they learned the proof for the theorem in the rudimentary case. The members of each dyad had been friends since before the experiment. For each group, the participant on the left side was named A and the one the right was named B. While an iPad was delivered to each group, a worksheet was prepared for each participant to record their ideas during the collaborative learning session. The position of these items was fixed as shown in Figure 2 (left).

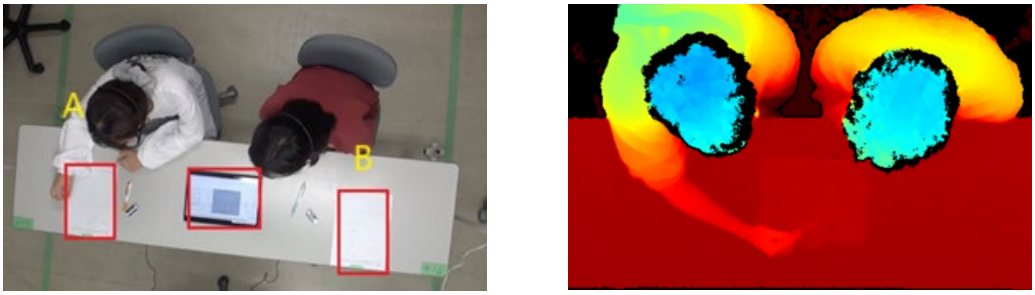


Figure 2. The setting of iPad and worksheets on a table (left) and the sample of the colored depth image derived from Intel RealSense depth camera (right)

Each participant wore a microphone through which their utterances were recorded. The recorded discourse was subsequently transcribed with time stamps. The transcribed text data was analyzed to identify when and how their thinking progressed. Also, the participants' behaviors were videotaped and the recorded image was imported into the behavioral analysis system Sportscode (https://www.hudl.com/products/sportscode). Their behaviors were classified into several categories including "manipulating dynamic content", "making gestures", "pointing to mathematical expressions", and "writing on the worksheet". The temporal transition of each group's behaviors was coded on a Sportscode time line. Moreover, the Intel RealSense depth camera (https://www.intelrealsense.com/stereo-depth), based on stereo vision technology, was used to measure the depth from the sensor to each picture element on the table and depth data were logged with time stamps. Figure 2 (right) shows a sample of colored depth image derived from RealSense. Using this log data, the movements of the participants' hands and heads intercepting light from the depth sensor over the specific regions including two worksheets and the iPad (outlined in red on the left of Figure 2) were visualized on a time line respectively. The time intervals during which progress in participants' thinking was identified from the transcribed text data were plotted on the Sportscode timelines, which were examined for changes in their behavioral patterns across these intervals. Additionally, RealSense timelines were compared with the Sportscode timelines to determine if any of those changes were portrayed on the RealSense data.

3. Results

The participants were assumed to argue about the following three cases: (1) $\alpha > \frac{\pi}{2}, \alpha + \beta < \pi$, (2) $\beta > \frac{\pi,}{2}, \alpha + \beta < \pi$, (3) $\alpha < \frac{\pi}{2}, \beta < \frac{\pi}{2}, \frac{\pi}{2} < \alpha + \beta < \pi$. Here we analyze the data from the CSCL of 8 groups I - VIII. Due to the confusion about the range of angles, groups II and VII reached an impasse during their session and completed the task for only one of these cases. Also, groups VI and VIII became confused due to their insufficient understanding of the trigonometric ratios for supplementary and complementary angles, though they finally accomplished the task for at least two cases. While group IV took as long to accomplish the

task as these groups, the participants' struggle was caused by their difficulty in reconciling the case when $\alpha = \frac{\pi}{2}$ in which some relevant triangle "collapses". Regarding the utterance of Participant B, "It is enough to choose the calculation method (addition and subtraction) in accordance with the positional relationship between PH and HQ" while pointing to the mathematical expression in rudimentary case (34:29), it can be seen that group IV got the whole picture of the proof at the last stage. However, except for the limited time intervals involving this sort of discussion, the participants' communication was seemingly at an impasse. The remaining 3 groups I, III, and V completed the task for the above mentioned three cases and recognized their similarities and differences. While a previous gesture study by Nathan (Nathan, et al., 2014) illustrated that task-relevant actions together with pedagogical language can facilitate students' gaining mathematical insight, another previous study by Alibali et al. (Alibali, et al., 2012) clarified that gestures manifest embodied mathematical knowledge in three distinct ways. In the case of this study, most of the observed gestures were pointing to the geometric elements in the dynamic geometry screen and the rest were mainly conceptual metaphors depicting the trigonometric ratios for supplementary and complementary angles.

Figure 3 shows timelines derived from Sportscode and RealSense in which the horizontal axes pointing to the right represent the passage of time and are synchronized for each group. In the Sportscode timelines, the behaviors – "manipulation of the dynamic content", "making gestures", "pointing to the symbolic expression", and "writing on the worksheet" - of each participant were coded from next to the top row to the bottom row. The time intervals of the progress in thinking were identified by analyzing the transcribed text data and are highlighted in yellow. It can be observed that, through these time intervals, participants' behaviors shifted their weight from manipulation and gesture (acquisition of visual information) to pointing at and writing mathematical expressions (conversion to formal deduction). In the RealSense timelines, the fluctuations of the depth from the sensor to the iPad region, left worksheet region (participant A), and right worksheet region (participant B) are visualized by the bar graphs in blue, green, and red respectively. As observed in the videotaped image, while writing behaviors usually occurred over the worksheet region, other ones occurred mainly over the iPad region which can be regarded as "the space for collaborative work". In accordance with this observation, red and green bars become conspicuous in RealSense timelines when participants A and B were engaged in writing on the worksheet respectively.

The analysis of videotaped images also indicates that the role of writing behavior changed as participants' thinking progressed. At the earlier stage, writing behavior and pointing to geometric objects often occurred alternately and participants wrote down some pieces of their observations on the worksheets. In contrast, at the later stage, they were solely engaged in writing behavior to describe the whole picture of proof. Regarding this change, the case of group V is exceptional. In fact, considerable weight was placed on writing compared to other groups and participants began to describe the whole process of proof at a very early stage with the utterance by B (02:06), "Writing may lead to our understanding". Judging from this utterance, the participants might have already recognized that a similar outline of proof can be applied to these general cases with minor changes in a few pieces.

4. Discussion and Conclusion

Comparison of Sportscode timelines between groups gives some insight into the RQ. (1). While the transition from visual information acquisition to formal deduction is observed in the cases of groups with relatively short thinking processes, that transition is not so apparent in other cases including Groups II and IV. This contrast strongly indicates that monitoring the changes in learners' behavioral patterns in the CSCL environment can lead to a diagnosis of their impasse. A group-wise comparison between the timelines of Sportscode and RealSense gives clearer insight into the RQ. (2). The movements of participants' hands and heads over worksheet regions associated with their writing behaviors are clearly portrayed on RealSense time lines. These findings suggest that, as in many other related studies (Sharma and Gianna-

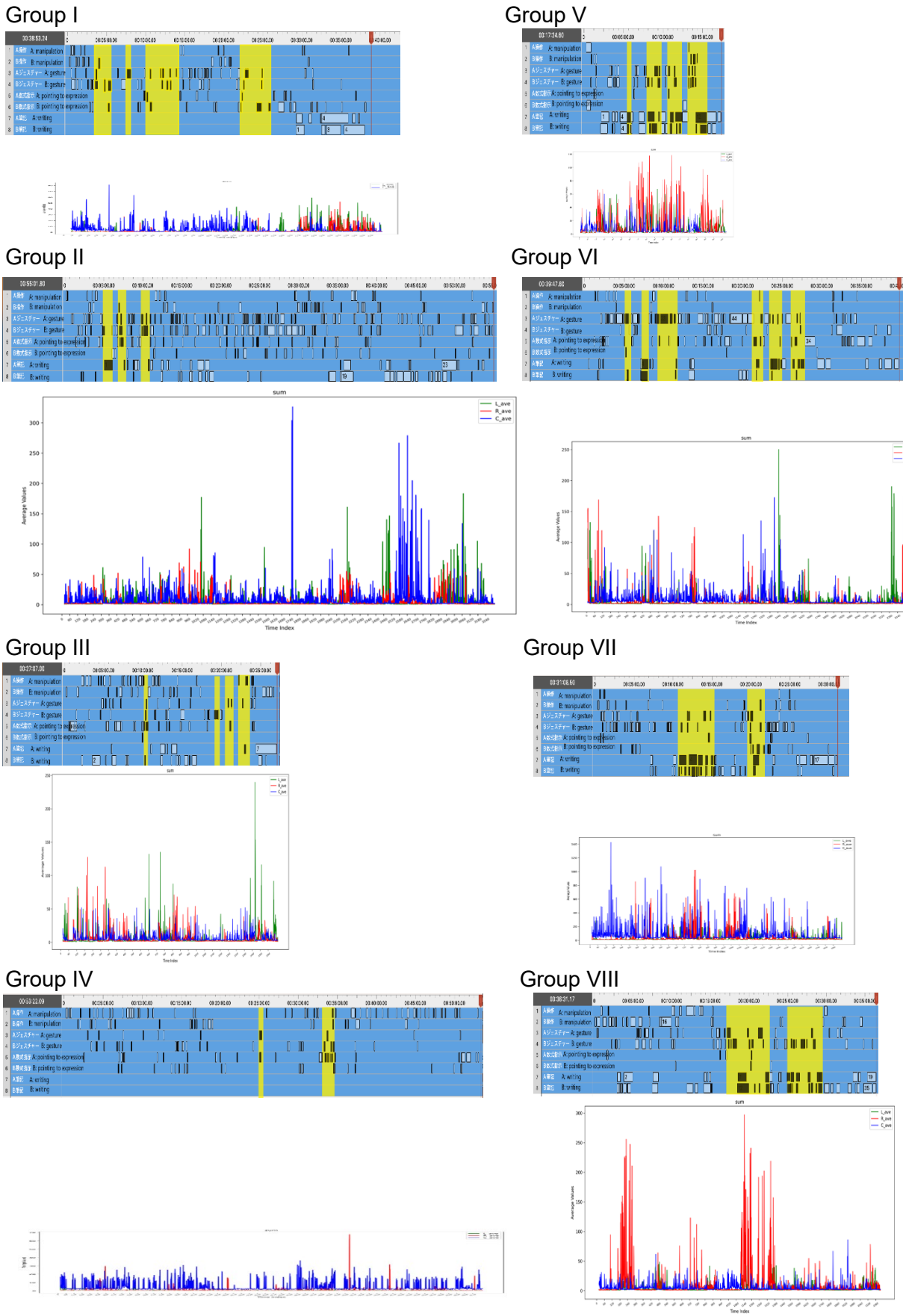


Figure 3. Time lines for each group derived from Sportscode (first figure) and RealSense (second figure). They are synchronized groupwise and their unit lengths are made uniform across groups. Writing behaviors are coded on the last two tiers in Sportscode time line.

kos, 2020), aligning different sort of data (sensing data and manipulation log in this case) can help instructors to assess learning performance or diagnose learners' struggle. In particular, the strong interrelation between interaction patterns from a touchscreen and high levels of cognitive workload illustrated in some previous research (Mock, et al., 2016) seems to be consistent with the findings of this study that participants' persistence in the acquisition of visual information indicates their struggle.

While the inexpensiveness of the depth camera offers a great potential for making the framework of this study scalable, there are several challenges associated with this research method, as noted in other related studies (Sharma and Giannakos, 2020; Ochoa, 2022). First, generalizing findings of this study to other learning themes and contexts is not straightforward. In fact, presumed thinking processes and behavioral patterns of learners heavily depend on the specific characteristics of these themes and contexts. Therefore, further refinement is necessary for the design and configuration of learning tools and sensing devices. Second, the participant population in this study is limited, and additional data are required to accurately analyze the thinking processes of learners from different populations. For instance, in Group IV of this study, the quality of participants' thinking was noticeably higher compared to other groups, despite their proof activity appearing stagnant. This suggests that aligning other types of learning data is essential to fully understand their learning processes. Further study across a wider range of learning themes and populations is needed.

Acknowledgements

This work was partially supported by JSPS Kakenhi 21K02752 and 22K02951.

References

Alibali, M. W., & Nathan, M. J. (2012). Embodiment in mathematics teaching and learning: Evidence from learners' and teachers' gestures. *The Journal of the Learning Sciences*, *21*, 247-286.

Blikstein, P., & Worsley, M. (2016). Multimodal learning analytics and educational data mining: Using computational technologies to measure complex learning tasks. *Journal of Learning Analytics, 3(2),* 220-238.

Dehaene, S., Spelke, E., Pinel, P., Stanescu, R., Tsivkin, S. (1999). Sources of mathematical thinking: Behavioral and brain-imaging evidence. *Science, 284*, 970-974.

Lakoff, G., & Nunez. R. E. (2001) *Where Mathematics Comes From*. Basic Books.

Mock, P., Gerjets, P., Tibus, M., Trautwein, U., Moller, K., & Rosenstiel, W. (2016). Using touchscreen interaction data to predict cognitive workload. *Proceedings of ICMI'18*, 349-356.

Nathan, M. J., Walkington, C., Boncoddo, R., Pier, E., Williams, C. C., & Alibali, M. W. (2014). Actions speak louder with words: The roles of action and pedagogical language for grounding mathematical proof. *Learning and Instruction, 33*, 182-193.

Ochoa, X. (2022). Multimodal learning analytics – Rationale, Process, Examples and Direction, in *Handbook of Learning Analytics* 22, 54-65.

Ochoa, X., Chiluiza, K., Mendez, G., Luzardo, G., Guaman, B., & Castells, J. (2013). Expertise estimation based on simple multimodal features. *Proceedings of ICMI'13*, 583-590.

Oviatt, S., Cohen, A., & Weibel, N. (2013). Multimodal learning analytics: Description of math data corpus for ICMI grand challenge workshop. *Proceedings of ICMI'13*, 563-568.

Paneth, L., Jeitziner, L. T., Rack, O., & Zahn, C. (2023). A multi-method approach to capture quality of collaborative group engagement. *Proceedings of CSCL'23,* 91-98.

Sharma, K., & Giannakos, M. (2020). Multimodal data capabilities for learning: What can multimodal data tell us about learning? *British Journal of Educational Technology*, 51-5, 1450-1484.

Spikol, D., Avramides, K., & Cukurova, M. (2016). Exploring the interplay between human and machine annotated multimodal learning analytics in hands-on STEM activities. *Proceedings of LAK'16,* 522-523.

Spikol, D., Ruffaldi, E., & Cukurova, M. (2017). Using multimodal learning analytics to identify aspects of collaboration in project-based learning. *Proceedings of CSCL'17,* 263-270.

Sweller, J., Ayres, P., & Kalyuga S. (2011). *Cognitive Load Theory.* Springer.

The Desirability of Peer-Assisted Learning Spaces for Online Student-Centered Error Correction: A Preliminary Evaluation

Fu-Yun YU
National Cheng Kung University, Taiwan
fuyun.ncku@gmail.com

Abstract: The educational value of error correction is well recognized. However, studies have found that students often need support to achieve optimal learning from this activity. In light of the documented benefits of collaborative learning, peer-assisted learning spaces were developed to augment a student-centered error-correction system that incorporates a generic seven-step framework as technological scaffolding. This paper briefly outlines the theoretical basis of the peer-assisted space before presenting a study assessing its desirability. The quantitative finding from the preliminary study—showing that significantly more participants preferred the peer-assisted over the individual mode, X^2= 11.115, p < .001—support the desirability of peer-assisted learning spaces for student-centered error correction. Students' explanations further form explicit suggestions for system design.

Keywords: Error correction, learning from errors, online learning space, peer-assisted learning, preferred learning mode, student-centered approach

1. Introduction

Through decades of scientific inquiry, psychologists and educational researchers have identified essential pedagogical components (Dick et al., 2022). Feedback to students following practice has been recognized as one of the most powerful instructional principles for promoting learning (American Psychological Association, Coalition for Psychology in Schools and Education, 2015).

With the distinctive affordances of modern technologies, systems that support student practice with feedback are now abundant. While providing clear, automatic, explanatory feedback by such systems is technically feasible and financially viable, this approach does not support self-explanation. The self-explanation principle posits that people learn more deeply and meaningfully when they generate explanations (Roy & Chi, 2005) through the externalization process of articulating or justifying what, why, and how answers are given (Chi et al., 1989). Furthermore, given the challenges of error correction—such as students generally engaging in shallow behaviors and lacking the knowledge and strategies needed for productive outcomes) (Suarez, 2013)—we developed the world's first online drill-and-practice system designed to anchor student-centered error-correction activities. In this system, a generic seven-step error-correction framework—serving as technological scaffolding—was devised to guide students through the process (Yu, 2024).

Later, realizing that students often seek support from experts despite the guidance provided in technology-supported learning environments (Mercer & Fisher, 1992), we further enhanced the system by adding social scaffolding in the form of peer-assisted error correction. As the desirability of the incorporated peer-assisted space is not yet known, this study aims to assess that desirability—specifically for supporting error-correction activities—an aspect considered important in product evaluation (Bland & Osterwalder, 2020).

2. The Development of an Online Student-Centered Error-Correction System

with Peer-Assisted Learning Spaces

The theoretical basis for incorporating a peer-assisted component into the system is Puntambekar and Kolodner's (1998) distributed scaffolding theory, which suggests that integrating various agents and multiple sources of scaffolding can maximize learning outcomes. In our system, students can solicit assistance from their peers through the incorporated call-out-for-help function. As illustrated in Figure 1, once a request is entered in the designated space during any of the error-correction steps, an automatic notification is sent to initiate timely, on-demand peer assistance.

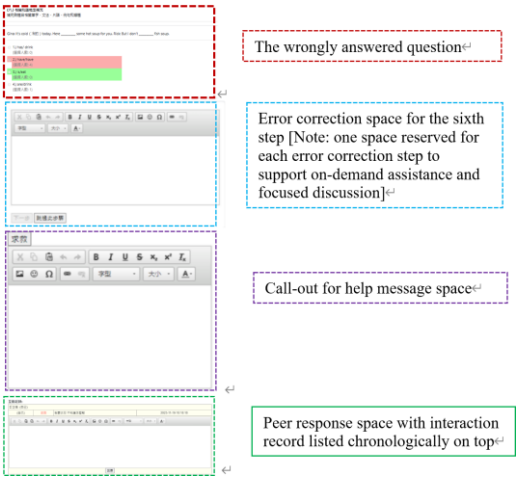


Figure 1. Peer-assisted interaction space enabling assistance seeking and provision during the error-correction process

3. A Preliminary Study on the Desirability of Peer-Assisted Learning Spaces for Online Student-Centered Error Correction

To examine the desirability of the incorporated peer-assisted space, four classes of sixth-grade students (*n*=104) taught by the same social science teacher at a primary school in central Taiwan were invited to participate. The error-correction activities were integrated into the classes' weekly drill-and-practice sessions to support learning. During this study, two of the three 40-minute instructional sessions each week were reserved for instruction delivery, and one was scheduled for online drill-and-practice and error correction.

In the first week, the session was used primarily to familiarize and equip students with the essential knowledge and skills needed to complete the tasks in the online drill-and-practice and error-correction systems. Starting in the second week, the weekly online activities followed a set routine: the teacher first provided whole-class feedback on student performance from the previous week, after which students engaged in online drill-and-practice, followed by online error-correction activities. For the online drill-and-practice activity, students answered 10-item multiple-choice questions on the instructional content covered that week. For the error-correction activity, students corrected their incorrectly answered questions by following the step-by-step procedures outlined in the system. To allow comparisons based on concrete personal experiences, students engaged in individual error correction in the second week before transitioning to peer-assisted error correction in the subsequent weeks. Because the size of interacting groups may affect students' experiences, a small group peer-assisted setting was used in the third week—with the system automatically forming groups of four—while a large group peer-assisted setting was used in the fourth week, in which all classmates could help each other in a single online interaction space. At the end of the fourth week, students responded to the question: *Which error-correction mode do you prefer—individual or peer-assisted? Please justify your answer*.

Quantitative data indicated that nearly two-thirds of the participants (66.35%) preferred the peer-assisted mode over the individual mode. A chi-square test further revealed a

statistically significant difference at the 0.001 level, X^2 = 11.115. Analysis of the explanatory descriptions revealed several major themes. Among those favoring the individual mode, the predominant theme was *accountability for one's own learning*—specifically, taking responsibility for finding the correct answer or source of information to promote deeper learning. For supporters of the peer-assisted mode, the most salient theme was *the value of mutual support among classmates in achieving deeper understanding and fostering close bonds*. Additionally, while many students appreciated *having the whole class at one's disposal* rather than limited to their group, quite a few reported feeling overwhelmed by numerous calls for help awaiting attention; as a result, they preferred the smaller-group arrangement within the peer-assisted mode.

4. Conclusion

Based on the findings, two suggestions are offered. First, online spaces for both individual and peer-assisted error correction should be built to accommodate students' differing needs while actualizing the concept of scaffolding theory—support-provision for those not yet ready for individual work and support-fading for those who have reached the desired state (Collins et al.,1987). Second, spaces supporting peer-assisted error correction with varying group sizes (e.g., small, large) may be desirable, ensuring that timely, on-demand assistance can be provided without overwhelming the help-giver and making the task more manageable for those involved.

Acknowledgements

This work was supported by the Ministry of Science and Technology, Taiwan, under Grant No. 111-2410-H-006-023-MY3.

References

American Psychological Association, Coalition for Psychology in Schools and Education. (2015). *Top 20 principles from psychology for preK–12 teaching and learning*. Retrieved from http://www.apa.org/ed/schools/cpse/top-twenty-principles.pdf

Bland, D. J., & Osterwalder, A. (2020). *Testing business ideas: A field guide for rapid experimentation*. Wiley.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*, 145–182.

Collins, A., Brown, J. S., & Newman, S. E. (1987). *Cognitive apprenticeship: Teaching the craft of reading, writing, and mathematics.* Center for the Study of Reading. University of Illinois at Urbana-Champaign.

Dick, W., Carey, L., & Carey, J. O. (2022). *The systematic design of instruction* (9th ed.). Pearson

Mercer, N., & Fisher, E. (1992). How do teachers help children to learn? An analysis of teachers' interventions in computer-based activities. *Learning and instruction, 2*(4), 339-355.

Puntambekar, S., & Kolodner, J. L. (1998). Distributed scaffolding: Helping students learn in a "learning by design" environment. In A. S. Bruckman, M. Guzdial, J. L. Kolodner, & A. Ram (Eds.), *Proceedings of the Third International Conference of the Learning Sciences* (ICLS'98) (pp. 35–41). Atlanta, GA: Association for the Advancement of Computing Education.

Roy, M., & Chi, M. T. H. (2005). The self-explanation principle. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning*. Cambridge University Press.

Suarez, A. R. (2013). *The effect of teacher's error correction strategies on students' writing* [Unpublished master thesis]. National Autonomous University of Nicaragua.

Yu, F. Y. (2024). Theory-driven design for the development of a student-centered error-correction online learning system. Kashihara, A., Jiang, B., Rodrigo, M. M., & Sugay, J. O. (Eds.), *Conference Proceedings of the 32nd International Conference on Computers in Education Volume I* (pp. 806-808), November 25-29, Manila, Philippines.

Development of a Maker Education Instructional Design Model: Delphi-based Analysis

Xinru LIU[a], Lei WANG[b] & Guang CHEN[c*]
[a] *Faculty of Psychology, Beijing Normal University, China*
[b] *Beijing Normal University Tianjin Eco City Affiliated School, China*
[c] *Faculty of Education, Beijing Normal University, China*
*guang@bnu.edu.cn

Abstract: This study aims to develop a Maker education instructional design model that can improve students' comprehensive ability and creativity. This study constructs a Maker education instructional design model comprising four stages: needs analysis, content selection, process design, and effect evaluation. The model's effectiveness is validated through the development and implementation of Computer Aided Manufacturing for Education (CAME) courses.

Keywords: Maker education instructional design model, Delphi Method, Computer Aided Manufacturing for Education

1. Introduction

Our learning environment has become intelligent learning environment. The combination of the STEAM education with interdisciplinary characteristics and the Maker education instructional design model with the purpose of cultivating creativity is of great help to the cultivation of compound innovative talents.

Researchers had put efforts on the development framework of the Maker education instructional design. Jia et al. (2021) integrated STEAM in engineering design and Maker education into primary school curriculum to promote students' comprehensive capabilities. Combining the 3D technology of Kinesthetic Education Mobile Studio (MAKE) and the mobile maker space platform, to attract novice users from all places to participate in the practical activities of art creation(Jordan et al., 2021) . However, for the practical needs of curriculum development, there is no consensus on an operational development model and development process.

We took Computer Aided Manufacturing for Education (CAME) as an example to build a complete Maker education instructional design model. We hope that these experiences can be used as reference by international peer.

2. Research design and method

This study used the revised Delphi Method to build the relevant indicator system of the Maker education instructional design model, and implements a round of semi-structured interviews and two rounds of questionnaires. Finally, experts reached an agreement and screen out the key indicators of the Maker education instructional design model.

This study constructed a Maker education instructional design model and applied it to develop the Computer Aided Manufacturing for Education course. Using the Delphi Method, 9 experts participated in interviews and 20 in questionnaire surveys. Interviews explored: (1) previous STEAM/maker curriculum development experience, (2) development processes, (3) influencing factors, and (4) evaluation approaches. Round 1 questionnaire: 31 items derived

from literature and interviews using 5-point Likert scales. Round 2 questionnaire: 19 retained items rated for importance (1=very unimportant to 5=very important). Round 2 questionnaire: 19 retained items rated for importance (1=very unimportant to 5=very important). Nineteen indicators were identified across three categories: needs analysis (13 indicators), development process (3 indicators), and evaluation (3 indicators). The model was then applied to develop the Computer Aided Manufacturing for Education course.

3. Research Findings

After two rounds of questionnaires, the consensus rate of experts on the second round of questionnaires was 89.47%. It is higher than the data 80% mentioned in study (Kim et al., 2022). This indicates that the experts have reached an agreement, so the investigation was stopped.

3.1 Construction of Maker education instructional design model

Maker education solves real-life problem environments by providing problem-solving methods (Kim et al., 2022). In this study, the Maker education instructional design model was carried out in four stages: needs analysis, content design, process design, and effect evaluation.

The first stage is needs analysis, examining objects from nations, society, schools, to students and teachers. This involves analyzing curriculum standards, students' knowledge base, school educational resources and class hour support. The second stage is content design, matching national curriculum standards of relevant subjects. The knowledge point map covers S, T, E, A, M perspectives, using different marks to show students' acquired knowledge points, their relationships, and target knowledge points. The third stage is process design, including learning activity design and learning path design. This emphasizes students' subjectivity and activities based on student characteristics, followed by collaborative work sharing. The fourth stage is effect evaluation, conducting formative evaluation of course effectiveness through work evaluation and team collaboration ability. Work evaluation focuses on three aspects: innovation, technology and exploration. If results are unsatisfactory, the process returns to the first stage for iterative improvement.

Based on this analysis, the Maker education instructional design model is shown in Figure 1. Solid rectangles represent necessary processes, dotted rectangles represent optional processes.

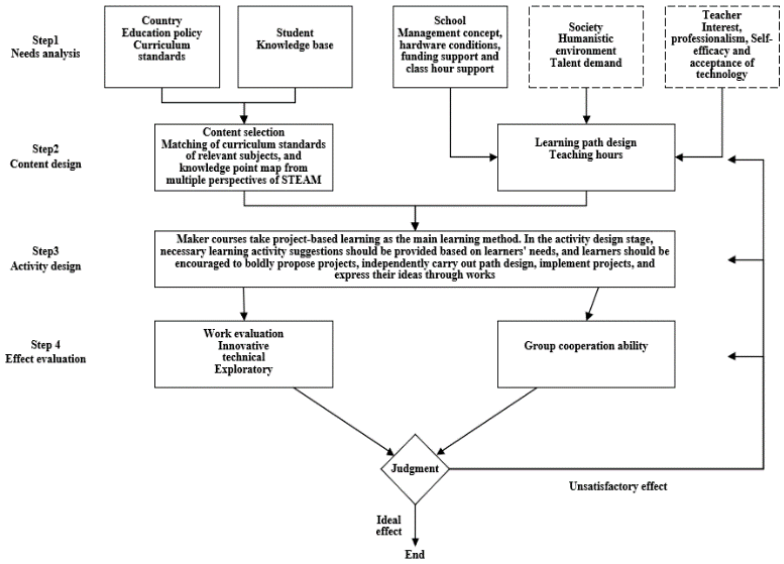


Figure 1. Maker education instructional design Model Diagram.

3.2 Computer Aided Manufacturing for Education (CAME)

The core goal of CAME is to introduce computer-aided manufacturing technology and related equipment into the classroom, and help students express their ideas through works. The content of the textbook is divided into three levels, mainly including the basis of Auto CAD and drawing commands, basic operations of drawing and graphic editing, word games and various object designs. However, the existing textbooks emphasize the operation process too much and how to design and implement scientific knowledge are not clearly reflected.

3.3 Curriculum reconstruction based on this research's curriculum development model

In order to deal with the problems existing in the available curriculum and improve the learning effect of students, we reconstructed the existing curriculum according to the previous Maker education instructional design model.

The needs analysis stage analyzes educational resources and learner characteristics. Schools need maker laboratories and computer rooms with laser cutting machines and class-related materials. The course targets students from fourth grade primary to third grade junior high school. Due to significant differences in cognitive development and independent exploration willingness between elementary and junior high students, two paths are needed: primary and advanced. In content design, knowledge points are extracted based on lesson content and students' existing knowledge base to select learning content. In process design, primary students receive a brief teacher demonstration before independent exploration, whereas junior students work with minimal guidance; typical project-based activities include sketch → CAD → laser-cut → test, with rotating roles to support collaboration. Effect evaluation focuses on artifact quality and team collaboration.

4. Conclusion

Using Delphi method, this study identified 19 indicators and constructed a Maker education instructional design model verified through CAME development.

While expert consensus eliminated some student analysis indicators, practical application revealed their necessity. This highlights that comprehensive needs analysis remains foundational to effective instructional design, even when challenging to implement. The model demonstrates that content selection must align with maker education principles through project-based, collaborative activities.

Future research should empirically verify the model's effectiveness and examine specific indicators' impact on implementation.

References

Y. Jia, B. Zhou, and X. Zheng, “A Curriculum Integrating STEAM and Maker Education Promotes Pupils’ LeJia, Y., Zhou, B., & Zheng, X. (2021). A Curriculum Integrating STEAM and Maker Education Promotes Pupils’ Learning Motivation, Self-Efficacy, and Interdisciplinary Knowledge Acquisition. Frontiers in Psychology, 12, 725525. https://doi.org/10.3389/fpsyg.2021.725525

Jordan, A., Knochel, A. D., Meisel, N., Reiger, K., & Sinha, S. (2021). Making on the Move: Mobility, Makerspaces, and Art Education. International Journal of Art & Design Education, 40(1), 52–65. https://doi.org/10.1111/jade.12333

Kim, J.-Y., Seo, J. S., & Kim, K. (2022). Development of novel-engineering-based maker education instructional model. Education and Information Technologies, 27(5), 7327–7371. https://doi.org/10.1007/s10639-021-10841-4

C3
Advanced Learning Technologies, Learning Analytics, Platforms and Infrastructure (ALT/LA/P/I)

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education.
AsiaPacific Society for Computers in Education

Impact of LLM Feedback on Learner Persistence in Programming

Yiqiu ZHOU[a*], Maciej PANKIEWICZ[b], Luc PAQUETTE[a] & Ryan S. BAKER[b]
[a]*University of Illinois Urbana-Champaign, USA*
[b]*University of Pennsylvania, USA*
*yiqiuz3@illinois.edu

Abstract: This study examines how Large Language Model (LLM) feedback generated for compiler errors impacts learners' persistence in programming tasks within a system for automated assessment of programming assignments. Persistence, the ability to maintain effort in the face of challenges, is crucial for academic success but can sometimes lead to unproductive "wheel spinning" when students struggle without progress. We investigated how additional LLM feedback based on the GPT-4 model, provided for compiler errors affects learners' persistence within a CS1 course. Specifically, we examined whether its impacts differ based on task difficulty, and if the effects persist after the feedback is removed. A randomized controlled trial involving 257 students across various programming tasks was conducted. Our findings reveal that LLM feedback improved some aspects of students' performance and persistence, such as increased scores, a higher likelihood of solving problems, and a lower tendency to demonstrate unproductive "wheel spinning" behavior. Notably, this positive impact was also observed in challenging tasks. However, its benefits did not sustain once the feedback was removed. The results highlight both the potential and limitations of LLM feedback, pointing out the need to promote long-term skill development and learning independent of immediate AI assistance.

Keywords: Autograding, Automated feedback, Programming, Persistence, GPT, LLM

1. Introduction

Persistence, the ability to maintain effort in the face of challenges, is a critical factor in student success and learning outcomes (Cloninger et al., 1993). This concept is closely linked to grit, productive struggle, and growth mindset that encourage persistent efforts (Cutts et al., 2010; Duckworth et al., 2007). Previous studies have demonstrated that higher persistence predicts course completion (Wang & Baker, 2018), college enrollment (Adjei et al., 2021) and academic success (McClendon et al., 2017). Persistence can be particularly relevant to computer science education, as programming often involves debugging. This process can be challenging due to various factors such as misconceptions, fragile knowledge, and the 'superbug' (the incorrect assumption that the programming language has intelligent interpretive powers) (McCauley et al., 2008). Debugging often requires students to engage in iterative testing to identify and resolve issues, demanding persistence as students must maintain their efforts despite the frustration and setbacks.

However, not all persistence in programming tasks is productive. Students may exhibit unproductive persistence, or "wheel spinning." This occurs when students make repeated unsuccessful attempts without making meaningful progress or achieving mastery within a reasonable timeframe (Beck & Gong, 2013). Wheel spinning is particularly prevalent among novice programmers who struggle with syntax and find it difficult to interpret compiler error messages (Hartmann et al., 2010). Without appropriate guidance, students may misinterpret error messages, apply ineffective debugging strategies, or repeatedly attempt incorrect solutions, hindering their learning (Nienaltowski et al., 2008).

Recent advancements in artificial intelligence, particularly Large Language Models (LLMs) like GPT, offer new opportunities for personalized feedback in CS education. Examples include Codehelp (Liffiton et al., 2023) and Codeaid (Kazemitabaar et al., 2024). LLMs have also been used to generate code explanations (Leinonen, Denny, et al., 2023), and debugging hints (Leinonen, Hellas, et al., 2023; McBroom et al., 2021). However, recent studies suggest that novices may struggle to interpret AI-generated feedback, leading to cognitive overload and ineffective problem-solving strategies (Leinonen, Hellas, et al., 2023). Concerns have also emerged about students' potential over-reliance, where students merely copy AI-suggested solutions without fully understanding them, hindering their ability to internalize programming concepts and independently solve problems (Becker et al., 2023; Finnie-Ansley et al., 2022). This highlights the need for further investigation into whether LLM feedback fosters genuine problem-solving or reinforces unproductive behaviors.

This study contributes to this discussion by examining whether LLM-generated feedback enhances learners' productive persistence in programming tasks and whether its benefits persist after AI assistance is withdrawn. We hypothesize that the personalized LLM feedback could help students engage in more effective debugging and problem-solving strategies while mitigating unproductive persistence (i.e., repeated unsuccessful attempts). To rigorously test this hypothesis, we conducted a randomized controlled experiment with 257 CS1 students using an automated assessment platform. Students were randomly assigned to receive either standard compiler error messages or standard compiler error messages plus LLM-enhanced feedback with error explanation, allowing us to compare their persistence behaviors across programming tasks of varying difficulty. Beyond assessing the immediate effects of AI-generated feedback, our study uniquely examines its long-term impact by analyzing student performance after LLM feedback is removed. Specifically, we examine:

RQ1: How does LLM feedback affect learners' persistence in programming tasks?

RQ2: Does LLM feedback show a significant effect on learners' persistence, specifically in challenging programming tasks?

RQ3: How does the removal of LLM feedback affect the performance of students who have previously received it? Are there lasting effects on their persistence?

2. Literature Review

2.1 Operationalization of Persistence

Persistence research has recently expanded to computer science domain (Flores & Rodrigo, 2020; Pinto et al., 2021), but the iterative nature of programming poses challenges for traditional models. Programming involves cycles of coding, testing, and debugging (McCauley et al., 2008). Common approaches like correct-in-a-row count (Gong & Beck, 2015) or Bayesian Knowledge Tracing (BKT) (Käser et al., 2016) often assume binary correctness and clear skill mapping, which do not reflect the multifaceted progress typical in programming tasks. In programming tasks, students often make incremental progress in various aspects of the code (e.g., syntax, logic, efficiency) simultaneously. Another challenge is the issue of credit assignment. While intelligent tutoring systems (ITSs) use well-defined skill structure to track learning (Fang et al., 2017; Gong & Beck, 2015), programming tasks often involve overlapping concepts, making it hard to pinpoint which specific skill or concept underlies errors. As a result, the criteria traditionally used to define mastery and differentiate between productive and unproductive persistence in ITS may not directly translate to the programming context. Thus, this work leverages the metrics defined in Pinto et al. (2021), which are tailored to measure persistence in programming tasks. This approach considers the characteristics of programming tasks and provides a more nuanced view of persistence in programming.

2.2 LLM in CS education and Research Motivation

LLM have shown promise in CS education by providing explanations for programming concepts and errors, and personalized hints (Kazemitabaar et al., 2024; Leinonen, Denny, et al., 2023; Liffiton et al., 2023; McBroom et al., 2021; Sit et al., 2024). In introductory programming courses, ChatGPT has been found to deliver consistent formative feedback (Kiesler et al., 2023) and improve task performance with positive student ratings (Pankiewicz

& Baker, 2023). However, limitations remain: GPT-based tools may struggle with complex queries and software testing questions (Qureshi, 2023) and can yield mixed results, especially when deeper understanding of programming concepts is required (Pankiewicz & Baker, 2024). Moreover, concerns around over-reliance have emerged (Xue et al., 2024). One risk is the illusion of understanding, where students believe they have grasped a concept simply because they can produce working solutions with AI (Becker et al., 2023). Tools like OpenAI Codex may enable students to bypass the problem-solving process instead of engaging in debugging and critical thinking (Finnie-Ansley et al., 2022). Research on Copilot has similarly shown that some students develop a habit of outsourcing their cognitive effort to AI (Prather et al., 2023). These concerns highlight that unregulated use may also undermine the acquisition of problem solving skills. Recent research has shown that students' intentions to use these tools are shaped not only by immediate needs such as academic stress and risk-taking tendencies (Samson et al., 2024) but also by broader psychological factors such as personal innovativeness, trust, and self-efficacy (Elinzano & Ching, 2024).

This study aims to bridge the gap between the need for persistence in programming and the potential of AI-generated feedback. Most existing studies on AI-generated feedback have focused primarily on immediate learning outcomes, with limited attention to whether these benefits continue after the AI support is withdrawn. This study examines how LLM feedback impacts student persistence both immediately and over the long term. Specifically, we investigate whether LLM feedback reduces unproductive persistence and whether this benefit persists once removed. By examining the interaction between LLM feedback and task difficulty, this study also aims to inform best practices regarding when and how AI assistance should be used. Ultimately, this research provides evidence-based insights into the conditions under which AI-generated feedback supports meaningful learning in programming.

3. Method

3.1 Participant and Learning Environment

The study involved 257 computer science students, enrolled in a mandatory first semester CS1 course at a large European university. Topics include types and variables, conditional statement, recursion, loops, and loops with arrays. Students were randomly assigned to the Experimental group (N=129) or the Control group (N=128). Participants comprised 182 males and 74 females, with gender information missing for one participant. Consent was obtained from students. Tasks were assigned on an online platform for automated assessment, through 14 modules containing between 5 to 14 problems each, with modules typically assigned on a weekly basis. The platform's assessment process involved compiling and testing the submitted code written in C# language, followed by providing students with feedback that included: compiler messages (line, error message, and id), results for each unit test (with input values and expected outcome), and overall score (the percentage of successful unit tests 0-100%).

3.2 Experiment Setting and Prompt Design

Both the control and experimental groups had access to 141 programming tasks distributed across 14 modules (as shown in Figure 1). For the first half of each module (72 tasks), the experimental group received additional feedback on compiler errors from OpenAI's GPT-4 model (gpt-4-0613), while the control group received only standard messages. For the remaining 69 tasks, the GPT-4 feedback was withdrawn for the experimental group to assess any lasting effects. Feedback was created via a request to the OpenAI Chat Completion API using the approach to prompt generation that consisted of the following parts: 1) general instructions for the assistant, 2) assignment text, 3) student code, and 4) results of the code evaluation. The prompt was extended by examples of an ideal response containing three elements (in English): an explanation of the error (example from one of generated hints: "The compiler message *;* expected means that a semicolon is missing in the line *int a = 2*b*"), a solution strategy (example: "To fix this error, you need to add a semicolon *;* at the end of the line int *a = 2*b*"), and an educational element regarding the underlying concept (example: "Remember that in C#, every statement must be terminated with a semicolon"). This structure was designed to provide scaffolded support that guided students toward a solution without

offering explicit code corrections. Hints were generated in approximately 20 seconds as a re-viewable pop-up and offered once per submission, with no limit on attempts. The researchers also conducted a qualitative review of a sample of hints to ensure their general accuracy and pedagogical value.

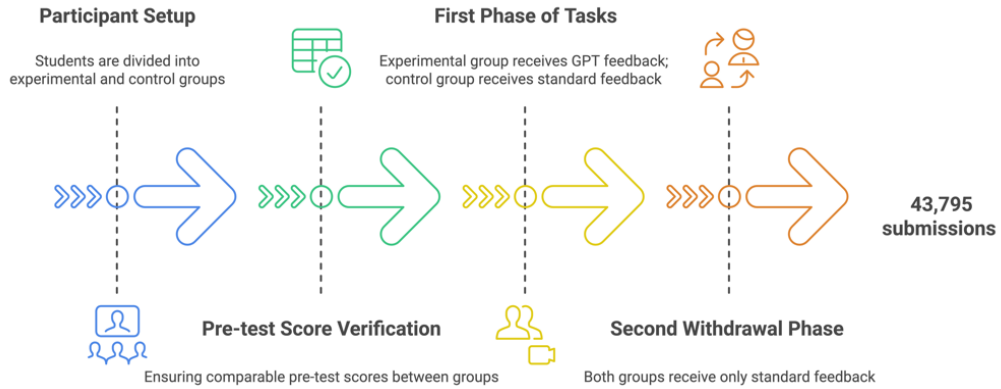


Figure 1. Experiment Setting

3.3 Persistence Elements

This study utilizes log data from 43,795 submissions to extract eight features, adapted from (Pinto et al., 2021), that quantify various aspects of persistence. The choice of thresholds—such as the 75th percentile—was informed by prior work (Pinto et al., 2021). For time-related features, we subtracted the actual LLM feedback generation time or used an average of 20 seconds when this record was unavailable. These features were computed per task and then aggregated to the student level by averaging across tasks. Prior to aggregation, all features were standardized within each task (i.e., z-scored across all students) to account for variation in task difficulty and baseline behavior. The final feature list consisted of:

Number of submissions: The total number of submission attempts made by the student for a specific task.

Beyond attempt threshold: The proportion of tasks for which a student exceeded the 75th percentile of submission attempts (calculated across all students for each task).

Solved: The proportion of attempted tasks that a student successfully solved, where a task is considered solved if the final submission received a perfect score.

Last submission score: The average score of a student's final submissions across all attempted tasks.

Time since first submission: The average time a student spent on tasks, calculated from their first submission to their last for each task. For tasks with LLM assistance, time is adjusted to account for LLM feedback generation.

Beyond time threshold: The proportion of tasks for which a student's total time spent exceeded the 75th percentile of time spent (calculated across all students for each task).

Taking a break: The proportion of tasks where a student had at least one gap of two hours or more between consecutive submissions. This two-hour threshold was determined based on an inspection of the distribution of submission intervals across all students, where breaks of this length represented a noticeable departure from typical short pauses.

Wheel spinning: The proportion of tasks where a student made 7 or more submissions and didn't solve the problem (final submission score was below 1). The 7-submission threshold corresponds to the mean plus one standard deviation of attempt counts in our dataset, aligning with prior work to identify patterns of unproductive persistence (Gong & Beck, 2015).

These features were designed to capture different aspects of student persistence, including effort (number of submissions), efficiency (time spent), success rate (solved tasks), and potential struggles (wheel spinning).

3.4 Analysis

We employed rank-based linear regression from the Rift package in R due to non-normality in the data (Kloke & McKean, 2012), which is robust to non-normal data distributions. To further assess the stability and reliability of the model estimates, we conducted Monte Carlo simulations with 1000 iterations for each model. These simulations allowed us to evaluate the

sensitivity of our results to sampling variability and ensure inference robustness. Additionally, we verified that there were no significant differences in the pre-test scores between the control and experimental groups using the Mann-Whitney U test (U = 4441.0, p = 0.121). The persistence metrics, outlined in Section 3.4, served as dependent variables. Each regression model included two independent variables: feedback type (additional LLM feedback vs. control) and pre-test scores, the latter serving as a covariate to control for prior knowledge differences. Our analysis addressed three specific research questions. **Impact on All Tasks (RQ1)**: The impact of LLM feedback was evaluated across all tasks. **Impact on Challenging Tasks (RQ2)**: Challenging tasks were defined by three criteria: average score below 100%, total time spent above the 75th percentile of all tasks, and total submission attempts above the 75th percentile. By analyzing these tasks separately, we aimed to gain nuanced insights into the impacts of LLM feedback on learners' persistence when faced with more difficult programming tasks. **Post-feedback Removal Effects (RQ3)**: The persistence of observed benefits was examined after LLM feedback was removed.

4. Results

4.1 RQ1: Impact on All Tasks

To investigate the impact of additional LLM feedback on learners' persistence, we employed rank-based regression to analyze each persistence metric, with pre-test scores as a covariate. The results are presented in Table 1, which reports regression coefficients as the estimated effect of group membership on persistence metrics, Monte Carlo standard errors as the measure of estimation precision, and p-values to indicate statistical significance. We also calculated Kendall's Tau τ as the rank-based measure of association strength and direction. This ranges from -1 (perfect negative correlation) to +1 (perfect positive correlation), with values closer to 0 indicating weaker relationships. Tau values near ±0.10 suggest weak associations, ±0.30 moderate associations, and ±0.50 or higher strong associations. The experimental group, for tasks with GPT feedback enabled, showed significantly higher last submission scores (β = 0.029, p = 0.041, τ = 0.137) and a marginally higher likelihood of solving the problems (β = 0.028, p = 0.061, τ = 0.095) compared to the control group. Furthermore, the experimental group exhibited significantly lower rates of going beyond the time threshold (β = -0.098, p =0.00, τ = -0.143), taking breaks (β = -0.035, p = 0.050, τ = -0.089), and wheel spinning (β = -0.025, p = 0.032, τ = -0.134) compared to the control group. These findings indicate that learners who received GPT feedback were less likely to spend excessive time on tasks, take breaks, or engage in unproductive persistence (i.e., wheel spinning) than those who received only standard compiler error messages.

Table 1. *Combined Regression Analysis for All Tasks*

Variable	**GPT Feedback Enabled (Phase 1)**			**GPT Feedback Disabled (Phase 2)**		
	Coefficient (Std.E)	MC Mean (Std.D)	p-value	Coefficient (Std.E)	MC Mean (Std.D)	p-value
Number of submissions (count)						
Intercept	7.192 (0.479)	7.256 (0.487)	< 0.001***	7.101 (0.446)	7.082 (0.518)	< 0.001***
Experiment	0.381 (0.417)	-0.439 (0.412)	0.363	0.460 (0.391)	0.441 (0.411)	0.242
Pre-test	-0.556 (0.090)	-0.550 (0.075)	< 0.001***	-0.438 (0.084)	-0.433 (0.085)	< 0.001***
Beyond Attempt Threshold (proportion)						
Intercept	0.141 (0.015)	0.141 (0.016)	< 0.001***	0.120 (0.019)	0.123 (0.020)	< 0.001***
Experiment	-0.014 (0.014)	-0.016 (0.014)	0.307	0.029 (0.018)	0.028 (0.020)	0.113
Pre-test	-0.019 (0.003)	-0.019 (0.003)	< 0.001***	-0.014 (0.004)	-0.014 (0.003)	< 0.001***
Solved (proportion)						
Intercept	0.866 (0.019)	0.868 (0.021)	< 0.001***	0.840 (0.018)	0.838 (0.018)	< 0.001***
Experiment	0.028 (0.015)	0.029 (0.016)	0.061	0.016 (0.016)	0.015 (0.016)	0.324

Pre-test	0.011 (0.003)	0.011 (0.003)	0.0005***	0.017 (0.003)	0.017 (0.003)	< 0.001***
Last Submission Score (0–1 scale)						
Intercept	0.913 (0.015)	0.911 (0.015)	< 0.001***	0.894 (0.020)	0.892 (0.022)	< 0.001***
Experiment	0.029 (0.014)	0.029 (0.014)	0.041*	0.016 (0.019)	0.015 (0.019)	0.402
Pre-test	0.009 (0.003)	0.009 (0.003)	0.003**	0.012 (0.004)	0.012 (0.004)	0.003**
Time Since First Submission (seconds)						
Intercept	459.217 (29.760)	464.509 (45.946)	< 0.001***	497.061 (39.856)	501.579 (44.586)	< 0.001***
Experiment	-32.823 (26.028)	-35.152 (29.838)	0.209	43.394 (34.758)	39.889 (35.614)	0.214
Pre-test	-43.127 (5.602)	-42.788 (7.031)	< 0.001***	-41.959 (7.440)	-42.536 (7.016)	< 0.001***
Beyond Time Threshold (proportion)						
Intercept	0.886 (0.034)	0.882 (0.028)	< 0.001***	0.852 (0.032)	0.871 (0.034)	< 0.001***
Experiment	-0.098 (0.031)	-0.101 (0.034)	0.002**	0.013 (0.028)	0.006 (0.032)	0.645
Pre-test	-0.047 (0.007)	-0.047 (0.007)	< 0.001***	-0.036 (0.006)	-0.042 (0.007)	< 0.001***
Taking a Break (proportion)						
Intercept	0.207 (0.023)	0.214 (0.035)	< 0.001***	0.190 (0.023)	0.198 (0.037)	< 0.001***
Experiment	-0.035 (0.018)	-0.036 (0.020)	0.050*	-0.012 (0.020)	-0.012 (0.022)	0.543
Pre-test	-0.025 (0.004)	-0.025 (0.005)	< 0.001***	-0.024 (0.004)	-0.024 (0.005)	< 0.001***
Wheel Spinning (proportion)						
Intercept	0.086 (0.012)	0.084 (0.015)	< 0.001***	0.088 (0.016)	0.084 (0.021)	< 0.001***
Experiment	-0.025 (0.012)	-0.028 (0.014)	0.032*	0.019 (0.015)	0.020 (0.019)	0.196
Pre-test	-0.011 (0.003)	-0.010 (0.003)	< 0.001***	-0.013 (0.003)	-0.012 (0.003)	< 0.001***

** Note: Coefficient (Std.E): The coefficient value and its standard error for the regression analysis. MC Mean (Std.D): The mean value and its standard deviation obtained from Monte Carlo simulations.*

4.2 RQ2: Impact on Challenging Tasks

The results for the challenging tasks with LLM feedback enabled for the experimental group are presented in Table 2. The analysis revealed that the experimental group had significantly higher last submission scores (β = 0.059, p = 0.013, τ = 0.183) and a marginally higher likelihood of solving the challenging problems (β = 0.097, p = 0.059, τ = 0.148) compared to the control group. These findings suggest that the addition of LLM feedback may have helped learners achieve better final outcomes and solve rates in challenging programming tasks. However, no significant differences were found between the two groups in terms of the number of submissions, time spent on the task, going beyond the time or attempt threshold, or taking breaks. Overall, these results suggest that the impact of LLM feedback on learners' persistence can be observed in challenging tasks, particularly in terms of achieving better final outcomes and solving the problems.

Table 2. *Combined Regression Analysis for Challenging Tasks*

Variable	**GPT Feedback Enabled (Phase 1)**			**GPT Feedback Disabled (Phase 2)**		
	Coefficient (Std.E)	MC Mean (Std.D)	p-value	Coefficient (Std.E)	MC Mean (Std.D)	p-value
Number of submissions (count)						
Intercept	11.716 (0.887)	11.689 (1.138)	< 0.001***	11.092 (1.144)	11.092 (1.144)	< 0.001***
Experiment	0.681 (0.823)	0.696 (0.879)	0.409	0.925 (1.061)	0.873 (1.190)	0.385
Pre-test	-1.043 (0.177)	-1.032 (0.177)	< 0.001***	-0.802 (0.230)	-0.802 (0.230)	0.001***
Beyond Attempt Threshold (proportion)						

Intercept	0.251 (0.032)	0.247 (0.044)	< 0.001***	0.257 (0.038)	0.244 (0.042)	< 0.001***
Experiment	0.030 (0.028)	0.030 (0.030)	0.277	0.019 (0.033)	0.020 (0.035)	0.570
Pre-test	-0.030 (0.006)	-0.029 (0.007)	< 0.001***	0.023 (0.007)	-0.022 (0.007)	0.002**
Solved (proportion)						
Intercept	0.638 (0.058)	0.650 (0.054)	< 0.001***	0.615 (0.049)	0.621 (0.047)	< 0.001***
Experiment	0.097 (0.051)	0.097 (0.050)	0.059	0.034 (0.043)	0.030 (0.045)	0.436
Pre-test	0.024 (0.011)	0.023 (0.010)	0.028*	0.043 (0.009)	0.042 (0.008)	< 0.001***
Last Submission Score (0–1 scale)						
Intercept	0.824 (0.029)	0.822 (0.032)	< 0.001***	0.842 (0.022)	0.841 (0.030)	< 0.001***
Experiment	0.059 (0.023)	0.062 (0.025)	0.013*	0.003 (0.020)	0.002 (0.024)	0.892
Pre-test	0.014 (0.005)	0.014 (0.005)	0.005**	0.019 (0.004)	0.018 (0.005)	< 0.001***
Time Since First Submission (seconds)						
Intercept	964.237 (78.881)	951.294 (101.410)	< 0.001***	1095.265 (103.008)	1107.270 (130.428)	< 0.001***
Experiment	56.962 (77.068)	47.408 (91.108)	0.461	28.874 (92.823)	22.494 (92.560)	0.756
Pre-test	-104.134 (16.567)	-99.793 (17.451)	< 0.001***	-97.333 (20.135)	-95.936 (22.104)	< 0.001***
Beyond Time Threshold (proportion)						
Intercept	1.034 (0.037)	1.034 (0.029)	< 0.001***	1.035 (0.039)	1.031 (0.027)	< 0.001***
Experiment	-0.007 (0.036)	-0.009 (0.042)	0.847	-0.084 (0.038)	-0.084 (0.041)	0.028*
Pre-test	-0.040 (0.008)	-0.040 (0.009)	< 0.001***	-0.029 (0.008)	-0.030 (0.008)	< 0.001***
Taking a Break (proportion)						
Intercept	0.382 (0.059)	0.381 (0.060)	< 0.001***	0.371 (0.069)	0.361 (0.075)	< 0.001***
Experiment	0.018 (0.055)	0.020 (0.060)	0.739	-0.034 (0.063)	-0.029 (0.061)	0.591
Pre-test	-0.055 (0.012)	-0.053 (0.011)	< 0.001***	-0.043 (0.014)	-0.042 (0.012)	0.002**
Wheel Spinning (proportion)						
Intercept	0.175 (0.048)	0.170 (0.050)	< 0.001***	0.209 (0.046)	0.210 (0.059)	< 0.001***
Experiment	-0.080 (0.045)	-0.078 (0.049)	0.080	-0.014 (0.045)	-0.013 (0.052)	0.752
Pre-test	-0.016 (0.010)	-0.014 (0.010)	0.105	-0.029 (0.010)	-0.028 (0.009)	0.003**

4.3 RQ3: Post-feedback Removal Impact

The analysis revealed that almost all the benefits observed in the experimental group were no longer present when the LLM feedback was disabled. The only statistically significant difference we found is that, for challenging tasks, the experimental group exhibited a lower tendency to go beyond the time threshold (β = -0.084, p = 0.028, τ = -0.091).

5. Discussion

Our findings extend previous work on persistence in programming (Flores & Rodrigo, 2020; Pinto et al., 2021) by demonstrating the potential of LLM feedback to reduce wheel spinning. Two key themes emerge from the findings: **the promises of LLM feedback in supporting students**, and the **challenges of sustaining the effects of AI-powered feedback**.

Empirical findings from RQ1 and RQ2 suggest that the addition of LLM feedback can promote learners' persistence and reduce unproductive behaviors. Specifically, we have observed improvements in last submission scores, higher likelihood of solving problems, lower rates of going beyond the time threshold (after adjusting for the time required to generate LLM feedback), and reduced wheel spinning. These findings align with previous studies on intelligent tutoring systems, such as (Maniktala et al., 2020), who found that hints in the format

of unsolicited help can encourage productive persistence. Similarly, our results echo the work of (Marwan et al., 2020) in CS education, where personalized feedback led to greater engagement and increased intention to persist in CS. However, the benefits of LLM feedback for compiler errors appear limited in certain scenarios, particularly for challenging tasks. While the experimental group demonstrated significantly higher last submission scores and a marginally higher likelihood of solving challenging problems, LLM feedback did not significantly reduce the probability of wheel spinning on challenging tasks compared to the control group. This limitation in complex problem-solving scenarios could be attributed to several factors: *(1) Inability to diagnose underlying misconceptions*: In complex tasks, students' difficulties often stem from fundamental misconceptions that are not easily identifiable from code alone (Qian & Lehman, 2017). LLM models, operating solely on submitted code, may struggle to effectively diagnose and address these underlying issues. *(2) Overgeneralization of feedback*: A major issue identified with ChatGPT is its tendency to provide overly general responses (Rahman et al., 2023; Ray, 2023). For highly complex tasks, LLM feedback might lack the specificity needed to overcome challenges, failing to disrupt wheel-spinning. *(3) Cognitive overload*: Challenging tasks already impose a high cognitive load on students. Additional feedback, even if accurate, might exacerbate this cognitive burden, making it difficult for students to effectively process and apply the suggestions. These observations align with previous research (Qureshi, 2023), which found that ChatGPT is effective primarily with simpler data structure problems but struggles with complex queries. This often requires multiple prompts and can potentially increase students' cognitive load. These results highlight the need for a nuanced approach to implementing LLM feedback in CS education. While it shows promise in supporting learners, particularly with simpler tasks, educators and researchers must consider strategies to leverage LLMs in more challenging programming tasks.

When disabled, the previously observed benefits of LLM feedback no longer existed, suggesting that the effects may rely on continuous availability. This contrasts with findings from Pankiewicz and Baker (2023), who reported that students exposed to LLM feedback continued to solve tasks more quickly even after its removal, possibly due to a potential learning effect. In our study, however, students were enrolled in an introductory CS1 course, where many lacked prior programming experience, which may have limited their ability to internalize the feedback. Moreover, LLM feedback targeted only compiler errors and was available for just the first half of each module, a design intended to mitigate over-reliance. Yet this limited exposure, combined with the relatively low frequency of compiler errors (under 10% of submissions; Pankiewicz & Baker, 2024), may have reduced opportunities for meaningful learning transfer. These contrasting findings emphasize the need for further research into the sustainability of the benefits of LLM feedback.

Several implications emerge. First, AI-generated feedback has the potential to support programming tasks by helping students overcome initial barriers to understand compiler error messages and achieve more efficient debugging. However, careful implementation is necessary to avoid over-reliance on AI assistance. Second, future research should explore feedback systems that adapt to the learner's current performance, potentially through hybrid systems that combine LLM feedback with data-driven insights about student's current knowledge. Lastly, there is a need to design feedback systems that explicitly foster the development of transferable problem-solving skills, given that the benefits of LLM feedback were not sustained when removed. This could involve gradually reducing the level of support provided, encouraging them to develop independent problem-solving strategies.

Our study has several limitations. First, the effectiveness of LLM feedback may vary depending on the training, prompts and specific models. Our study used GPT-4, and results may differ with other versions or similar language models. Second, these tasks are optional problems and participants were free to engage as little or as much as they wanted with the problems. Third, existing methods for detecting productive persistence and wheel spinning do not account for quality of code revisions or debugging strategies, which could provide valuable insights and should be considered in future research. Lastly, our quantitative metrics can be complemented by qualitative data, such as interviews or surveys, to better understand how students perceive LLM-generated feedback and to contextualize the modest effect sizes.

6. Conclusion

This study provides empirical evidence on the impact of LLM feedback on programming tasks. LLM feedback for compiler errors has shown to enhance certain aspects of student

performance and persistence, especially in challenging tasks. However, its benefits appear to diminish once the feedback is removed in our study context. The transient nature of the benefits emphasizes the need for careful integration of AI tools in educational settings. It emphasizes the importance of developing strategies that not only leverage immediate LLM assistance but also foster students' independent problem-solving skills. Future research should focus on developing LLM feedback systems that can balance immediate learning support with the promotion of long-term skill development and learning. While LLM feedback presents exciting opportunities for CS education, the ultimate goal is to leverage the strengths of LLM feedback while cultivating transferable debugging and programming skills that persist beyond the immediate context of AI assistance.

References

Adjei, S. A., Baker, R. S., & Bahel, V. (2021). Seven-year longitudinal implications of wheel spinning and productive persistence. *International Conference on Artificial Intelligence in Education*, 16–28.

Beck, J. E., & Gong, Y. (2013). Wheel-spinning: Students who fail to master a skill. *Artificial Intelligence in Education: 16th International Conference, AIED 2013, Memphis, TN, USA, July 913, 2013. Proceedings 16*, 431–440.

Becker, B. A., Denny, P., Finnie-Ansley, J., Luxton-Reilly, A., Prather, J., & Santos, E. A. (2023). Programming is hard-or at least it used to be: Educational opportunities and challenges of ai code generation. *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, 500–506.

Cloninger, C. R., Svrakic, D. M., & Przybeck, T. R. (1993). A psychobiological model of temperament and character. *Archives of General Psychiatry*, *50*(12), 975–990.

Cutts, Q., Cutts, E., Draper, S., O'Donnell, P., & Saffrey, P. (2010). Manipulating mindset to positively influence introductory programming performance. *Proceedings of the 41st ACM Technical Symposium on Computer Science Education*, 431–435.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*(6), 1087.

Elinzano, G. B. O., & Ching, M. R. (2024). Determinants of ChatGPT Adoption in Academe & Other Fields – A Review on Theoretical Perspective. *International Conference on Computers in Education*. https://doi.org/10.58459/icce.2024.4969

Fang, Y., Nye, B., Pavlik, P., Xu, Y. J., Graesser, A., & Hu, X. (2017). Online Learning Persistence and Academic Achievement. *Internatonal Educational Data Mining Society*.

Finnie-Ansley, J., Denny, P., Becker, B. A., Luxton-Reilly, A., & Prather, J. (2022). The robots are coming: Exploring the implications of openai codex on introductory programming. *Proceedings of the 24th Australasian Computing Education Conference*, 10–19.

Flores, R. M., & Rodrigo, M. M. T. (2020). Wheel-spinning models in a novice programming context. *Journal of Educational Computing Research*, *58*(6), 1101–1120.

Gong, Y., & Beck, J. E. (2015). Towards detecting wheel-spinning: Future failure in mastery learning. *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, 67–74.

Hartmann, B., MacDougall, D., Brandt, J., & Klemmer, S. R. (2010). What would other programmers do: Suggesting solutions to error messages. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1019–1028.

Käser, T., Klingler, S., & Gross, M. (2016). When to stop? Towards universal instructional policies. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, 289–298.

Kazemitabaar, M., Ye, R., Wang, X., Henley, A. Z., Denny, P., Craig, M., & Grossman, T. (2024). Codeaid: Evaluating a classroom deployment of an llm-based programming assistant that balances student and educator needs. *Proceedings of the 2024 Chi Conference on Human Factors in Computing Systems*, 1–20.

Kiesler, N., Lohr, D., & Keuning, H. (2023). Exploring the potential of large language models to generate formative programming feedback. *2023 IEEE Frontiers in Education Conference (FIE)*, 1–5.

Kloke, J. D., & McKean, J. W. (2012). Rfit: Rank-based Estimation for Linear Models. *The R Journal*, *4*(2), 57–64. https://doi.org/10.32614/RJ-2012-014

Leinonen, J., Denny, P., MacNeil, S., Sarsa, S., Bernstein, S., Kim, J., Tran, A., & Hellas, A. (2023). Comparing code explanations created by students and large language models. *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1*, 124–130.

Leinonen, J., Hellas, A., Sarsa, S., Reeves, B., Denny, P., Prather, J., & Becker, B. A. (2023). Using large language models to enhance programming error messages. *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, 563–569.

Liffiton, M., Sheese, B. E., Savelka, J., & Denny, P. (2023). Codehelp: Using large language models with guardrails for scalable support in programming classes. *Proceedings of the 23rd Koli Calling International Conference on Computing Education Research*, 1–11.

Maniktala, M., Cody, C., Barnes, T., & Chi, M. (2020). Avoiding help avoidance: Using interface design changes to promote unsolicited hint usage in an intelligent tutor. *International Journal of Artificial Intelligence in Education*, *30*, 637–667.

Marwan, S., Gao, G., Fisk, S., Price, T. W., & Barnes, T. (2020). Adaptive immediate feedback can improve novice programming engagement and intention to persist in computer science. *Proceedings of the 2020 ACM Conference on International Computing Education Research*, 194–203.

McBroom, J., Koprinska, I., & Yacef, K. (2021). A survey of automated programming hint generation: The hints framework. *ACM Computing Surveys (CSUR)*, *54*(8), 1–27.

McCauley, R., Fitzgerald, S., Lewandowski, G., Murphy, L., Simon, B., Thomas, L., & Zander, C. (2008). Debugging: A review of the literature from an educational perspective. *Computer Science Education*, *18*(2), 67–92.

McClendon, C., Neugebauer, R. M., & King, A. (2017). Grit, growth mindset, and deliberate practice in online learning. *Journal of Instructional Research*, *8*, 8–17.

Nienaltowski, M.-H., Pedroni, M., & Meyer, B. (2008). Compiler error messages: What can help novices? *Proceedings of the 39th SIGCSE Technical Symposium on Computer Science Education*, 168–172.

Pankiewicz, M., & Baker, R. S. (2023). Large Language Models (GPT) for automating feedback on programming assignments. *Proceedings of the 31st International Conference on Computers in Education. Asia-Pacific Society for Computers in Education*.

Pankiewicz, M., & Baker, R. S. (2024). Navigating Compiler Errors with AI Assistance-A Study of GPT Hints in an Introductory Programming Course. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1* (pp. 94–100).

Pinto, J. D., Zhang, Y., Paquette, L., & Fan, A. X. (2021). Investigating elements of student persistence in an introductory computer science course. *5th Educational Data Mining in Computer Science Education (CSEDM) Workshop*.

Prather, J., Reeves, B. N., Denny, P., Becker, B. A., Leinonen, J., Luxton-Reilly, A., Powell, G., Finnie-Ansley, J., & Santos, E. A. (2023). "It's weird that it knows what i want": Usability and interactions with copilot for novice programmers. *ACM Transactions on Computer-Human Interaction*, *31*(1), 1–31.

Qian, Y., & Lehman, J. (2017). Students' misconceptions and other difficulties in introductory programming: A literature review. *ACM Transactions on Computing Education (TOCE)*, *18*(1), 1–24.

Qureshi, B. (2023). Exploring the use of chatgpt as a tool for learning and assessment in undergraduate computer science curriculum: Opportunities and challenges. *arXiv Preprint arXiv:2304.11214*.

Rahman, M. M., Terano, H. J., Rahman, M. N., Salamzadeh, A., & Rahaman, M. S. (2023). ChatGPT and academic research: A review and recommendations based on practical examples. *Rahman, M., Terano, HJR, Rahman, N., Salamzadeh, A., Rahaman, S.(2023). ChatGPT and Academic Research: A Review and Recommendations Based on Practical Examples. Journal of Education, Management and Development Studies*, *3*(1), 1–12.

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*.

Samson, B., Luriaga, R., & Ebardo, R. (2024). Do Academic Stress and Risk Propensity Affect Behavioral Intention to Use ChatGPT among University Students? *International Conference on Computers in Education*. https://doi.org/10.58459/icce.2024.4938

Sit, C. Y. E., Yang, Y., Yeung, W. K., & Kong, S. C. (2024). Developing a LLMs-Driven System Based on Human-AI Progressive Code Generation Framework to Assist Mathematics Learning. *International Conference on Computers in Education*. https://doi.org/10.58459/icce.2024.4815

Wang, Y., & Baker, R. (2018). Grit and intention: Why do learners complete MOOCs? *International Review of Research in Open and Distributed Learning*, *19*(3).

Xue, Y., Chen, H., Bai, G. R., Tairas, R., & Huang, Y. (2024). Does ChatGPT Help With Introductory Programming? An Experiment of Students Using ChatGPT in CS1. *Proceedings of the 46th International Conference on Software Engineering: Software Engineering Education and Training*, 331–341.

Too Detailed to Share? Towards Risk-Based Privacy Protection of Fine-Grained Educational Data

Hibiki ITO[a*], Chia-Yu HSU[b] & Hiroaki OGATA[b]
[a]*Graduate School of Informatics, Kyoto University, Japan*
[b]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
*hibiki.itoo@gmail.com

Abstract: Due to the recent increase in the use of digital learning platforms, fine-grained digital trace data has been growing in the education sector. However, despite the potential of such micro-level log data for understanding and personalising individual learning processes, its secondary use is limited due to privacy concerns. A key to advancing data sharing for the secondary use while protecting individual privacy is effective risk assessments. Nevertheless, prior research predominantly focuses on privacy risks of structured tabular data, leaving fine-grained digital trace data underexplored. To fill this gap, we conduct a comprehensive risk analysis of fine-grained educational data using the unicity framework. Employing two real-world datasets reflecting on secondary and higher education settings and two open datasets on self-paced language learning, we demonstrate that fine-grained educational data is highly susceptible to re-identification through timestamps. In addition, we show that the effectiveness of naïve coercing of timestamps depends on the number of students in the dataset and the diversity of educational contexts where the data is collected. Our findings help practitioners to make risk-based decisions to choose appropriate privacy protection strategies.

Keywords: Data sharing, privacy, fine-grained data, unicity

1. Introduction

The last few decades have witnessed the rapid increase in the use of digital platforms in education, leading to growing data about learning and its environment. Particularly, educational activities that were once transient and confined to the involved leaners and educators now leave fine-grained traits of log data that reveal their learning profiles. Educational researchers in learning analytics (LA) and educational data mining (EDM) are urging to take advantage of these micro-level data to understand individual learning processes and personalise learning.

However, despite the potential of increasing fine-grained digital trace data in education, the secondary use of such data by researchers has been limited due to privacy concerns (Fischer et al., 2020), leaving sensitive, but potentially useful data in enclaves (Baker & Hutt, 2025). Although protecting the privacy of learners is quintessential in the use of educational data, data sharing is also required for advancing open science and developing reliable educational technologies (Baker et al., 2024).

To ensure an appropriate trade-off between the preservation of individual privacy and the pursuit of collective societal benefits, careful risk assessments and choosing appropriate privacy protection techniques are of paramount importance (Joksimović et al., 2022). Prior research in the field of LA/EDM has investigated privacy risks associated with educational data such as the re-identification risk (Prasser & Kohlmayer, 2015). To mitigate these risks, privacy-preserving data sharing methods such as k-anonymity (Angiuli et al., 2015), synthetic data generation (Liu et al., 2025) and differential privacy (Gursoy et al., 2017) have been studied in the context of education. However, these studies primarily focus on structured

tabular data, and less attention has been paid to fine-grained, set-valued educational data such as log data from learning platforms and behavioural sensors—the gap we aim to fill. Particularly, the risk assessment of fine-grained log data is missing in the literature, which is essential for advancing the privacy-aware secondary use.

To this end, this paper focuses on the re-identification risk of fine-grained educational data, providing a first comprehensive risk assessment through the unicity framework (de Montjoye et al., 2013) in the realm of education. Informally, the unicity metric estimates how many evet-level records are sufficient to single out individual data subjects in a set-valued dataset. This intuitive and realistic metric of privacy risk arguably offers useful evidence for data custodians to make risk-based decisions on what privacy protection methods to apply for sharing educational data.

2. Related works

2.1 Risk assessment on educational data

Effective risk assessment is key to balancing individual privacy concerns with the societal benefits derived from the secondary use of educational data (Joksimović et al., 2022). An inaccurate estimation of privacy risk—be it an underestimation or overestimation—can lead to the implementation of suboptimal protective strategies. Such misalignment may result in either latent vulnerabilities or an unwarranted loss in data utility. The former undermines stakeholder trust and potentially disrupts educational activities as symbolised by the failure of inBloom (Bulger et al., 2017), while the latter may produce severe outcomes, including harmful intervention results (Fredrikson et al., 2014).

Among various types of information-theoretic privacy risks to consider when choosing appropriate anonymisation techniques, we focus on the re-identification risk of pseudonymous data because such data is practically common and beneficial in education research, and the re-identification risk is an area of active research in the domain of education while leaving fine-grained data underexplored. It should be noted that by information-theoretic risk we mean to pay particular attention to quantifiable re-identification risks such as uniqueness of records in a dataset; otherwise, assessing the feasibility of attackers who have access to anonymised data would necessarily depend on the context.

In the literature, the ARX (Prasser & Kohlmayer, 2015) is perhaps the most used assessment tool of re-identification risk of educational data. Having been developed primarily for health data, it allows for estimating re-identification risk defined through the uniqueness of records within specific populations. For example, Kyritsi et al. (2019) employed the ARX tool to estimate the re-identification risk of aggregated tabular data of LMS logs. The study showed that total number of logs for each learner exhibited the highest re-identification risk with 24.39% records unique in the sample dataset and 3.48% unique in the population, according to ARX's definition of population uniqueness.

Another tool for assessing re-identification risk is the Re-identifier Risk Ready Reckoner (R4) developed at the Commonwealth Scientific and Industrial Research Organisation (CSIRO, 2019). Unlike the algorithm of ARX, it quantifies re-identification risk using Markov models accounting for not only uniqueness but also uniformity, i.e., consistency of individuals throughout the course (Vatsalan et al., 2022).

Nonetheless, both these tools take as input data in tabular form and are not applicable to set-valued data. Although in the secondary use, set-valued educational data is usually aggregated into features like the total number of logs in the above example, aggregating and conducting risk assessment every time when a third-party analyst creates a request for data sharing is burdensome and not a sustainable solution. Therefore, evaluating the re-identification risk of fine-grained data would play a crucial role in allowing for a wider range of analyses in the secondary use of the data, which is missing in the literature of the education domain.

2.2 Unicity framework

Proposed by de Montjoye et al. (2013), *unicity* is a framework to evaluate the re-identification risk based on the uniqueness of individuals in a set-valued dataset. Using large-scale location data, de Montjoye et al. (2013) demonstrated that given four geographical points, on average over 95% of individuals can be uniquely determined. It should be noted that re-identifiability and uniqueness are distinct concepts, and unicity is a metric of uniqueness in the first place, potentially overestimating re-identification risk (Barth-Jones et al., 2015). However, as mentioned before, we focus on the re-identification risk of educational data quantified by unicity as other factors such as how an adversary gains auxiliary information necessarily depend on the context.

There has also been a critique that unicity should be estimated based on the uniqueness in the population as the number of individuals in a dataset potentially impacts its unicity (Sánchez et al., 2016). A rebuttal to this claim by Farzanehfar et al. (2021) demonstrated that unicity remains high for larger, population-level location data. That said, unlike location data, educational data are typically collected, stored and used within a single institution or even a single module by a single teacher. Hence, we argue that assessing the unicity of micro-level data collected within a specific context such as a single institution and a single module (see Section 3.1) is indeed meaningful in the domain of education. Overall, our research question (RQ) is formulated as follows:

RQ: *What is the re-identification risk evaluated by the unicity framework in fine-grained educational data?*

3. Materials and methods

3.1 Data

For the experiments we use two private real-world datasets and two open datasets publicly available online. Table 1 shows the summary of the datasets used in the experiments. All data in our experiments are pseudonymous.

Table 1. *Description of the datasets*

Dataset	# records	# students	Period	Description
BookRoll University	66,259	51	4 m	Reading behaviour in a bachelor-level academic reading module
BookRoll Secondary	6,486,986	752	4 m	Reading behaviour in multiple secondary school classes (mostly mathematics and English)
Duolingo	12,854,226	115,222	2 w	Duolingo vocabulary lessons for multiple learning languages
EdNet-KT4	131,441,538	297,915	1y 3m	English reading and listening exercises

To empirically assess the re-identification risk in sharing fine-grained educational data, we employed log data collected via BookRoll, an e-book system through which learning materials are distributed to students as PDF files (Ogata et al., 2015). As shown in Figure 1, interaction logs (e.g. open/close materials, highlight texts, create handwritten notes) are sent to the learning record store (LRS) in the form of the xAPI standard. Each log includes the timestamp, the actor, the xAPI verb, the context ID indicating the class/module and the device type on which the event was operated.



Figure 1. An example of xAPI log collected by the BookRoll system

Additionally, each xAPI log is associated with a BookRoll-specific operation name, which indicates more granular description than xAPI verbs. Table 2 shows some examples of operation names and corresponding xAPI verbs.

Table 2. *Examples of BookRoll operations and xAPI verbs*

xAPI verb	Operation Name	Function
`read`	`NEXT`	Go forward to the next page
	`PREV`	Go back to the previous page
`noted`	`ADD_MEMO`	Add a note
	`DELETE_MEMO`	Delete a note
`highlighted`	`ADD_MARKER`	Add a marker highlight
	`DELETE_MARKER`	Delete a marker highlight

We use two datasets consisting of BookRoll logs. First, the BookRoll University dataset contains log data of undergraduate students within an academic reading module at a Japanese public university. The context is specific to a single module with medium class size (n=51), and the duration of data collection ranges over a semester, thus being a common data unit for LA/EDM analyses. We also employ this dataset to reflect the scope of a typical primary use of educational data at higher education institutions, as we are interested in sharing such data for the secondary use. Typical analyses on this dataset would include temporal learning processes or collaborative learning within a specific context.

Second, the BookRoll Secondary dataset consists of Japanese secondary school students' log data (n=752) across multiple classes (mostly mathematics and English). The scope of the data is limited to a single school, which is a typical unit of educational data in K-12 settings for primary use. Again, this is because we are interested in the risk of sharing such data that are often not released due to privacy concern. As the dataset is larger than the previous dataset in terms of volume and represents more contextual diversity, it would be suitable for studying cross-context learning behaviour in a secondary school setting. For fair comparison, we set the period of data collection to a single semester, the same as for the BookRoll University dataset. This serves as a lower bound of more longitudinal data since the uniqueness of individual trajectories generally increases as data becomes more longitudinal.

To further generalise the results on the previous two private datasets that represent more formal curriculum-based education for the young, we conduct additional experiments with two public datasets that reflect on more self-paced life-long learning. The Duolingo dataset (Settles, 2017) is open data for replication of a study on second language learning by Settles and Meeder (2016). It contains Duolingo vocabulary lesson results for each learner, where each record includes the timestamp, the lexeme (i.e. the target word) and the learning language. Although the dataset spans only a two-week period, it reflects more diverse self-paced learning behaviours, as learners engage with Duolingo lessons independently—unlike

the BookRoll data, where learner activity is shaped by regular, scheduled classes in formal educational settings.

Lastly, the EdNet-KT4 (call EdNet for short) dataset is a large-scale open dataset consisting of exercise results on Santa, a multi-platform self-study app for preparation of the TOEIC (Test of English for International Communication) test (Choi et al., 2020). The data provides more longitudinal learning processes of a larger number of learners, complementing the Duolingo dataset. The results derived from these public datasets also exemplifies a common form of educational data that are gathered and stored by educational technology companies, yet such data often remain inaccessible to researchers. These findings could serve as evidence to encourage companies to adopt effective privacy protection methods for data sharing, which is essential for advancing the broader use of educational data (Fischer et al., 2020).

3.2 Methods

We apply the unicity framework (de Montjoye et al., 2013) to the four datasets[1]. Algorithm 1 shows the process of calculating the unicity, given a dataset and a set of quasi-identifiers (QIs). Note that in each dataset, every *student* contributes multiple *events*. We estimate the unicity by taking the average of the outputs over ten random seeds. For the BookRoll University and Secondary datasets, we include the entire datasets in each sample (i.e. the sample size *m* equals the number of students in each dataset). For the Duolingo and EdNet datasets, we set *m*=2500.

ALGORITHM 1: CALCULATE UNICITY

Input: $\mathfrak{D}$ (dataset), Q (set of QIs), ε (number of events available to the attacker), m (sample size)
Output: *Unicity*

$D \leftarrow$ subset $\mathfrak{D}$ by quasi-identifiers in Q // only quasi-identifiers are used
$S \leftarrow$ randomly choose m students from D
Unicity $\leftarrow 0$
for t in S do // go through every student in the sample
  Trajectory $\leftarrow$ all events of t in D
  Observations $\leftarrow$ randomly choose ε events in *Trajectory*
  Candidates $\leftarrow$ choose from D students whose events include *Observations*
  if $|Candidates| = 1$ do // check if the target is unique
    Unicity $\leftarrow$ *Unicity* $+1$
  end if
end for
return *Unicity* $/ m$

In our experiments, we focus on timestamps as a quasi-identifier, as they are common in set-valued log data and also play an important role in temporal analyses of learning processes (Knight et al., 2017), thereby being a key to balance privacy and utility. One may consider a re-identification attack in the example scenario as illustrated in Figure 2. Having access to pseudonymous log data, an adversarial third-party analyst can search for auxiliary information e.g. on online social media and link the information to record in the dataset, re-identifying the target student. Here, auxiliary information does not have to be public, but can be, for instance, physical observations or inferred from other information, and we assume that the attacker knows that the target student is in the dataset. In this case, the private information regarding the target student's learning behaviour is inadvertently disclosed to a third party without detection. In addition, all student information associated with the same ID, including possibly sensitive information such as final marks of every module, would be revealed to the third-party analyst. The unicity framework provides a means of evaluating the average-case likelihood of this type of privacy violation occurring.

[1] Code available at https://github.com/hibiki-i/too-detailed-to-share.

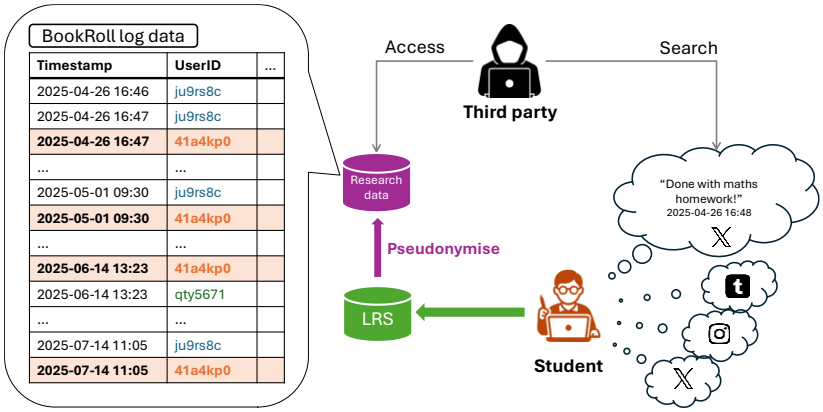


Figure 2. Example re-identification attack by the uniqueness of log trajectory

4. Results

4.1 Timestamps are personally identifiable information (PII)

Figure 3 shows the estimated unicity for different datasets with 95% confidence intervals (CI) computed by bootstrapping (the subsequent plots also show CIs in the same way). Here, ε represents the number of observations available to the attacker and the timestamps are the only quasi-identifier. That is, an attacker only knows whether a target student is active on the learning platform at certain time points (i.e. one-minute time windows). For example, the unicity of the BookRoll Secondary dataset for ε=4 is 0.913, meaning that given four time points an attacker can determine on average 91.3% of the students in the dataset. For smaller ε, the BookRoll Secondary exhibits larger unicity than the BookRoll University, while they converge to the high unicity as ε increases. This is perhaps because the BookRoll University dataset reflects less diversity by focusing on a specific context, where students regularly attend lectures at the same time, reducing the uniqueness of individual trajectories. Nonetheless, both BookRoll datasets exhibit high unicity with a few observations available to the attacker, implying that pseudonymous log data of learning behaviour is highly susceptible to re-identification.

In addition, despite learners' potential behavioural diversity due to the nature of self-paced learning, the Duolingo and EdNet datasets exhibit lower unicity compared to the BookRoll datasets, converging below 0.6 and 0.3, respectively. This is probably due to the larger numbers of individuals in these datasets. Nonetheless, these values must be interpreted with caution, as the unicity metric only evaluates the *average-case* re-identification risk. In other words, the unicity framework quantifies the risk averaged over all students, without capturing the variability in individual vulnerability—some students face higher re-identification risks than others. When protecting individual privacy, we are typically interested in the *worst-case* vulnerability—sometimes referred to as the prosecutor scenario. From this perspective, even though the unicity of the Duolingo and EdNet datasets are relatively small, sharing these datasets would still require strong security and privacy protection measures.

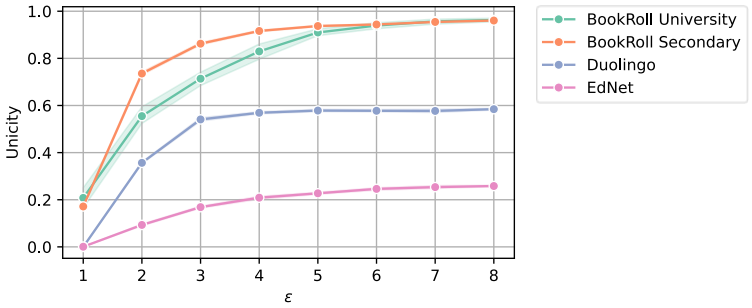


Figure 3. The unicity of different datasets with one-minute timestamps

As the literature suggests that the dataset size impacts its unicity (Barth-Jones et al., 2015; Farzanehfar et al., 2021), Figure 4 plots the unicity of each dataset for ε=4, for which we grouped the Duolingo dataset by learning languages and the unicity is separately estimated for each of them. Here, and in the subsequent plots as well, we fix ε=4, because Figure 3 tells us that unicity by and large converges at this point for all the datasets. Though it can be observed that unicity tends to decrease with the number of individuals in a dataset with a mostly convex curve (Farzanehfar et al., 2021), the trend cannot be generalised over different contexts. For example, while the BookRoll Secondary dataset consists of more students than the BookRoll University, the former shows higher unicity perhaps due to the cross-contextual nature of log data, promoting the uniqueness of each student's learning behaviour. Additionally, the EdNet dataset entails exceptionally low unicity for its number of students included. This gives us a practical implication that dataset size does not solely determine the vulnerability.

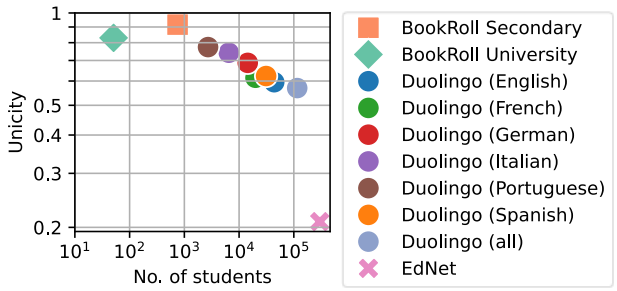


Figure 4. The unicity of different datasets for ε=4

4.2 Generalisation of timestamps

As fine-grained timestamps can be strong PII, we apply naïve coercing: the unit of timestamps are generalised to longer time windows such as quarters, hours and dates (see Figure 5). Here we assume that an attacker only has rough observation that a target student is active on the learning platform within a time window and that a time window is attributed as active if there are one or more logs within that window in the dataset. For example, if there is only one student whose (possibly multiple) BookRoll logs fall in a certain time window, the student can be re-identified by observing that the student uses BookRoll within that time window.

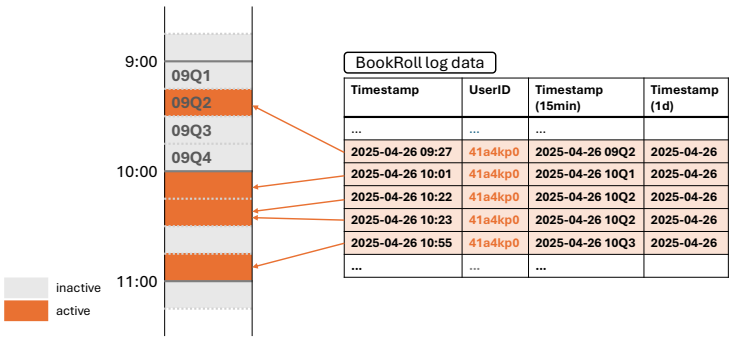


Figure 5. Generalisation of timestamps

Figure 6 illustrates the changes in unicity of each dataset when generalised timestamps with different levels (one-minute, quarter and date) are used as a quasi-identifier. Overall, the effectiveness of naïve coercing depends on the nature of a target dataset. For the BookRoll University dataset, several students are not protected from re-identification with a few timestamps even if the timestamps are generalised to dates. This is probably because students do not engage with the learning materials every day, increasing the uniqueness of

dates when each student is active on BookRoll. On the other hand, for the other three datasets, the generalisation of timestamps to dates protects almost all learners from re-identifying by at least up to eight dates, implying that this naïve coercing effectively mitigates the re-identification risk (i.e. unicity near zero).

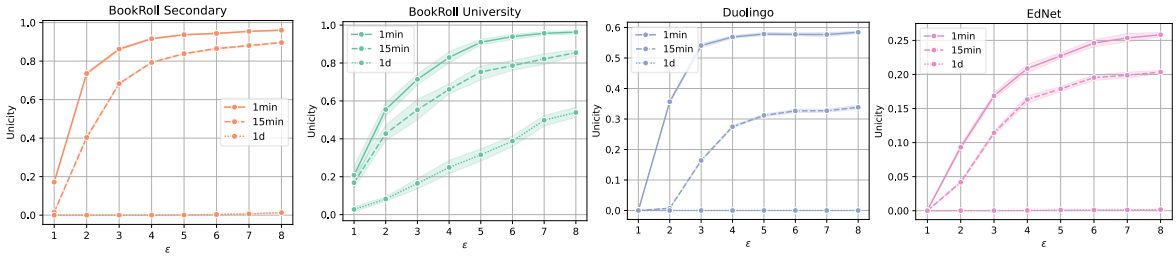


Figure 6. The unicity of different datasets with timestamp generalisation

4.3 Combining auxiliary information

While timestamps entail high risk of re-identification as a quasi-identifier, the risk can be even higher when an attacker have more auxiliary information that can be linked to the log data in question. Figure 7 illustrates how additional quasi-identifiers impact the unicity of each dataset. For the BookRoll datasets, timestamps combined with xAPI verbs would more easily determine unique students. The effect of adding the xAPI verbs to the set of quasi-identifiers is greater than adding device and context information (e.g. 10th grade maths). Moreover, adding the operation, more granular description of a log event than the verb, increases the unicity of the BookRoll University to 0.982 for ε=4. That is, almost all students are uniquely identified given four data points with timestamps and operation names.

The unicity of the Duolingo and EdNet datasets also increases with auxiliary information. Especially, the effect of identifying which target word a learner is engaged at the timestamp is remarkable, raising the unicity over 0.8. Furthermore, the information of platform (either mobile or web) and actions (like xAPI verbs) of EdNet records contributes to the elevated re-identification risk, albeit to a comparatively small extent.

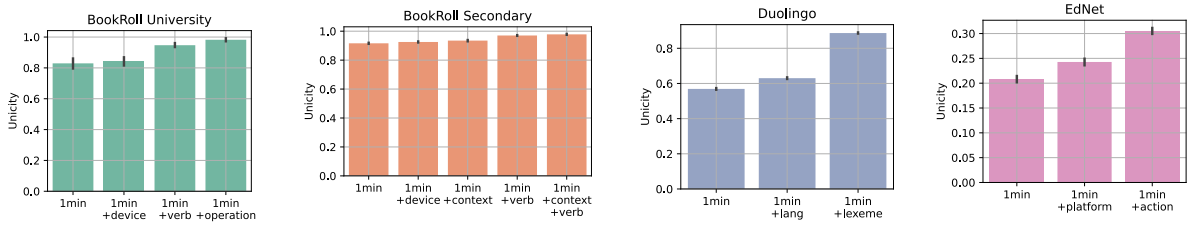


Figure 7. Unicity for ε=4 with different sets of quasi-identifiers

5. Discussion

5.1 Implications

Employing the unicity framework, we investigated the re-identification risk of micro-level educational data. Our contributions are twofold: First, the results demonstrate that educational digital trace data of relatively small size (up to several hundred individuals) are highly vulnerable to re-identification using timestamps as a quasi-identifier while larger datasets also remain unsafe to share by mere pseudonymisation. Second, naïve coercing is potentially effective for cross-contextual or large-scale data but remains ineffective for small-scale, context-specific data.

In practice, our findings help data custodians to make risk-based decisions on choosing proper security and privacy-protection measures. Data administrators and analysts

must recognise the high re-identification risk associated with pseudonymous fine-grained educational data—underscored by the unicity results in this study—as mere pseudonymisation is very common in the secondary use of educational data (Baker et al., 2024). Additionally, our findings suggest that appropriate privacy protection measures depend on the dataset size and whether the data reflects multiple educational contexts. Specifically, for both large-scale fine-grained educational data and smaller datasets covering multiple contexts, generalising timestamps can be an effective privacy-preserving strategy, provided that timestamps are the only quasi-identifier. However, our results illustrate that smaller datasets remain vulnerable despite the generalisation of timestamps, and that other information about each log such as xAPI verbs increases the vulnerability to re-identification, thereby requiring stronger privacy protection techniques like noise addition and synthetic data generation.

5.2 Limitations

A technical remark on our results is that the unicity for these datasets should be seen as lower bounds for re-identification risk. As attacker's observations are chosen from each individual's log trajectory uniformly at random (see Algorithm 1), this unicity estimation potentially biases attacker's observations toward *popular* time windows when many people are active on the learning platform, thereby possibly underestimating unicity (Achara et al., 2015). Nonetheless, our findings demonstrate that even the lower bound for unicity is high, supporting the claim that fine-grained educational data carries high re-identification risk.

Another limitation inherent to the unicity framework is that it only accounts for the uniqueness of individual trajectories. It might be possible to re-identify individuals by, for example, analysing correlations or patterns of log trajectories. Thus, in practice, unicity should not be the sole information to make risk-based decision for sharing educational data in privacy-aware manner. Nonetheless, since uniqueness is a major factor enabling direct re-identification, the assessment of unicity plays a pivotal role in the risk analysis of fine-grained educational data.

Finally, although our datasets reflect typical digital trace data in education, the generalisability of our findings should be tested in future research with various educational contexts. Particularly, data curators are encouraged to apply the unicity framework to assess the re-identification risk by themselves, when sharing pseudonymous log data with third parties.

6. Conclusion

Overall, this paper provides a first comprehensive analysis of the re-identification risk of fine-grained educational data through the unicity framework. The findings demonstrate that, despite pseudonymisation, educational log data is highly susceptible to re-identification through timestamps, and that the effectiveness of timestamp generalisation as a privacy protection strategy depends on dataset size and contextual diversity. Acknowledging the limitations of the unicity framework, our work contributes to inform the community about the re-identification risk of fine-grained educational data, encouraging data custodians to make risk-based decisions to share digital-trace data in privacy-preserving manner.

Acknowledgements

This work was partly supported by Council for Science, 3rd SIP JPJ012347 and JSPS KAKENHI Grant Number 23H00505, 25KJ1515.

References

Achara, J. P., Acs, G., & Castelluccia, C. (2015). On the Unicity of Smartphone Applications. *Proceedings of the 14th ACM Workshop on Privacy in the Electronic Society*. CCS'15: The 22nd ACM Conference on Computer and Communications Security, Denver Colorado USA. https://doi.org/10.1145/2808138.2808146

Angiuli, O., Blitzstein, J., & Waldo, J. (2015). How to de-identify your data. *Communications of the ACM*, *58*(12), 48–55.

Baker, R. S., & Hutt, S. (2025). MORF: A Post-Mortem. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, (pp. 797–802).

Baker, R. S., Hutt, S., Brooks, C. A., Srivastava, N., & Mills, C. (2024). Open science and Educational Data Mining: Which practices matter most? In *Proceedings of the 17th International Conference on Educational Data Mining* (pp. 279–287). International Educational Data Mining Society. https://doi.org/10.5281/ZENODO.12729816

Barth-Jones, D., El Emam, K., Bambauer, J., Cavoukian, A., & Malin, B. (2015). [Review of *Assessing data intrusion threats*]. *Science (New York, N.Y.)*, *348*(6231), 194–195.

Bulger, M., McCormick, P., & Pitcan, M. (2017). *The Legacy of inBloom*. https://datasociety.net/wp-content/uploads/2017/02/InBloom_feb_2017.pdf

Choi, Y., Lee, Y., Shin, D., Cho, J., Park, S., Lee, S., Baek, J., Bae, C., Kim, B., & Heo, J. (2020). EdNet: A large-scale hierarchical dataset in education. In *Lecture Notes in Computer Science* (pp. 69–73). Springer International Publishing.

CSIRO. (2019, April 4). *Re-Identification Risk Quantification*. Privacy Technology Research Group. https://research.csiro.au/isp/research/privacy/r4/

de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports*, *3*(1), 1376.

Farzanehfar, A., Houssiau, F., & de Montjoye, Y.-A. (2021). The risk of re-identification remains high even in country-scale location datasets. *Patterns (New York, N.Y.)*, *2*(3), 100204.

Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., Slater, S., Baker, R., & Warschauer, M. (2020). Mining big data in education: Affordances and challenges. *Review of Research in Education*, *44*(1), 130–160.

Fredrikson, M., Lantz, E., Jha, S., Lin, S. M., Page, D., & Ristenpart, T. (2014). Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. *USENIX Security Symposium*, *2014*, 17–32.

Gursoy, M. E., Inan, A., Nergiz, M. E., & Saygin, Y. (2017). Privacy-preserving learning analytics: Challenges and techniques. *IEEE Transactions on Learning Technologies*, *10*(1), 68–81.

Joksimović, S., Marshall, R., Rakotoarivelo, T., Ladjal, D., Zhan, C., & Pardo, A. (2022). Privacy-driven learning analytics. In *Manage Your Own Learning Analytics* (pp. 1–22). Springer International Publishing.

Knight, S., Friend Wise, A., & Chen, B. (2017). Time for change: Why learning analytics needs temporal analysis. *Journal of Learning Analytics*, *4*(3), 7–17.

Kyritsi, K. H., Zorkadis, V., Stavropoulos, E. C., & Verykios, V. S. (2019). The Pursuit of Patterns in Educational Data Mining as a Threat to Student Privacy. *Journal of Interactive Media in Education*, *2019*(1), 2.

Liu, Q., Shakya, R., Jovanovic, J., Khalil, M., & de la Hoz-Ruiz, J. (2025). Ensuring privacy through synthetic data generation in education. *British Journal of Educational Technology: Journal of the Council for Educational Technology*. https://doi.org/10.1111/bjet.13576

Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K., & Yamada, M. (2015). E-book-based learning analytics in university education. *IEEE International Conference on Consumer Electronics*, 401–406.

Prasser, F., & Kohlmayer, F. (2015). Putting statistical disclosure control into practice: The ARX data anonymization tool. In *Medical Data Privacy Handbook* (pp. 111–148). Springer International Publishing.

Sánchez, D., Martínez, S., & Domingo-Ferrer, J. (2016). Review of *Comment on 'Unique in the shopping mall: On the reidentifiability of credit card metadata'*. *Science (New York, N.Y.)*, *351*(6279), 1274.

Settles, B. (2017). *Replication data for: A trainable spaced repetition model for language learning* [Dataset]. Harvard Dataverse. https://doi.org/10.7910/DVN/N8XJME

Settles, B., & Meeder, B. (2016). A Trainable Spaced Repetition Model for Language Learning. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1848–1858.

Vatsalan, D., Rakotoarivelo, T., Bhaskar, R., Tyler, P., & Ladjal, D. (2022). Privacy risk quantification in education data using Markov model. *British Journal of Educational Technology: Journal of the Council for Educational Technology*, *53*(4), 804–821.

Logged In and Mindset On: Log Data Based Insights on Engagement and Learnings in Programming CBLEs

Indrayani NISHANE*, Ramkumar RAJENDRAN & Sridhar IYER
Centre for Educational Technology, Indian Institute of Technology Bombay, India
*indrayani.nishane@iitb.ac.in

Abstract: Mindset has been shown to influence learners' motivation, persistence, and academic outcomes. While extensive research supports the benefits of a growth mindset in traditional classroom settings, little is known about how these beliefs shape learning in autonomous digital environments. This study explores the relationship between students' mindsets (growth vs. fixed) and their performance and engagement in a computer-based learning environment (CBLE) designed for Python programming, called *PyGuru*. Sixty-seven first-year undergraduates participated in a two-day workshop involving mindset surveys, pre- and post-tests, and three hours of CBLE interaction. Learners were grouped into fixed and growth mindset categories using a tertile split. Learning gains were computed via Hake's formula; behavioural data from log traces were analysed using Mann-Whitney U tests with Benjamini-Hochberg correction. Growth mindset students demonstrated significantly higher learning gains despite similar baseline scores. They also engaged more selectively - performing fewer quizzes, readings, and assessments. Whereas fixed mindset learners showed more frequent error-checking. Time-on-task did not differ significantly, indicating that engagement quality, not duration, distinguished the groups. These results highlight that mindset influences not only outcomes but also learning behaviours in CBLEs, even without teacher support. Findings support the integration of mindset-sensitive features into digital learning tools and call for future research across larger, more diverse samples to inform adaptive, growth-oriented learning design.

Keywords: Learner's Mindset, Mindset in Programming Education, Learning Gain, Computer-based Learning Environment

1. Introduction

Mindset - the belief about whether abilities are fixed or malleable, plays a crucial role in learners' motivation, persistence, and academic success (Dweck, 2006). Students with a growth mindset, who see intelligence as developable through effort, tend to embrace challenges and persist through setbacks, consistently outperforming their fixed mindset peers (Blackwell et al., 2007; Yeager & Dweck, 2012). Growth mindset interventions have improved outcomes across educational levels and demographics (Paunesku et al., 2015; Claro et al., 2016; Kattoum & Baillie, 2025), while fixed mindsets are linked to avoidance behaviors and superficial engagement (Dweck & Leggett, 1988; Lopez & Nguyen, 2024).

In programming, a domain requiring iterative problem-solving, debugging, and abstraction, mindset significantly influences learning behaviors. Growth-minded learners persist through logic errors, engage in strategic exploration, and welcome feedback. In contrast, fixed mindset learners often avoid complex tasks, rely on repetition, and disengage when challenged (O'Dell, 2017; Morales-Navarro et al., 2024). These tendencies affect how students approach key programming tasks such as decomposition and code refactoring (Brewer, 2018; Payne et al., 2018).

While most mindset research has occurred in traditional classrooms, where teacher presence influences motivation, less is known about how mindset operates in self-directed settings. Computer-Based Learning Environments (CBLEs) require learners to self-regulate without real-time instructional support. They also offer unique research opportunities through

automated log data that captures granular, moment-to-moment behaviors (Singh & Rajendran, 2022; Nishane et al., 2021). However, few studies have leveraged this potential, relying instead on self-reports or observations that may miss subtle behavioral differences (Zhang et al., 2024; Payne et al., 2018).

This study addresses these gaps by investigating how growth and fixed mindsets influence learning gains and interaction patterns in *PyGuru*, a CBLE for Python programming. Sixty-seven first-year undergraduates participated in a two-day workshop involving mindset surveys, pre/post-tests, and interaction with the platform. Using Hake's formula for learning gain and Mann-Whitney U tests for behavior analysis (with Benjamini-Hochberg correction), we compared action frequencies and time-on-task between mindset groups. Growth mindset students achieved significantly higher learning gains ($r = 0.52$) and engaged more selectively, performing fewer resource-intensive actions. Fixed mindset learners showed more frequent error-checking and surface-level repetition. Time spent on tasks did not differ significantly, highlighting that engagement quality - not duration - distinguishes learner outcomes. These findings underscore the need for designing CBLEs that recognize and support mindset-driven learning behaviors through targeted scaffolds and reflective strategies.

2. Literature Review

Mindset, defined as beliefs about the malleability of intelligence, is a powerful predictor of motivation, resilience, and academic success (Dweck, 2006). A fixed mindset assumes intelligence is static, leading individuals to avoid challenges and give up when faced with setbacks. In contrast, a growth mindset promotes the idea that abilities can be developed through effort, strategy, and feedback, encouraging persistence and adaptive learning (Dweck, 2006; Dweck & Leggett, 1988). Empirical studies have confirmed these distinctions. For example, Blackwell, Trzesniewski, and Dweck (2007) demonstrated that middle school students taught about growth mindset achieved higher mathematics grades and motivation. Similarly, Paunesku et al. (2015) showed that brief online interventions significantly improved performance, especially for at-risk students. Yeager and Dweck (2012) further highlighted that growth mindset fosters resilience and academic gains across educational contexts.

Programming requires a range of complex problem-solving skills that closely align with learners' underlying mindsets. Students with a growth mindset tend to perceive challenges as opportunities to improve, while those with a fixed mindset often view difficulty as indicative of limited ability (Dweck, 2006; Dweck & Leggett, 1988). This distinction is evident across various programming tasks. For instance, in debugging, growth-minded learners persist through errors, refine their strategies, and treat bugs as opportunities for insight, whereas fixed-minded peers often disengage after repeated failures (O'Dell, 2017; Morales-Navarro et al., 2024). Similarly, in problem decomposition and algorithm design, growth-oriented students are more willing to tackle complexity and experiment with alternatives, while fixed-mindset learners may avoid such tasks due to fear of failure (Yeager & Dweck, 2012; Paunesku et al., 2015). This pattern extends to logic errors and code refactoring, where growth-minded students embrace iteration and continuous improvement, whereas fixed-mindset individuals often perceive the need to revise code as evidence of inadequacy (Brewer, 2018; O'Dell, 2017). When learning new concepts, growth-minded learners actively explore diverse strategies, while fixed-minded students often shy away from unfamiliar material (Blackwell et al., 2007). These differences also shape collaborative programming behaviours: the former group typically welcomes feedback, while the latter may become defensive or disengaged (Morales-Navarro et al., 2024). Even in interpreting error messages or tracing complex code, growth-minded learners show greater persistence, believing that understanding improves with effort, unlike fixed-minded peers who often withdraw early (Payne et al., 2018). Taken together, these alignments suggest that cultivating a growth mindset through strategies such as normalizing failure, encouraging iteration, and fostering peer learning, can significantly enhance students' engagement, resilience, and success in programming. Because learning programming requires the development of all these interrelated skills, novices often struggle to master it. Embedding learning strategies that reflect and reinforce a growth mindset can support students in overcoming these challenges and building long-term proficiency.

Research consistently shows that adopting a growth mindset is associated with higher academic achievement and persistence. Brewer (2018) found that online mindset workshops led to greater perseverance and success in computer-based environments. Similarly, Payne, Babb, and Abdullat (2018) reported better coding performance among growth-minded students. Morales-Navarro et al. (2024) observed that in high school physical computing classes, growth mindset students engaged more actively in debugging and diverse problem-solving strategies. Some more studies reaffirm that a growth mindset positively influences learners' behaviour across educational contexts. In STEM settings, students who perceived instructors as endorsing growth mindsets reported greater comfort and academic success (Kattoum & Baillie, 2025). Growth mindset was also linked to improved learning well-being among high school students, with motivation and grit as mediators (Zhang, Chen, & Li, 2024). College students exposed to mindset interventions reported reduced stress and enhanced academic motivation (Lopez & Nguyen, 2024). Similarly, a large-scale survey showed that growth mindset predicted better mental health and reduced stress impacts among undergraduates (Zhang, Wang, & Zhao, 2022). Together, these studies highlight mindset's role in promoting persistence, emotional resilience, and adaptive learning behaviours. Together, these studies emphasise growth mindset's link to engagement, resilience, and achievement across diverse settings.

Despite these insights, most prior research has been conducted in instructor-led environments, where teachers offer personalised feedback and motivation. In such settings, disentangling mindset effects from teacher influence is difficult. It remains unclear whether growth mindset alone can enhance learning in self-directed, computer-based environments (CBLEs), where teacher support is absent and learners must independently manage their learning processes (Nishane et al., 2023).

Furthermore, existing mindset studies have largely relied on self-reports and observations (Blackwell et al., 2007; Yeager & Dweck, 2012), which fail to capture granular, moment-to-moment behaviours. Log data from CBLEs offers an alternative - recording every click, code submission, error check, and navigation step which can give us detailed record of learner activities (Singh & Rajendran, 2022; Singh et. al., 2022; Nishane et al., 2021). This enables researchers to directly map mindset-related behaviours, such as debugging frequency or exploration breadth, providing a richer understanding of learning processes.

To address these gaps, our study investigates how mindset impacts learning gains and engagement in Python programming within a CBLE. Using detailed clickstream data and learning gain analysis, we explore how growth- and fixed-mindset learners differ in action frequency and time spent on various learning activities. By analysing behavioural traces rather than self-reports, we aim to offer clearer insights into how mindset influences autonomous learning and inform the design of adaptive, mindset-sensitive learning environments.

3. Research Questions

To address the identified gaps, this study investigates how learners' mindsets influence their behaviour and performance in a fully computer-based learning environment (CBLE). Specifically, it examines the relationship between mindset and learning gains in Python programming (RQ1), as well as how mindset shapes interaction patterns, including frequency of clicks and time spent within the CBLE (RQ2).

- RQ1: How much does learners' mindset (growth vs. fixed) affect their learning gains in Python programming when using a CBLE without the presence of a teacher to influence the learning process and learning outcomes?
- RQ2: How does learners' mindset (growth vs. fixed) influence the programming learning in a computer-based learning environment (CBLE) using the log data beyond what can be gathered using self-report and researcher observation data?
    - What is the impact of learners' mindset on the number of actions performed and
    - How does the learners' mindset influence the time spent on those actions in a CBLE?

4. Background : CBLE - PyGuru

We developed PyGuru, a computer-based learning environment aimed at enhancing Python programming education (Singh & Rajendran, 2022; Singh et al., 2022). As depicted in Figure 1, PyGuru comprises four key components.

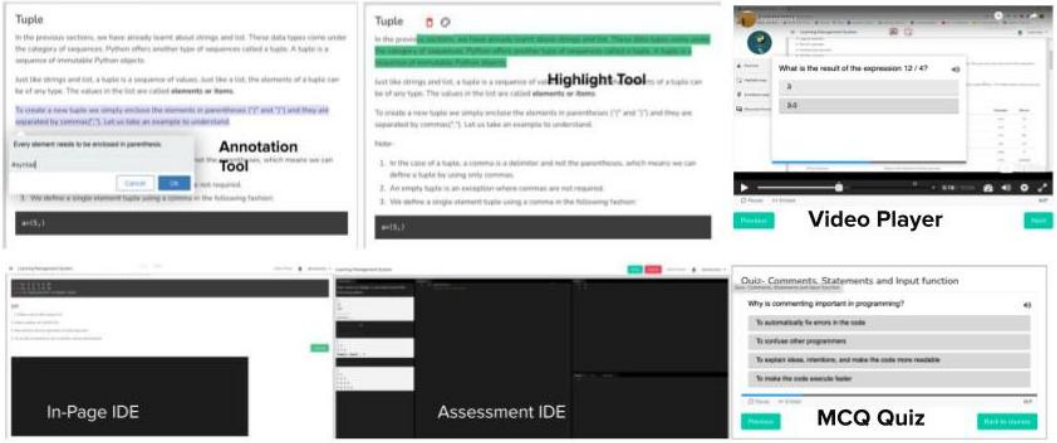


Figure 1. PyGuru environment includes book reader components, such as annotation tool, highlight tool, video player, in-page IDE, assessment IDE, and MCQ quiz.

The Book Reader allows learners to navigate textual content, highlight important sections, and add annotations with personalised notes and tags. The Video Player presents interactive videos embedded with questions that pause playback until answered, promoting active learner engagement. PyGuru also includes two integrated development environments (IDEs): an In-page IDE for immediate code experimentation alongside textual content and an Assessment IDE, which formally evaluates learners' programming skills by running their code against predefined test cases. Additionally, the platform offers an MCQ Assessment Module, providing multiple-choice questions that reinforce conceptual understanding and offer instant feedback. The learners' interactions with the CBLE comprised several key actions. First, they "Read" textual content and "Watch_Video" tutorials to acquire foundational information on each topic. While viewing videos, they paused to answer In-video questions, which served to verify their understanding of the material. After completing a topic, students attempted a "Quiz" covering the same topics to further assess their comprehension. To practice coding concepts, they used the "Practice_Code" environment within the page, entering syntax and experimenting with examples. More formal "Assessment" code tasks required learners to expand or modify pseudocode and "Verify_assessment" code against the test cases.

5. Methodology

5.1 Study Design

The study was conducted over two days in a workshop-style format with first-year non-CS undergraduate students. All participants were from non-CS major background with no prior programming experience using Python. On Day 1, participants completed a demographic survey (age, degree, year of study, perceived programming proficiency, and prior experience), provided consent for log data collection, filled out the 20-item Growth Mindset Scale (Dweck, 1999, 2006), and undertook a 30-minute pre-test. This was followed by 1.5 hours of interaction with the PyGuru environment. On Day 2, learners continued using PyGuru for another 1.5 hours and concluded with a 30-minute post-test. The process is shown in Figure 2.

The pre- and post-tests, adapted from Lewis (2016), assessed Python programming knowledge and were scored out of 25 marks: 13 from multiple-choice questions (1 mark each) and 4 from three open-ended questions (3 marks each), such as code completion and output prediction.

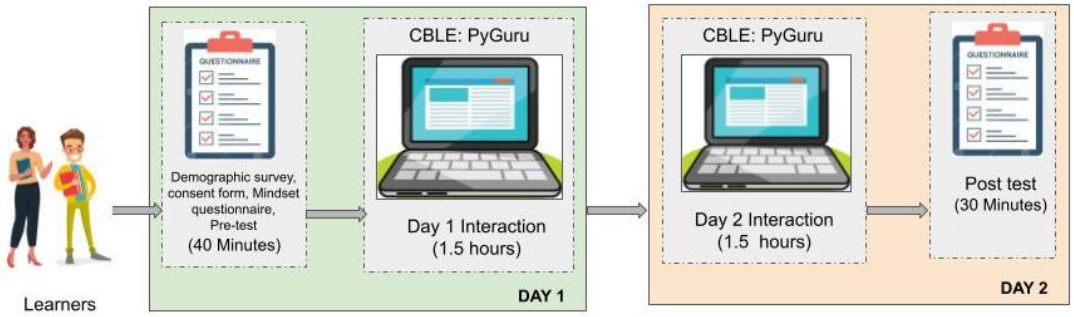


Figure 2. Block diagram of the study design

5.2 Data Collection Procedure

Data were collected from 72 first-year undergraduate students. After excluding five students with missing pre-test, post-test, mindset, or demographic data and the learners with fewer than 10 actions, the final sample comprised 67 participants. The mindset survey demonstrated acceptable internal consistency with Cronbach's alpha value of 0.72. Using a tertile-split approach based on mindset questionnaire scores, students were categorised into three groups: growth mindset (top 33%, n = 24), fixed mindset (bottom 33%, n = 22), and mixed mindset (middle 33%, n = 21).

This grouping enabled us to focus on the extremes of the mindset continuum, a method commonly used in educational research. For example, Ma and Xu (2004) used tertile splits to study math anxiety, and Guthrie et al. (2012) applied a similar approach to examine reading engagement. In this study, only the growth and fixed mindset groups were retained for comparative analysis.

5.3 Data Analysis

Learning gain, a commonly used metric in educational research, was employed to assess the impact of learners' mindsets on their programming performance. It was computed using Hake's formula (Hake, 1998):

$$\textit{Learning Gain} = \frac{(Post{-}test - Pre{-}test)}{(100 - Pre{-}test)}$$

This formula quantifies the proportion of possible improvement a learner achieves, accounting for ceiling effects and enables comparison across students with different starting scores. This metric captured knowledge improvement after interacting with the CBLE, *PyGuru*, and allowed us to compare learning outcomes between growth and fixed mindset groups. We also analysed differences in action frequencies and average time spent on learning activities. Actions with mean frequency below 1 in either group were excluded. Due to non-normal data and unequal group sizes (growth: 24; fixed: 22), we used the non-parametric Mann-Whitney U test for group comparisons. To address multiple comparisons, we applied the Benjamini-Hochberg correction, which controls the False Discovery Rate and is suitable for identifying meaningful behavioural differences in educational datasets.

6. Results

6.1 Learning Gain (RQ1)

To assess the impact of mindset on academic performance, we compared pre-test scores using the Mann-Whitney U test, which revealed no significant difference between growth and fixed mindset groups, $U(23, 22) = 352$, $p = .053$, $r = .29$. The mean pre-test score for growth mindset students was 37.50 (Median = 40), and 30.55 (Median = 24) for fixed mindset

students. Figure 3 shows comparison of the learning gain across the groups of students with fixed and growth mindset. Boxplot shows the differences in distribution of learning gain and highlights the significance in the difference.

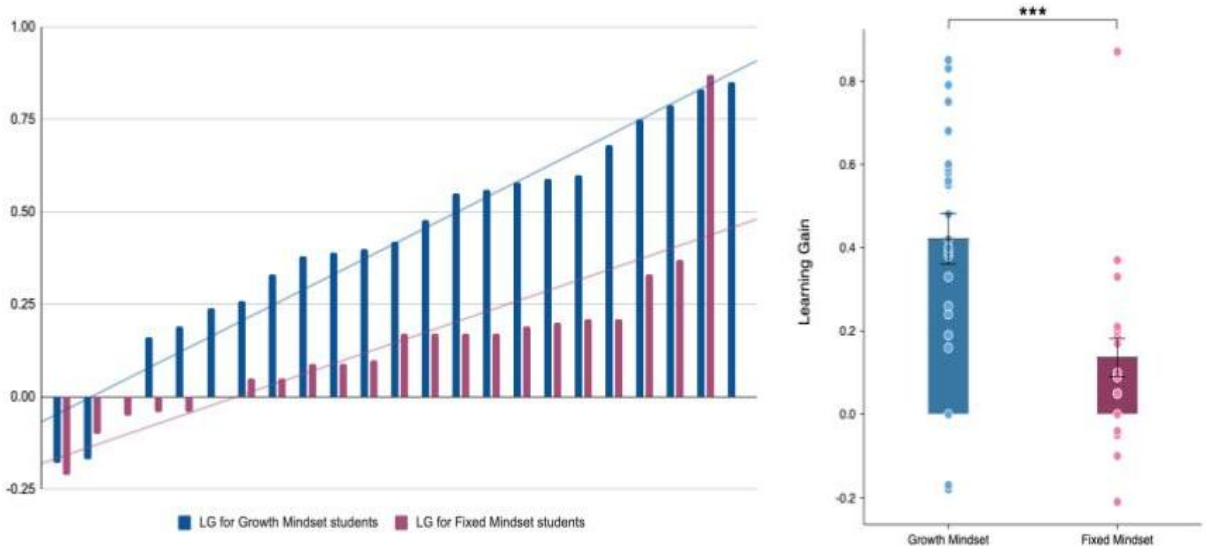


Figure 3. Comparison between the learning gains of students with Growth and Fixed mindset student-wise and group-wise.

Due to unequal group sizes and non-normal distributions, we again used the Mann-Whitney U test. Results showed a significant difference in learning gains, $U = 406.5$, $p = .0005$, $z = 3.49$, with a large effect size ($r = 0.52$, Cohen's $d = 1.11$), favouring growth mindset students. Their mean learning gain was 0.421 (SD = 0.136), compared to 0.291 (SD = 0.214) for fixed mindset students. Growth mindset students achieved significantly higher learning gains compared to fixed mindset students, with a large effect size, highlighting a strong relationship between mindset and academic improvement.

6.2 Differences in the action counts performed by the students with Growth and Fixed mindsets (RQ 2a)

We present the visual comparison of action counts of both the groups in Figure 4. Differences can be seen for the actions such as reading, video watching, quiz, assessment, and practice coding. Apart from the frequency of "Verify_assessment" against the test cases per student, frequency of other actions seem to be different for group of students with growth and fixed mindset. A Mann-Whitney U test was conducted to compare the frequency of various actions between growth and fixed mindset students. The Benjamini-Hochberg procedure was applied to control for multiple comparisons, with an FDR (False Discovery Rate) threshold of 0.05.

A Mann-Whitney U test was conducted to compare the frequencies of various learning actions between growth and fixed mindset students. After applying the Benjamini-Hochberg correction, several actions showed statistically significant differences between the groups as shown in Table 1. Growth mindset students performed significantly fewer quiz ($U = 61.5$, $p < .001$, $r = -0.66$, $Mean_{Growth}=3.38$, $Std_{Growth}= 3.98$, $Mean_{Fixed}= 11.86$, $Std_{Fixed}= 5.96$), video watching ($U = 69.5$, $p < .001$, $r = -0.63$, $Mean_{Growth}= 12.33$, , $Std_{Growth}= 38.96$, $Mean_{Fixed}= 16.05$, $Std_{Fixed}= 8.69$), reading ($U = 95$, $p = .00021$, $r = -0.55$, $Mean_{Growth}= 15.38$, $Std_{Growth}= 14.42$, $Mean_{Fixed}= 36.18$, $Std_{Fixed}= 18.03$), and assessment ($U = 118$, $p = .00119$, $r = -0.47$, $Mean_{Growth}= 2.58$, $Std_{Growth}= 3.31$, $Mean_{Fixed}= 7.27$, $Std_{Fixed}= 5.69$) actions compared to fixed mindset students.

However, no significant differences were found for actions like practice coding, and verifying assessments, as indicated by non-significant corrected p-values. Details are presented in the Table 1.

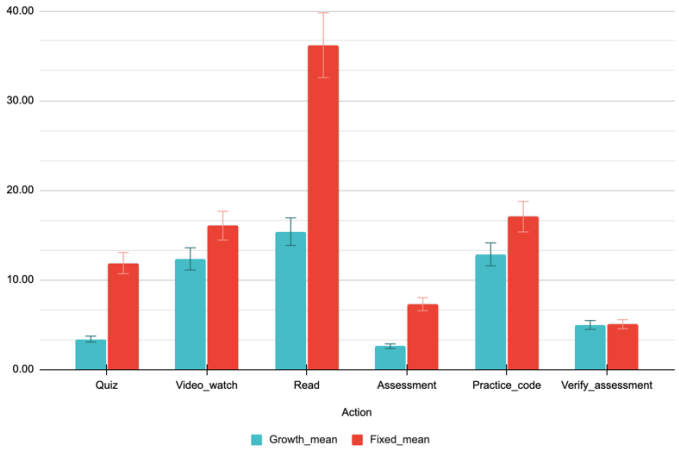


Figure 4. Graph comparing mean value of action counts/frequency for groups of students with Growth mindset and Fixed mindset.

Table 1. *Results of the Mann-Whitney U test along with the Benjamini-Hochberg post-hoc test results for comparison of frequency for each action performed for groups of students with Growth and Fixed mindsets.*

Action	MW_statistic	p_value	z_value	effect_size_r	cohens_d	corrected_p_value	significance
Quiz	61.5	7.55E-06	-4.45	-0.66	-1.68	6.04E-05	TRUE
Video_watch	69.5	1.84E-05	-4.28	-0.63	-0.13	7.37E-05	TRUE
Read	95	0.00021	-3.72	-0.55	-1.27	0.000554	TRUE
Assessment	118	0.00119	-3.21	-0.47	-1.01	0.002383	TRUE
Practice_code	178.5	0.06106	-1.88	-0.28	-0.27	0.081410	FALSE
Verify_assessment	222	0.34807	-0.92	-0.14	-0.01	0.397793	FALSE

Overall, growth mindset students engaged less frequently in certain resource-heavy activities like quizzes, reading, and assessments compared to fixed mindset students.

6.3 Differences in the average time spent on various actions by the students with Growth and Fixed mindsets (RQ2b)

To explore how mindset affects learning behaviour, the Figure 5 below compares the average time spent on key educational actions by students with growth and fixed mindsets. It highlights where their engagement strategies differ across the CBLE - PyGuru. For the actions like – quiz, video watching, reading and assessment, the time spent seems to be different for both the groups.

A Mann-Whitney U test was conducted to compare the average time spent on various actions between growth and fixed mindset students. The Benjamini-Hochberg procedure was applied to control for multiple comparisons, maintaining an FDR (False Discovery Rate) threshold of 0.05.

A Mann-Whitney U test was conducted to compare the average time spent on various learning actions between growth and fixed mindset students as shown in Table 4. After applying the Benjamini-Hochberg correction, no statistically significant differences were observed across any of the actions. Although quiz activities ($U = 173$, $p = .045$, $r = -0.30$, $Mean_{Growth} = 141.5$, $Std_{Growth} = 409.38$, $Mean_{Fixed} = 122.56$, $Std_{Fixed} = 108.56$)) initially showed low p-values, they did not remain significant after correction (corrected $p = .120$). Similarly, no

significant differences were found in time spent on assessments, reading, video watching, or practising code. These results suggest that mindset groups did not differ significantly in the amount of time they allocated to individual learning actions. Detailed results are in Table 2.

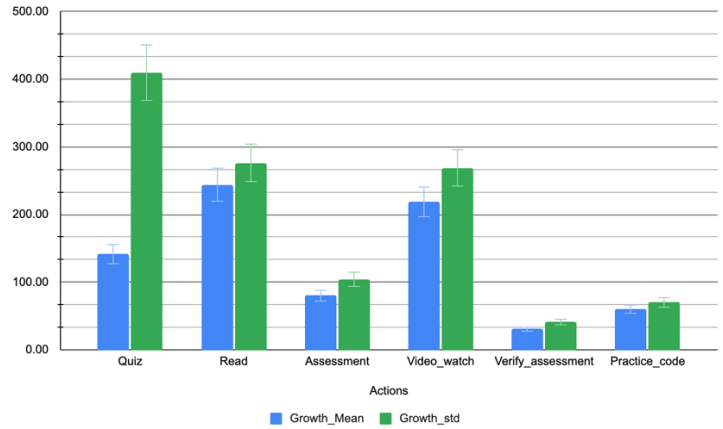


Figure 5. Graph comparing mean value of average time spent on the actions by the groups of students with Fixed and Growth mindsets.

Table 2. *Results of the Mann-Whitney U test for comparison of times spent in seconds for each action performed for groups of students with Growth and Fixed mindsets.*

Action	MW_statistic	p_value	z_value	effect_size_r	cohens_d	corrected_p_value	Significance
Assessment	179	0.0602	-1.8691	-0.2756	-0.3992	0.1204	FALSE
Quiz	173	0.0451	-2.0011	-0.2950	0.0632	0.1204	FALSE
Read	352	0.0543	1.9351	0.2853	0.6560	0.1204	FALSE
Video_watch	200	0.1615	-1.4074	-0.2075	-0.1674	0.2295	FALSE
Verify_assessment	203	0.1721	-1.3414	-0.1978	-0.4246	0.2295	FALSE
Practice_code	218.5	0.3219	-1.0005	-0.1475	-0.1719	0.3678	FALSE

7. Discussion

Our findings show that learners' mindsets significantly influence academic performance in a computer-based learning environment (CBLE) in the absence of teacher/instructor. Growth mindset students achieved notably higher learning gains than their fixed mindset peers, as shown by a statistically significant Mann-Whitney U result (U = 406.5, p = 0.0005, r = 0.52, Cohen's d = 1.11).

This aligns with psychological theories asserting that a growth mindset which believe that intelligence can be developed through effort - enhances motivation, persistence, and performance (Dweck, 2006; Blackwell et al., 2007; Paunesku et al., 2015). Our study extends these insights to programming in CBLEs, where growth mindset learners not only started at similar baselines but progressed further. These findings suggest the value of early mindset assessments to personalise instruction. More broadly, they reinforce the importance of understanding not just *what* students learn, but *how* they engage, particularly in self-paced digital environments.

Beyond learning gains, our results reveal behavioural differences between mindset groups. Growth mindset learners engaged in significantly fewer quizzes, videos, readings, and assessments than fixed mindset learners (all $p < .001$, $r \geq 0.47$), while fixed mindset students

checked errors more frequently (p = .028, r = –0.27). No significant differences were found in coding practice, verification, or submission actions. This suggests growth mindset students interact more selectively and efficiently, reflecting mastery-oriented behaviours (Morales-Navarro et al., 2024; Murphy & Thomas, 2008), while fixed mindset learners may rely on repetition and surface-level engagement (Dweck, 2006). Frequent error-checking among fixed mindset learners may indicate trial-and-error strategies or avoidance of deeper conceptual effort (Yeager & Dweck, 2012). Despite these action differences, time-on-task did not significantly differ after correction. This suggests mindset effects are better reflected in engagement quality than in the amount of time spent supporting earlier findings (Claro et al., 2016).

Together, the data highlight that growth mindset learners optimize their time and actions, while fixed mindset learners show more frequent but potentially less effective behaviours. These distinct engagement strategies underlie the stronger outcomes seen among growth mindset students. This study reinforces the impact of mindset on how learners engage with CBLEs. Growth mindset students demonstrated better learning with fewer but more strategic actions, while fixed mindset learners showed heavier, less productive engagement. These patterns suggest that mindset affects not only outcomes, but also how cognitive resources are allocated. For educators and instructional designers, this underscores the importance of mindset-informed CBLE design. Systems should track not just what learners do, but how and why they engage. Embedding metacognitive prompts and adaptive feedback may help fixed mindset learners adopt more effective strategies.

8. Conclusion

This study investigated how learners' mindsets influence their performance and engagement in a self-regulated computer-based learning environment (CBLE) for Python programming. By combining mindset surveys with log data and learning gain analysis, we demonstrated that students with a growth mindset achieved significantly higher learning gains and exhibited more strategic engagement patterns compared to those with a fixed mindset. While fixed mindset learners engaged in more frequent actions like quizzes and error-checking, growth mindset students demonstrated greater efficiency and selectivity in their interactions by emphasizing quality over quantity.

The findings underscore the importance of integrating mindset-aware instructional strategies into CBLEs. By identifying how learners with different mindsets behave, educators and designers can create adaptive systems that scaffold productive engagement, encourage metacognitive reflection, and support the development of growth-oriented strategies.

Although our results have broader implications for self-directed learning environments beyond programming, their generalizability is shaped by certain constraints. The study focused on a specific CBLE (PyGuru), a limited sample of first-year non-CS undergraduates, and short-term learning sessions. Moreover, the mindset categorization relied on a tertile split, which simplifies the mindset continuum. These assumptions may influence the observed behavioural patterns and learning outcomes.

Future research should also leverage log data analysis more deeply to uncover nuanced patterns in learner behaviour. Specifically, pattern mining techniques can help identify recurring sequences of actions that distinguish growth and fixed mindset learners, offering a richer understanding of how mindset manifests in real-time interactions. This could reveal not only what learners do, but when and how they engage with different components of the CBLE. Such insights can guide the design of targeted interventions that adaptively respond to learners' behavioural patterns and support mindset development through system-level scaffolding. Future work should extend this analysis to larger and more diverse learner populations, explore longitudinal learning effects, and evaluate how similar mindset-behaviour relationships manifest across other domains. Additionally, deeper log analysis and qualitative data (e.g., interviews or think-aloud protocols) could provide richer insight into the mechanisms through which mindset shapes self-regulated learning in digital environments. This can help in development of more robust, mindset-sensitive digital learning environments.

Acknowledgements

This work is supported by the TIH Foundation for IoT and IoE under grant ID RD/0123-TIH0018-003.

References

Dweck, C.S. (2006), Mindset: The New Psychology of Success, Random House, New York, NY

Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality, and development.* Psychology Press.

Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development*, 78(1), 246-263.

Morales-Navarro, S., García, M., & López, J. (2024). Debugging behaviours of students with different mindsets in physical computing courses. International Journal of STEM Education, 11(2), 55–73.

Payne, J., Babb, J., & Abdullat, A. (2018). Enhancing programming education through growth mindset approaches. *Journal of Computer Science Education*, 32(4), 319–340.

O'Dell, D. H. (2017). The Debugging Mindset: Understanding the psychology of learning strategies leads to effective problem-solving skills. *Queue*, *15*(1), 71-90.

Brewer, J. (2018). Teaching growth mindset in online learning environments: A mixed-methods study. *Journal of Educational Psychology*, 110(8), 1122–1139.

Paunesku, D., Walton, G. M., Romero, C., Smith, E. N., Yeager, D. S., & Dweck, C. S. (2015). Mind-Set Interventions Are a Scalable Treatment for Academic Underachievement. *Psychological Science*, 26(6), 784-793.

Claro, S., Paunesku, D., & Dweck, C. S. (2016). Growth mindset tempers the effects of poverty on academic achievement. *Proceedings of the National Academy of Sciences*, 113(31), 8664–8668. https://doi.org/10.1073/pnas.1608207113

Yeager, D. S., & Dweck, C. S. (2012). Mindsets That Promote Resilience: When Students Believe That Personal Characteristics Can Be Developed. *Educational Psychologist*, 47(4), 302–314. https://doi.org/10.1080/00461520.2012.722805

Guthrie, J. T., Wigfield, A., Klauda, S. L., & Barbosa, P. (2012). Influence of concept-oriented reading instruction on reading comprehension and engagement: A cluster-randomized trial. *Journal of Educational Psychology*, *104*(2), 397–418.

Ma, X., & Xu, J. (2004). The causal ordering of mathematics anxiety and mathematics achievement: A longitudinal panel analysis. *Journal of Adolescence*, *27*(2), 165–179.

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review, 95*(2), 256–273.

Kattoum, R. N., & Baillie, M. T. (2025). A more positive mindset context is associated with better student outcomes in STEM, particularly for traditional-age students. *International Journal of STEM Education*, *12*(1), 15.

Hake, R.R.(1998). Interactive-Engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. American Journal of Physics,66(1),64-74.

Nishane, I., Sabanwar, V., Lakshmi, T. G., Singh, D., & Rajendran, R. (2021, July). Learning about learners: Understanding learner behaviours in software conceptual design TELE. In *2021 International Conference on Advanced Learning Technologies (ICALT)* (pp. 297-301). IEEE.

Singh, D., & Rajendran, R. (2022). Investigating learners' Cognitive Engagement in Python Programming using ICAP framework. In *Proceedings of the 15th International Conference on Educational Data Mining* (p. 789).

Nishane, I., Singh, D., Rajendran, R., & Iyer, S.(2023). Does learner mindset matter while learning programming in a computer-based learning environment. In *2023 International Conference on Technology for Education (T4E)*. IEEE.

Singh, D., Subramaniam, H., & Rajendran, R.(2022). PyGuru: a programming environment to facilitate measurement of cognitive engagement. *In International Conference on Computers in Education*.

Zhang, Y., Chen, H., & Li, J. (2024). The impact of a growth mindset on high school students' learning subjective well-being: The mediating role of achievement motivation and grit. *Frontiers in Psychology, 15*, 1399343.

Lopez, C. M., & Nguyen, T. (2024). The effect of mindset interventions on stress and academic motivation in college students. *Innovative Higher Education, 49*(1), 25–42.

Zhang, L., Wang, X., & Zhao, Y. (2022). The influence of growth mindset on the mental health and life events of college students. *Frontiers in Psychology, 13*, 821206.

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

How do students design? A data-driven approach for understanding students' engineering design behaviors

Shruti JAIN*, Namrata SRIVASTAVA, Clayton COHN, Naveeduddin MOHAMMED & Gautam BISWAS
Vanderbilt University, USA
**shruti.jain@vanderbilt.edu*

Abstract: Engineering design is a natural addition to STEM+C education, as it motivates students to integrate science and computing while engaging in problem-solving tasks. However, K-12 students face challenges in applying reasoning and critical thinking skills to find "optimal" design solutions. In this paper, we investigate students' behaviors when they are engaged in engineering design tasks by analyzing the sequences of the design solutions they generate to find an "optimal" solution. We collected data from sixth-grade students (N=88) working in pairs in an NGSS-aligned earth sciences curriculum. Using Markov Models, we visualized students' solution progressions and identified instances where they became stuck in sub-optimal states. Our results show that (1) students' scores improved significantly from pre-test to post-test; (2) students' problem-solving behaviors could be categorized as *design space explorers* and *conservative designers*; and (3) students' problem-solving behaviors influenced the quality of their final designs. Additionally, our findings underscore the importance of supporting design space exploration as a key aspect of scaffolding engineering design activities.

Keywords: STEM+C, Engineering design, NGSS, Learning Analytics

1. Introduction

Engineering design (ED) helps advance STEM+C education, enabling students to combine scientific and computational concepts and practices to solve real-world problems. ED builds 21st-century skills, such as critical thinking, creativity, collaboration, and communication (Ali & Tse, 2023). The National Research Council (2012) and the Next Generation Science Standards (NGSS Lead States, 2013) underscore the importance of integrating ED into K–12 science curricula to foster problem-solving skills, design thinking, and deeper engagement with core scientific practices.

Science-integrated engineering curricula offer a promising approach to helping students learn by applying science concepts to meaningful design tasks (Kelly & Cunningham, 2019). Through computational modeling and simulation-based learning environments, students can apply their disciplinary knowledge in authentic contexts and develop strategies to search the design space and find an optimal design solution that meets specified constraints (Zhang et al., 2019). However, students often find engineering design tasks challenging. They must understand the scientific principles and also navigate complex design spaces where trade-offs, such as cost versus performance, must be carefully considered (Purzer et al., 2015). Prior research shows that students struggle to identify and manage these trade-offs and often rely on trial-and-error approaches instead of systematic reasoning or science-based justifications (Purzer et al., 2015; Zhang et al., 2019; Hutchins et al., 2021). Compounding this challenge, teachers often lack formal training in engineering design, making it difficult for them to support their students (Radolff & Capobianco, 2021). As a result, students often find it difficult to develop and apply systematic principles and critical thinking skills that support generating high-quality or optimal engineering design solutions.

In this paper, we examine the design solutions created by sixth-grade students participating in the Water Runoff Challenge (WRC), a three-week STEM+C curriculum that seamlessly combines earth science, engineering, and computational thinking (Basu et al., 2022). Students investigate the water absorption and runoff characteristics of different surface materials after heavy rainfall. They learn through hands-on experiments and then develop conceptual and computational models that estimate water runoff for various materials and rainfall amounts, using their absorption capacities and the principle of conservation of matter.

The students work together in pairs to extend their computational models to redesign a schoolyard, aiming to reduce water runoff while meeting constraints related to cost, accessibility, and functionality for the play areas. They are encouraged to explore the design solution space and submit their "best" designs that satisfy all specified criteria. By analyzing the sequence of design solutions students generate as they strive for their "best" solutions, our research seeks to understand how students navigate the solution space, their awareness of the design constraints, and how their problem-solving patterns affect the quality of their final designs. Specifically, we analyze data collected from the WRC learning environment to address the following research questions:

RQ1: How effective is the WRC curriculum in helping students develop engineering design knowledge? To answer RQ1, we analyzed students' learning gains derived from their pre- and post-test scores.

RQ2: How do students' search patterns for design solutions vary as they work toward the optimal solution? To address RQ2, we (i) applied an Expectation Maximization Clustering algorithm to characterize their search patterns, and (ii) studied their problem-solving behaviors using Markov models.

RQ3: How are the different problem-solving behaviors related to the quality of the design solution reported by the students? To address RQ3, we developed scoring schemes to compare the final design solutions of students from the two different clusters.

The results demonstrate that the WRC curriculum effectively supports student learning in science, computational thinking, and engineering design. Findings highlight the need for scaffolding design space exploration and trade-off analysis to help students optimize their solutions. Insights from this work can inform curriculum design, and the development of personalized scaffolds to foster effective integration of engineering design into K–12 classrooms.

2. Literature Review

NGSS advocates for integrating engineering practices in K-12 education (NGSS Lead States, 2013); research on students' engineering design (ED) processes primarily focuses on high school students. For instance, Yu et al. (2020) used structural equation modeling to show that the design process mediates the link between scientific knowledge, critical thinking, and design quality. Purzer et al. (2015) studied high school students designing energy-efficient buildings, highlighting challenges in trade-off analysis. Xing et al. (2023) applied learning analytics to identify challenges related to behavioral features in an "energy-plus" model that was used to design a house with minimal energy requirements.

The limited research on how middle schoolers engage with ED includes Bowen et al. (2016), who noted that while prior knowledge affects initial performance in virtual bridge simulations, iterative practice can help fill these knowledge gaps. Similarly, Du et al. (2025) have studied the dynamics between science concepts and design behaviors of different performance groups using Markov Chain analysis. Montgomery et al. (2020) used click-stream data to calculate sub-behaviors– exploration and systematicity– for understanding students' optimization behaviors. Still, these cumulative metrics fail to capture temporal transitions from exploratory to systematic optimization. Prior work by Zhang et al. (2019) used 3-D visualizations of individual students' trajectories to demonstrate how middle school students struggle with navigating trade-offs. While trying to satisfy one constraint, they end up compromising on another. Zhou et al. (2021) observed that high-performing middle school students effectively integrated all verbal, visual, and physical modalities in a toy design workshop. However, such analysis needs to be done individually and is not scalable (e.g., Kelly & Cunningham, 2019;

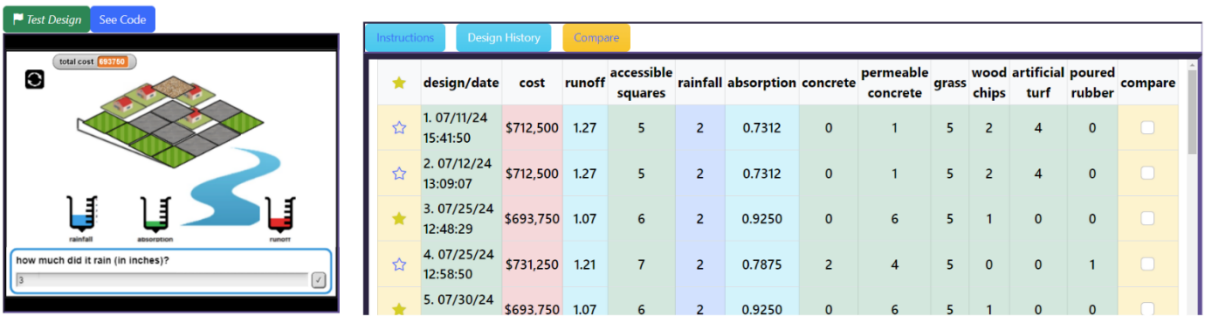
Wendell et al., 2017). We need to look at groups of students as a whole to find generalizable trends in students' problem-solving behaviors. To our knowledge, the process of evolution of students' designs in a simulation-based design environment remains underexplored.

Evaluating and providing feedback on ED solutions is challenging, as the ED activities comprise several objective and subjective design criteria. Goldstein et al. (2016) developed a scoring criterion for ED projects, taking the subjectivity into account by adopting a Trade-off Value approach that compares designs based on their performance on technical, economic, and human factors. However, the suggested approach uses percentile calculation, which can only evaluate an ED solution relative to other students' solutions. On the other hand, the scoring criteria in previous works (Zhang et al., 2019) did not fully consider human factors, such as accessibility, of the ED challenge in the WRC curriculum. Hence, we adopted the two frameworks to develop an objective metric to evaluate the ED solutions while accounting for technical, economic, and human factors. It is important to understand how middle school students' design behaviors evolve over time, and how we can analyze these behaviors on a scale. In this research, we aim to extend our understanding of middle school students' problem-solving behaviors in ED activity by studying the progression of students' design solutions.

3. The WRC Curriculum

The WRC curriculum (Basu et al., 2022) includes 12 lessons over three weeks, with daily 45-minute classes conducted in-person, combining traditional pen-and-paper tasks with activities in a computer-based learning environment. Students start with a pre-test and finish with a post-test to evaluate their learning gains. The curriculum covers earth science concepts such as the conservation of mass, rainfall, absorption, and runoff, helping students understand the challenges of urban runoff. They investigate how different surfaces manage rainwater and engage in an Engineering Design Challenge to design a schoolyard that reduces runoff after a rainfall of 2 inches. Students create designs and develop models to simulate and improve water runoff solutions based on the given criteria on the ED interface, as shown in Figure 1.

Students design a low-runoff schoolyard within a $750,000 budget in an interactive computer environment. The schoolyard is visualized as a 4×4 grid with 12 editable squares selectable from six materials: concrete, permeable concrete, natural grass, artificial turf, wood chips, and poured rubber. Each material varies in cost, absorption, and accessibility. Poured rubber is expensive (cost=$187,500), but it offers high absorption (1.2 inches) and is fully accessible. In contrast, wood chips (cost=$37,500) provide moderate absorption (1.0 inches) but lack accessibility. To solve the design challenge, students need to navigate these trade-offs. As students adjust their designs, the cost updates in real time. They can evaluate their design by clicking the "Test Design" button and entering a rainfall amount to visualize total absorption and runoff values. The Design History table (see Figure 1) tracks all tested designs. The engineering design challenge has 6^{12} potential solutions, and approximately 6^{6} meet the runoff, cost, and block constraints detailed in Table 1.



	design/date	cost	runoff	accessible squares	rainfall	absorption	concrete	permeable concrete	grass	wood chips	artificial turf	poured rubber	compare
☆	1. 07/11/24 15:41:50	$712,500	1.27	5	2	0.7312	0	1	5	2	4	0	
☆	2. 07/12/24 13:09:07	$712,500	1.27	5	2	0.7312	0	1	5	2	4	0	
★	3. 07/25/24 12:48:29	$693,750	1.07	6	2	0.9250	0	6	5	1	0	0	
☆	4. 07/25/24 12:58:50	$731,250	1.21	7	2	0.7875	2	4	5	0	0	1	
★	5. 07/30/24	$693,750	1.07	6	2	0.9250	0	6	5	1	0	0	

Figure 1. WRC Engineering Design (ED) Environment. The design on the left represents the optimal solution containing 6 permeable concrete, 5 natural grass, and 1 woodchips square.

4. Methodology

This section presents the study design, data collection, measures, and analytical methods, including scoring pre- and post-tests to calculate learning gains, qualitative states, and quantitative scores for analyzing students' engineering design solutions, and the statistical and clustering methods used for understanding design behaviors.

4.1 Study Design and Participants

Our study involved 170 sixth-grade students from an urban public school in the southeastern U.S., comprising 63% White, 17% Black/African American, 7% Hispanic, 9% Asian, and 4% other races, with 51% female and 49% male students. The Vanderbilt University Institutional Review Board approved the study with parental and student consent. The two participating science teachers underwent professional development before the study and co-taught the WRC curriculum with researcher support. The curriculum featured individual and collaborative tasks where students worked in pairs. For engineering design activities, pairs used a shared user ID in the web environment and submitted their final designs in the shared notebook. Students could freely create and test as many solutions as they wished during three class periods. The pairs generated 2 to 47 design solutions (*M=19.86, SD=10.99*). We analyzed pre-post tests, submitted designs, and design sequences from the design history logs collected in the WRC environment. After excluding incomplete records, our dataset comprises 88 students' data (44 pairs) and 874 generated engineering design solutions.

4.2 Measures

Learning gains: To answer RQ1, we analyzed students' responses to validated pre- and post-test included in the WRC curriculum. The tests measure students' knowledge in Science, Computational Thinking (CT), and Engineering Design (ED), and the tests have a section on each topic (Basu et al., 2022). Responses were graded by a researcher and an external grader using validated rubrics. After two rounds, the graders achieved an average Cohen's Quadratic Weighted Kappa (QWK) of 0.9, indicating strong inter-rater reliability (McHugh, 2012).

States of Engineering Design (ED) Solutions: To answer RQ2 and assess the quality of students' ED solutions, we organized their designs hierarchically by runoff levels and compliance with the design constraints. Designs were categorized based on runoff generated from two inches of rainfall: those below the median runoff value across all design solutions were classified as Low Runoff, while others were labeled as High Runoff. Next, we assessed the designs by determining their compliance with the design constraints. Designs that met neither the cost criterion nor the blocks criteria were labeled as Nothing Satisfied. Those that met one of the two criteria were identified as Cost Satisfied or Blocks Satisfied. Designs fulfilling both criteria were classified as All Satisfied. This process resulted in 8 categories, but we combined the high runoff designs into one state due to their smaller number.

In summary, we classified the designs into five states: (i) *High Runoff* (above-median runoff); (ii) *NS-Low Runoff* (low runoff, no design constraints met); (iii) *CostS-Low Runoff* (low runoff, only budget criterion met); (iv) *BlockS-Low Runoff* (low runoff, meeting blocks criteria only); and (v) *AllS-Low Runoff*, for designs meeting all constraints, deemed high quality. This categorization provided insights into students' search patterns, including how they navigate the design space, prioritize constraints, and approach design challenges. While these states primarily reflect solution quality in terms of constraint satisfaction and do not directly measure cognitive processes, analyzing transitions between them using Markov Models allows us to infer consistent problem-solving patterns and strategies.

Final Design Solutions Score: To answer RQ3, we developed a scoring scheme to evaluate students' final designs systematically. Following Goldstein et al (2016), we calculated sub-scores for the design challenge's technical (minimizing runoff under two inches of rainfall), economic (keeping redesign costs below $750,000), and human factors (ensuring accessibility, grassy field, play area, and parking).

Figure 1 shows the optimal design for this challenge, achieving the lowest possible runoff of 1.07 inches with a cost of $693,750 while meeting the *accessibility* and *facilities criteria*, which is collectively termed the *blocks criteria*. Our 12-point scoring system assigns up to 5 points for the runoff value, with full points for the optimal runoff and 0 for incomplete solutions. The *cost criterion* contributes up to 3 points: 2 points for staying under budget, plus one for improving cost-efficiency towards the optimal design cost. The *blocks criteria* are scored binarily: 1 point for meeting the *accessibility criterion* (at least six accessible squares) and 1 point each for satisfying the grassy field, play area, and parking/hardcourt requirements. This scoring scheme, detailed in Table 1, was used to assess the final schoolyard designs submitted by the students at the end of the Engineering Design activity. As it was developed post-hoc, students relied on their judgment to evaluate their design solutions during the activity.

Table 1. *Design constraints of the ED task and the proposed scoring scheme.*

Criteria	**Criteria and Score Description**	**Sub-score**
Runoff	Design should minimize runoff for 2 inches of rainfall. Scored on a linear scale (0-5), with the optimum runoff value getting a score of 5, and higher runoff solutions getting lower scores.	5
Cost Criterion	Design must not exceed the budget of $750,000. 2 points for meeting the budget constraint, and a linear scale (continuous, 0-1) for getting closer to the optimal cost.	3
Accessibility Criterion	Design must have 6 squares made of accessible materials (concrete, permeable concrete, artificial turf, or poured rubber). Binary score for meeting the accessibility criterion.	1
Facilities Criteria	**(a) Field**: Design must have 4 squares of grassy field made of grass or artificial turf. **(b) Play area**: The Design must have 1 square for the play area made of wood chips or poured rubber. **(c) Parking and Hardcourt**: Design must have 5 squares for parking and hardcourt made of concrete or permeable concrete. Awarded binary scores for meeting each of the Facilities Criteria.	1 1 1
	Total Score	12

4.3 Data Analysis Methods

4.3.1 Effectiveness of the WRC Curriculum - Statistical Tests

To evaluate the effectiveness of the curriculum, we assessed whether the scores for each section (Science, CT, and ED) on the pre-post tests followed a normal distribution using the Shapiro-Wilk test. Because the data was not normally distributed, we used the non-parametric Wilcoxon signed-rank test to compare student performance on pre- and post-tests.

4.3.2 Identification of Problem-Solving behaviors - FOMM and Clustering analysis

We identified clusters based on problem-solving behaviors using a two-step method involving a First-Order Markov Model (FOMM) and Expectation-Maximization (EM) clustering, following a methodology inspired by prior process mining research (Gatta et al, 2017; Srivastava et al, 2022). First, we classified students' designs into the five states, representing the quality of design solutions (detailed in Section 4.2). We then calculated the sequences of design states for each student pair – the sequences represent the progression of quality of the designs generated iteratively over multiple sessions. Next, using the pMineR package in R (Gatta et al, 2017), we trained a First-Order Markov Model (FOMM) to estimate the transition probabilities between the states, creating transition matrices for each pair's sequence. These transition matrices were input to an EM clustering algorithm to identify distinct problem-solving profiles. The Elbow method suggested an optimal solution of $k=2$ clusters. Model convergence was monitored using the log-likelihood change across iterations, and cluster performance was

evaluated using within-cluster and between-cluster distance metrics derived from the FOMM-based representations. To further interpret the identified clusters, we employed descriptive statistical analysis to summarize key characteristics and differences among the clusters, and FOMM process mining to investigate the transitions between different design states within each cluster. The transition matrices and subsequent clustering reveal consistent patterns in design changes, suggesting intentional problem-solving rather than random modifications.

4.3.3 Influence of Design Behaviors on Final Design Solutions - Statistical Tests

We used descriptive analysis and Mann-Whitney U tests to investigate how different design behaviors influenced the final design solutions that students submitted in their notebooks.

5. Results

5.1 RQ1: Effectiveness of the WRC curriculum

The pre- and post-test analysis for the 88 students showed that the intervention produced significant learning gains ($p<0.001$) for all three domains− Science, CT, and ED. The accumulated average scores (*maxscore=40*) improved from *25.47 (SD=5.95)* to *30.34 (SD=4.64)* points, showing a significant improvement *(W=264.5, r=−0.751, p<0.001)*. Science scores *(maxscore=12)* improved from *7.76 (SD=2.59)* to *8.83 (SD=2.33) (W=599.5, r=−0.603, p=0.001);* CT scores (*maxscore=13*) improved from *8.06 (SD=2.96)* to *10.55 (SD=2.02) (W=268.5, r=−0.749, p<0.001)*; and ED scores *(maxscore=15)* from *9.64 (SD=2.46)* to *10.97 (SD=2.14) (W=612.0, r=−0.597, p<0.001).*

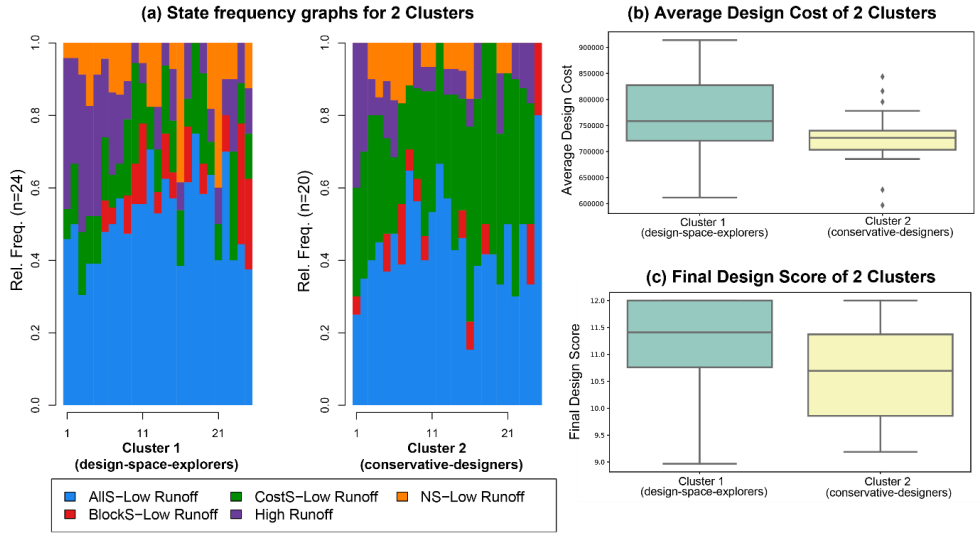


Figure 2. (a) State frequencies over design iterations for the two clusters. Box plots of (b) average cost of generated designs; and (c) Final Design Scores of students in the two clusters.

5.2 RQ2 (i): Identifying Problem-Solving Behaviors

Using the clustering approach (see Section 4.3.2), we identified two distinct clusters of student pairs characterized by their exploration strategies. Figure 2(a) shows state frequencies over design iterations for the two clusters. Notably, pairs in Cluster 2 spent more time in the *CostS-Low Runoff* state, indicating a stronger emphasis on cost optimization. The box plot (Figure 2(b)) also shows that the average cost of designs in Cluster 2 was lower (*M=$726,023.45, SD=55,513.41*) than Cluster 1 (*M=$776,305.04, SD=73,957.01*). This difference was statistically significant (*p=0.02, Mann–Whitney U=339.0, Z=2.333, r=0.352*). Given their cost-focused and cautious strategy, we refer to Cluster 2 as *conservative designers*. In contrast, students

in Cluster 1 explored a wider range of design states, demonstrating a more exploratory approach; we therefore refer to this cluster as *design space explorers*.

5.3 RQ2 (ii): Understanding Students' Engineering Design Process

To analyze the exploration strategies of the two clusters—*conservative designers* (*N=20*) and *design space explorers* (*N=24*)—we constructed First-Order Markov models (FOMM) based on the sequences of their design solutions. Figure 3 presents a comparative FOMM for the two clusters, where states (see Section 4.2) represent the quality of design solutions, while edges indicate transitions between states with corresponding probabilities of transition.

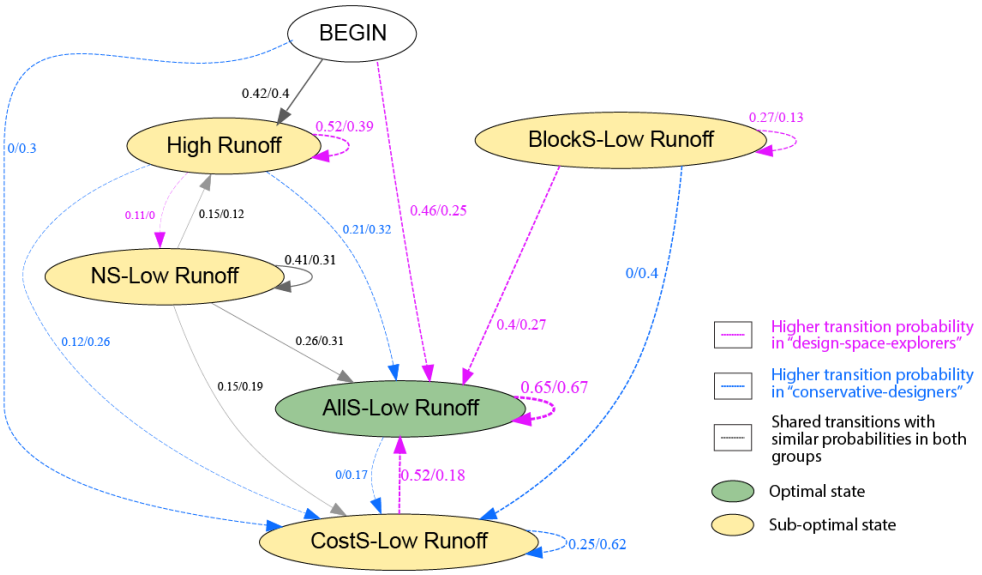


Figure 3. First-order Markov models for "*design space explorers*" and "*conservative designers*". Only transitions with probabilities > 0.1 are shown. Labels show the transition probability for *design space explorers* / *conservative designers*. Colored edges indicate probability differences >0.1 (pink: higher for *design space explorers*; blue: higher for conservative designers).

The *conservative designers* typically began from suboptimal states, such as *High Runoff* (*prob=0.4*), and only a few student pairs achieved the optimal state (*AllS-Low Runoff*) early in their search process (*prob=0.25* of generating the optimal design in the first step). The students initially generated a sequence of *High Runoff* solutions, often not succeeding in improving their solutions (*prob=0.39*) or moving to another sub-optimal state before eventually finding an optimal solution. Interestingly, there are no incoming edges to the *NS-Low Runoff state*, suggesting students achieved low runoff through various approaches and did not backtrack once they had satisfied at least one of the design constraints.

A recurring theme among the *conservative designers* is their focus on cost-driven transitions. The *CostS-Low Runoff* state is the only sub-optimal state with incoming edges from every other state, indicating that students consistently prioritized optimizing the cost of their designs regardless of their starting point. Moreover, the only outgoing edge from the *CostS-Low Runoff* state leads to the optimal state, suggesting that students made a concerted effort to improve cost by either experimenting within the *CostS-Low Runoff* state (*prob=0.62*) or they transitioned to the optimal state (*prob=0.18*) by meeting the *blocks criteria* as well. The back and forth between the optimal *AllS-Low Runoff* state and the suboptimal *CostS-Low Runoff* state shows their cost optimization behavior.

On the other hand, *design space explorers* began their exploration with either optimal(*prob=0.46*) or *High Runoff* states (*prob=0.42*). Once in the *AllS-Low Runoff* or *High Runoff* states, *design space explorers* were likely to remain there (with probabilities of *0.65* and *0.52*, respectively). The *design space explorers* were more willing to experiment with different designs early in their process, as evidenced by the loop between the *NS-Low Runoff* and *High Runoff* states, suggesting that they were more willing to sacrifice low runoff in their pursuit of

satisfying the design constraints. A commonality in both the *conservative designers* and the *design space explorers* is that there are no incoming edges to *the BlockS-Low Runoff state,* indicating that both groups prioritized the other design constraints, like the *cost criterion* over the *blocks criteria*. The Markov model in Figure 3 also suggests that *design space explorers* prefer to minimize runoff while addressing design constraints simultaneously, for instance, in the *NS-Low Runoff* state, where *design space explorers* had a strong tendency to remain (*prob=0.41*) before eventually transitioning directly to the optimal state (*prob=0.26*).

5.4 RQ3: Influence of design behaviors on Final Design Solutions

We graded the final designs submitted in their notebooks using the scoring criteria described in Section 4.2. We found that *design space explorers* (*M=11.23, SD=0.91*) scored higher than the *conservative designers (M=10.62, SD=0.98)* (see Figure 2(b)*).* The Mann-Whitney U test showed a significant difference (*p=0.046, U=323.5, Z=1.968, r=0.297*) among the clusters. Further, analysis revealed that 10 of 20 *conservative designers* (50%) and 19 of 24 *design space explorers* (79.2%) submitted an *AllS-Low Runoff* designs; the remainder submitted sub-optimal designs with various constraint combinations. 10 *conservative designers* (50%) and three *design space explorers* (12.5%) submitted designs that achieved low runoff and met the *cost criterion*; 1 *design space explorers* pair (4.2%) submitted designs that only met *blocks criteria*; and one *design space explorers* pair (4.2%) submitted a design that met both the *cost* and *blocks criteria* but had *High Runoff*. Overall, only four *conservative designer pairs* (20%), while 11(45.8%) *design space explorer pairs* submitted the optimal design achieving 12 points.

6. Discussion

Our results demonstrate significant learning gains ($p < 0.001$, see Section 5.1), indicating that the WRC curriculum effectively facilitated students' understanding of science, computational thinking, and engineering design concepts. However, only 29 pairs (65.9%) submitted a design meeting all constraints with low runoff, suggesting that while most students grasp the primary concepts of the WRC curriculum, several struggle to apply these concepts effectively in real-life problem-solving scenarios.

Markov analysis revealed students' diverse approaches in their search for optimal design solutions. Students struggled to integrate multiple constraints into their decision-making process– for instance, the loop between *CostS-Low Runoff* and *AllS-Low Runoff* suggests that students' attempts to optimize one constraint often came at the expense of another– a trend also seen in prior research (Zhang et al, 2019; Purzer et al, 2015). While trade-offs are inherent in design tasks, the WRC curriculum does not explicitly emphasize the role of trade-off analysis in the problem-solving process. Incorporating this analytical perspective into the curriculum could help students develop critical design thinking skills.

Analysis of the submitted designs reveals that *design space explorers* indeed reported better design solutions than *conservative designers*. More *design space explorers* submitted an *AllS-Low runoff* solution, sometimes also the optimal solution, as reported in section 5.4. This finding also supports the claim that to optimize a design truly, engineers must sufficiently explore the possible design space (Montgomery et al, 2020) and underscores the need to encourage such exploration in design activities. We can assert that the patterns observed are likely intentional design changes rather than random exploration as if students' actions were random, optimal solutions would be unlikely or take considerably longer to reach as there are 6^{12} possible solutions to this engineering design challenge.

Additionally, we found that students' final designs prioritized low runoff and the *cost criterion,* values that are visually emphasized on the design interface*,* while overlooking the *blocks criteria*. "Informed" designers use words and graphics when considering plans or justifying decisions (Purzer et al, 2015). Our findings support this, and we conjecture that displaying additional criteria (e.g., the *blocks criteria*) within the ED interface may encourage students to focus more holistically on all design constraints. In future studies, we will incorporate this to display the design criteria on the interface.

Some students generated better solutions than the ones they reported as their final submissions. Our previous investigations (Zhang et al, 2019) found that fewer than 10% of students reported their best design. Recognizing this, we added a “design history” table in the simulation environment, allowing students to compare their current designs to ones they generated earlier. We conjecture that this feature supported students in identifying their best design, as 15 pairs (34.1%) submitted the “optimal design”, which was the best design they generated. This finding suggests that supporting the recording and reviewing of design solutions can scaffold students’ optimization processes.

While the results are specific to the WRC curriculum, our approach to analyzing design strategies is generalizable. The specific states depend on the problem statement and associated design constraints, the methods for categorizing solutions, and identifying strategies from sequences of design solutions are broadly applicable. For other multi-constraint challenges (e.g., Du et al., 2025; Montgomery et al., 2020; Bowen et al., 2016), solutions can similarly be divided into states and analyzed for strategy patterns. Future work will test and validate these methods across other engineering design and open-ended learning environments.

7. Conclusion and Future Work

This paper presents an effective approach to unpacking the problem-solving behaviors of middle school students by analyzing the sequences of their generated designs. This contributes to the limited research on the design behaviors of middle school students. Furthermore, the WRC curriculum is unique as it combines Science, CT, and ED while studying the relationships between science and engineering design in a manner proposed in the NGSS curricula (Lilly et al, 2022). The results demonstrate the effectiveness of the WRC curriculum and emphasize the need for scaffolding exploration to successfully engage with Engineering Design activities.

This research has implications for incorporating Engineering Design tasks into STEM+C learning environments. Our findings suggest that students require support in performing tradeoff analysis and learning to compare and evaluate their design solutions. One potential way to assist with tradeoff analysis in these exploratory environments is to display students' design sequences and the design constraints that must be met on the interface. Students tend to focus more on the constraints visible on their screens. Future work should concentrate on effectively scaffolding students to systematically explore design solutions.

This work has limitations. The design histories used in this analysis did not capture intermediate actions between tests, so the Markov analysis misses nuances of students’ problem-solving strategies, such as reasons for specific changes from one design solution to the next. Although clustering analysis suggested distinct groups and FOMM revealed consistent patterns, further validation of problem-solving strategies will require analysis of student conversations during the design sessions. As this was a retrospective analysis, further research is needed to assess how our approach can improve instruction and scaffolding in classroom settings. In the future, we aim to leverage online analysis of students’ design solutions to identify struggling students and to provide targeted support for those using less effective exploration strategies during the design process.

Acknowledgements

We thank the National Science Foundation (#1742195), participating students, and teachers.

References

Ali, M., & Tse, A. W. C. (2023). Research Trends and Issues of Engineering Design Process for STEM Education in K-12: A Bibliometric Analysis. International Journal of Education in Mathematics, Science and Technology, 11(3), 695-727.

Basu, S., McElhaney, K. W., Rachmatullah, A., Hutchins, N. M., Biswas, G., & Chiu, J. (2022). Promoting computational thinking through science-engineering integration using computational modeling. In Proceedings of the 16th International Conference of the Learning Sciences-ICLS 2022, pp. 743-750. International Society of the Learning Sciences.

Bowen, B. D., DeLuca, V. W., & Franzen, M. M. S. (2016). Measuring how the degree of content knowledge determines performance outcomes in an engineering design-based simulation environment for middle school students. Computers & Education, 92, 117-124.

Du, H., Zhu, G., Xing, W., & Xie, C. (2025). What Distinguishes Students' Engineering Design Performance: Design Behaviors, Design Iterations, and Application of Science Concepts. Journal of Science Education and Technology, 34(2), 314-326.

Gatta, R., Vallati, M., Lenkowicz, J., Rojas, E., Damiani, A., Sacchi, L., ... & Valentini, V. (2017, December). Generating and comparing knowledge graphs of medical processes using pMineR. In Proceedings of the 9th Knowledge Capture Conference (pp. 1-4).

Goldstein, M. H., Mejia, C. V., Adams, R., Purzer, Ş., & Zielinski, M. (2016, October). Developing a measure of quality for engineering design artifacts. In 2016 IEEE Frontiers in Education Conference (FIE) (pp. 1-7). IEEE.

Hutchins, N. M., Basu, S., McElhaney, K., Chiu, J., Fick, S., Zhang, N., & Biswas, G. (2021). Coherence across conceptual and computational representations of students' scientific models. Computer-supported collaborative learning.

Kelly, G. J., & Cunningham, C. M. (2019). Epistemic tools in engineering design for K-12 education. Science Education, 103(4), 1080-1111.

Lilly, S., McAlister, A. M., Fick, S. J., Chiu, J. L., & McElhaney, K. W. (2022). Elementary teachers' verbal supports of science and engineering practices in an NGSS-aligned science, engineering, and computational thinking unit. Journal of Research in Science Teaching, 59(6), 1035-1064.

McElhaney, K. W., Zhang, N., Basu, S., McBride, E., Biswas, G., & Chiu, J. L. (2020, January). Using computational modeling to integrate science and engineering curricular activities. In M. Gresalfi & IS Horn (Eds.). The Interdisciplinarity of the Learning Sciences, 14th International Conference of the Learning Sciences (ICLS) 2020 (Vol. 3).

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. Biochemia medica, 22(3), 276-282.

Montgomery, R., Greenwald, E., Crane, S., Krakowski, A., Barber, J.: Operationalizing optimization in a middle school virtual engineering internship. Journal of Science Education and Technology 29(3), 409–420 (2020). https://doi.org/https://doi.org/10.1007/s10956-020-09826-8

National Research Council, Division of Behavioral, Board on Science Education, & Committee on a Conceptual Framework for New K-12 Science Education Standards. (2012). A framework for K-12 science education: Practices, crosscutting concepts, and core ideas. National Academies Press.

NGSS Lead States. (2013). Next generation science standards: For states, by states. National Academies Press.

Purzer, Ş., Goldstein, M. H., Adams, R. S., Xie, C., & Nourian, S. (2015). An exploratory study of informed engineering design behaviors associated with scientific explanations. International Journal of STEM Education, 2, 1-12.

Radloff, J., & Capobianco, B. M. (2021). Investigating elementary teachers' tensions and mitigating strategies related to integrating engineering design-based science instruction. Research in Science Education, 51, 213-232.

Srivastava, N., Fan, Y., Rakovic, M., Singh, S., Jovanovic, J., Van Der Graaf, J., ... & Gasevic, D. (2022, March). Effects of internal and external conditions on strategies of self-regulated learning: A learning analytics study. In LAK22: 12th international learning analytics and knowledge conference (pp. 392-403).

Wendell, K. B., Wright, C. G., & Paugh, P. (2017). Reflective decision-making in elementary students' engineering design. Journal of Engineering Education, 106(3), 356-397.

Xing, W., Pei, B., Li, S., Chen, G., & Xie, C. (2023). Using learning analytics to support students' engineering design: the angle of prediction. Interactive Learning Environments, 31(5), 2594-2611.

Yu, K. C., Wu, P. H., & Fan, S. C. (2020). Structural relationships among high school students' scientific knowledge, critical thinking, engineering design process, and design product. International Journal of Science and Mathematics Education, 18(6), 1001-1022.

Zhang, N., Biswas, G., Chiu, J. L., & McElhaney, K. W. (2019). Analyzing students' design solutions in an NGSS-aligned earth sciences curriculum. In Artificial Intelligence in Education: 20th International Conference, AIED 2019, Chicago, IL, USA, June 25-29, 2019, Proceedings, Part I 20 (pp. 532-543). Springer International Publishing.

Zhou, N., Pereira, N., Chandrasegaran, S., George, T. T., Booth, J., & Ramani, K. (2021). Examining middle school students' engineering design processes in a design workshop. Research in Science Education, 51, 617-646.

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Quantifying the Benefits of the Reread-Before-Answer Strategy in Japanese High-School EFL

Isanka WIJERATHNE[a*], Brendan FLANAGAN[b] & Hiroaki OGATA[c]
[a]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
[b]*Center for Innovative Research and Education in Data Science, Kyoto University, Japan*
*Wijerathne.isanka.6z@kyoto-u.ac.jp

Abstract: Digital e-book platforms collect detailed logs of students' reading behaviors, yet it remains unclear whether brief revisits of previously viewed pages before answering open-book quiz questions significantly improve outcomes. This study investigates whether "Reread-Before-Answer" (RBA) loops, brief backward navigation immediately preceding quiz responses, impact quiz performance. Analyzing one month of BookRoll logs from 263 Japanese high-school English learners, we applied sequence mining to detect RBA loops and logistic regression to quantify their effects, controlling for session duration, navigation length, and prior English proficiency. Results show RBA loops occurred in approximately 40% of sessions and consistently predicted better performance, with learners who engaged in these loops showing significantly higher quiz accuracy compared to strictly linear readers. Pattern analysis further revealed that successful students preferred concise review behaviors over prolonged rereading. These findings highlight that encouraging short, intentional revisits to relevant content just prior to assessment can effectively enhance students' quiz performance. Consequently, integrating structured rereading prompts within digital textbooks could serve as a practical intervention for educators and instructional designers to optimize learning outcomes.

Keywords: Reread-before-answer (RBA), Navigation Motifs, Learning Analytics, Sequence Mining, Mixed-Effects Logistic Regression, Japanese High-school EFL

1. Introduction

Educational e-book platforms provide rich data on how students navigate digital texts. Analyzing these navigation patterns can reveal strategies that contribute to learning success. Prior research on systems like BookRoll, a digital textbook reader, has shown that logs of student reading behaviours (page flips, backtracking, annotations) can predict performance (Flanagan et al., 2022). The Learning and Evidence Analytics Framework (LEAF) integrates such e-book data for learning analytics research (Kannan et al., 2022). However, much of the early work focused on summative measures (e.g., total pages read or total time) rather than the sequence of actions (Valle Torre et al., 2024). Recent studies argue that learning is a temporal process and that simply aggregating click counts is limited in explanatory power (Valle Torre et al., 2024). Instead, examining how students move through content, i.e., their navigation sequences, can yield deeper insight into study strategies and their effectiveness (Woollaston et al., 2025).

One particular sequence pattern of interest is the reread-before-answer (RBA) loop, defined as a navigation loop where a learner, before answering a quiz question, goes back to review earlier pages and then returns to answer. Formally, an RBA loop can be represented as: *page i → page (i-k) → page i → answer*, where $k \geq 1$ denotes the number of pages the learner jumped back. This pattern reflects a deliberate backtracking to previously read material just before attempting an assessment item. Theoretically, such "constructive re-processing" aligns with findings from reading research: rereading text can foster deeper comprehension

and better recall, especially for second language (L2) learners. For example, Thomas and Healy (2012) found that rereading improved understanding in both first and second language reading tasks, supporting the idea that reviewing content strengthens memory and clarification of concepts. In an open-book learning context, an RBA loop may indicate a metacognitive strategy where the student verifies facts or concepts right before answering, potentially leading to higher accuracy. We hypothesize that learners who engage in RBA loops will achieve higher mastery on within-unit multiple-choice questions (MCQs) than those who follow a strictly linear reading path. Specifically, preliminary observations from our e-book platform suggested a performance gap: sessions containing an RBA loop had on average about 10% higher quiz accuracy than linear sessions with no backtracking. *Figure 1* illustrates this difference in our dataset: learners who backtracked at least once before answering tended to score higher on the unit quiz, as evidenced by the higher median and narrower high-performance spread for the RBA group (orange) compared to the linear group (blue). This motivates a deeper, controlled analysis of the RBA effect. To systematically investigate this phenomenon, we pose the following research questions (RQs):

1. **RQ1:** Is there a relationship between learners' reread-before-answer (RBA) behaviour and their multiple-choice-question (MCQ) performance?
2. **RQ2:** To what degree do RBA behaviours correlate with MCQ performance after controlling for time-on-page, sequence length, and prior English ability?
3. **RQ3:** Which navigation behaviours uniquely and significantly distinguish sessions with top-quartile MCQ performance?

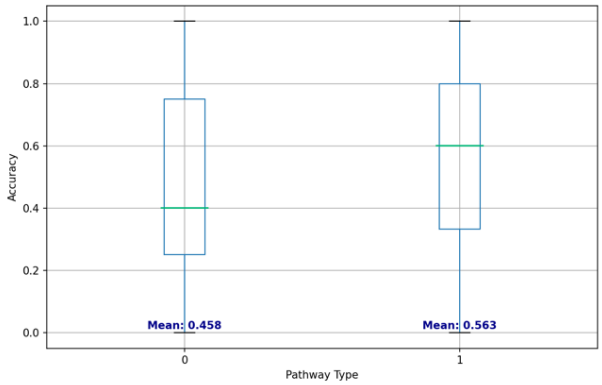


Figure 1. MCQ Accuracy by Pathway Type

In this study, we analyze anonymized e-book reading logs collected with informed consent from a Japanese high-school English course (Wijerathne et al., 2024) from a high school English course to answer these questions. We use sequence mining to detect RBA loops and other navigation motifs (Jacquemont et al., 2009) and apply mixed-effects logistic regression (with student and unit as random effects) (Ozechowski et al., 2007) to estimate the impact of RBA on quiz performance while controlling for potential confounds. Our findings confirm that RBA loops correspond to higher quiz success (about a 10.5 percentage-point gain raw, and ~6% gain after controls), and we identify specific navigation patterns prevalent among the top-performing students. We also discuss how these insights can inform the design of adaptive e-book features to support better study habits. The remainder of the paper is organized as follows: we first describe the dataset and methodology, then present results for each research question along with robustness checks, and finally discuss implications, limitations, and future work.

2. Methodology

2.1 Context and Data Collection

The data for this study come from a Japanese high school e-book learning environment in an EFL course. The platform consists of the BookRoll digital reading system integrated with the

LEAF learning analytics framework (Chen & Su, 2019; Flanagan & Ogata, 2018). BookRoll offers PDF educational resources and meticulously records user interactions. These records adhere to the xAPI standard, documenting events such as PDF access, page navigation, text highlighting, and quiz responses, along with their respective timestamps. Our analysis encompasses one month of interaction data from 263 high-school students engaging with 56 distinct PDF content units. Each content unit generally concludes with multiple-choice assessments designed to evaluate the students' comprehension of the material presented. Given the "open-book" format of these assessments, the logs reflect students' navigation patterns during quiz completion.

The raw log data comprised 57,532 events and 6,023 quiz answers, with a correct response rate of 55.4%. The logs were preprocessed for reliability before analysis. Timestamps were adjusted to Japan Standard Time, sorted, and any duplicates or errors were eliminated. Subsequently, the data was sessionized, grouping interactions into sessions defined by 30-minute inactivity periods. Consistent with widely adopted practice in learning-analytics sessionization (e.g., Kim et al., 2014; Kovanovic et al., 2015; Ming et al., 2019), we defined a new session after 30 minutes of inactivity in the BookRoll logs. Various session gap thresholds were evaluated for robustness (5, 15, 30, 45 mins), with 30 minutes selected for primary analysis due to its conventionality and logical representation of a session unit. After sessionizing, each event in the log was labeled with a session ID (*Table 1*).

Within each session, the navigation sequence of pages for each learner's attempt at the content was reconstructed. This sequence represents the ordered list of page IDs accessed by a student during that session before submitting their quiz answers. Additionally, quiz outcomes (correct/incorrect) were recorded alongside page transitions for the sequence. For instance, a simplified navigation sequence may appear as: [*Page 1 → Page 2 → Page 3 → Page 2 → Page 3 → Answer*]. In this case, the student accessed pages 1 through 3, returned to *page 2*, went back to *page 3*, and subsequently answered the questions. Such sequences are encoded as necessary to facilitate pattern detection.

2.2 Identifying RBA Loops

An RBA loop is defined as any purposeful backtracking that occurs immediately before a quiz attempt. For a session with ordered page numbers $(\boldsymbol{P_1}, \boldsymbol{P_2}, , \ldots, \boldsymbol{P_n})$ and a quiz event *Q*, we inspect the k pages that precede ***Q (k = 10)***. If within this window there is at least one index ***j*** where $\boldsymbol{P_j} > \boldsymbol{P_{j+1}}$, the learner has "jumped back" $\boldsymbol{\Delta = P_j - P_{j+1} + 1}$ pages. When $\boldsymbol{\Delta \leq k}$, and the quiz page either matches $\boldsymbol{P_{j+1}}$ or re-appears later in the window, the session is flagged ***has_RBA = 1***; otherwise ***has_RBA = 0***. The actual implementation is a single *O(n)* scan: using a 30-min inactivity rule, we obtained 1,843 learner-content sessions; 39.6% contained at least one RBA loop, implying that roughly two of every five study bouts involved rereading before answering. *Figure 1* contrasts MCQ accuracy by navigation type. RBA sessions show a higher median and narrower inter-quartile range, with a gap of ≈ 10 percentage points over linear sessions, an initial affirmative answer to RQ1. To test robustness, we repeated the detection with progressively stricter back-jump caps *(**k = 1…5**)*. Although the proportion of flagged sessions drops as ***k*** shrinks, the accuracy advantage remains virtually unchanged, indicating that even a single-page revisit is pedagogically beneficial (§3.3). For subsequent models (RQ1-RQ2) we analyze only the binary indicator, presence versus absence of an RBA loop, because the theoretical claim concerns whether strategic rereading occurs, not how many times or how far learners back-track.

2.3 Measures and Variables

The primary outcome variable is MCQ Accuracy, defined as the proportion of quiz questions a student answered correctly for each learner-content pair. In cases where a quiz contained only one question, this measure is binary (0 or 1), while quizzes with multiple questions yielded fractional scores between 0 and 1. Each learner's attempt on a given unit constitutes one data point. The main explanatory variable is the **presence of an RBA loop**, classifying sessions

as either RBA or linear. To control for potential confounds, we included three covariates. **Total Time-on-Page** represents the total time spent reading, calculated by summing the time between page events, with a cap on excessively long gaps to minimize idle-time inflation. This value was standardized using z-scores. **Sequence Length**, also standardized, refers to the number of distinct page transitions during a session. While linear readings of an n-page unit produce (n-1) transitions, sessions with backtracking result in longer sequences, which may reflect more thorough engagement or confusion. Finally, we included **Prior English Ability**, drawn from students' pre-course placement test scores (out of 100), standardized (mean=0, SD=1) for analysis. Missing values were addressed via multiple imputation. These covariates allow us to isolate the specific effect of RBA on MCQ accuracy, answering RQ2 comprehensively.

2.4 Statistical Analysis

To quantify the impact of RBA loops on quiz performance (RQ1 and RQ2), we employed a mixed-effects logistic regression model. Logistic regression is appropriate since our outcome can be considered binary at the quiz attempt level. We chose a mixed-effects approach to account for the nested structure of the data: multiple observations per student (some students attempted several contents) and per content unit (multiple students per the same content). By including random intercepts for Learner and Content Unit, we control for unobserved heterogeneity in individual ability (beyond the prior score) and question difficulty. This approach is similar to modeling each student and each unit having a baseline performance level, and we focus on the fixed effects that explain deviations from those baselines.
Our main regression model (Model 1) can be described as:

$$\boldsymbol{logit}\big(\boldsymbol{P}(\boldsymbol{Correct})\big) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} \cdot \boldsymbol{RBA_{present}} + \boldsymbol{\beta_2} \cdot \boldsymbol{TimeOnPage} + \boldsymbol{\beta_3} \cdot \boldsymbol{SeqLength} + \boldsymbol{\beta_4} \cdot \boldsymbol{PriorAbility} + (\mathbf{1} \mid \boldsymbol{Learner}) + (\mathbf{1} \mid \boldsymbol{Unit})$$

Here β_1 is the coefficient testing our hypothesis, a positive β_1 would indicate higher odds of a correct answer when an RBA loop is present, even after controlling for the other factors. We report this effect as an Odds Ratio (OR) for interpretability. An OR > 1 for RBA means the odds of quiz success are higher with an RBA loop than without. We also examine β_2 (time-on-page) and β_3 (sequence length) to ensure RBA isn't simply a proxy for "spent more time" or "looked at more pages."

Due to the absence of prior ability scores for all students, a Multiple Imputation (MI) method was employed to address missing data. Ten imputed datasets were generated, predicting missing scores through a regression-based imputer, and a logistic model was applied to each, with results aggregated via Rubin's rules (Royston, 2004). This approach provides a more reliable estimate without omitting students lacking prior scores. Additionally, a simplified model (Model 0) was executed without prior ability to evaluate the extent to which the RBA effect could be attributed to higher-skilled students' propensity to utilize RBA. The outcomes of these models are illustrated in forest plots. (see Results, *Figure 3*).

2.5 Pathway Pattern Mining

To address RQ3, we went beyond the RBA/non-RBA dichotomy and explored frequent navigation motifs in the sequences. We leveraged sequence pattern mining to identify which specific page transition patterns were over-represented among top-performing sessions. We first divided sessions by performance level: for example, "top quartile" sessions are those in the highest 25% of unit quiz scores (which in many cases means they answered all questions correctly), whereas "bottom quartile" would be low-scoring sessions. We then applied two complementary analyses: **Frequent Subsequence Mining:** Using the PrefixSpan algorithm (Bermudez et al., 2020), we mined for common subsequences of length 3-5 pages that occurred in at least 2% of sequences. PrefixSpan finds patterns like *"Page 2 → Page 3 → Page 2*" (a back-and-forth) or "*Page 1 → Page 2 → Page 3*" (a linear read) that appear frequently. **Motif Occurrence by Performance:** For the top 10 most frequent motifs, we computed how often each motif appeared in sessions of different performance quartiles. This

is visualized as a heatmap (*Figure 6*) where each row is a navigation motif and columns correspond to quartile groups, with cell color indicating the percentage of sessions in that group containing the motif. We also highlight which motifs have markedly higher presence in the top quartile compared to lower quartiles.

Additionally, we constructed a navigation transition network (pages as nodes, transitions as directed edges) and computed PageRank centrality for pages, as well as visualized overall flows with a Sankey diagram. While these network visualizations (e.g., *Figure 2*) are not the core of RQ3, they provided context, identifying which pages or transitions were "hubs" in the reading activity. For instance, if certain pages are frequently returned to by many students, that page might contain crucial content that warrants emphasis by instructors. By combining the statistical modeling and the motif analysis, we aim to paint a comprehensive picture: not only confirming if RBA loops help but also understanding what other navigation patterns characterize successful learning sessions. All analyses were conducted in Python, using libraries such as pandas for data handling, statsmodels for logistic regression, and custom scripts for sequence mining. The significance level was set at $\alpha = 0.05$ for hypothesis testing, and we report 95% confidence intervals for model coefficients.

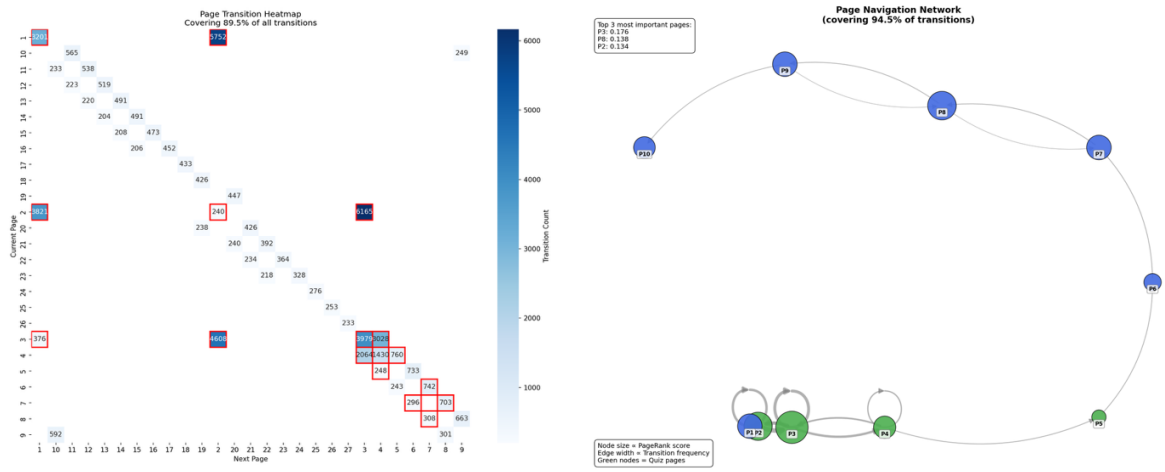


Figure 2. L: Page Transition Heatmap | R: Page Navigation Network

Table 1. *Raw MCQ Accuracy by Navigation Type and Session-Gap Threshold*

Gap (min)	RBA *n*	RBA *Acc*	Linear *n*	Linear *Acc*	Diff (pp)	Cohen's *d*
5	685	0.5651	188	0.4748	+9.0	0.299
15	720	0.5643	153	0.4576	+10.7	0.354
30	729	0.5629	144	0.4582	+10.5	0.347
45	730	0.5627	143	0.4584	+10.4	0.346

3. Results

3.1 RQ1: Relationship between RBA behaviour and MCQ accuracy

Sessions that include an RBA loop yield significantly higher MCQ accuracy compared to linear navigation sessions. Specifically, sessions containing at least one reread-before-answer loop had an average quiz accuracy of 56.4%, compared to 46.3% for purely linear sessions (*Table 1*), an absolute difference of roughly 10 percentage points in favor of RBA. In practical terms, a student who reviewed pages before answering got roughly 1 extra question correct out of 10, compared to a student who never looked back. This gap is illustrated in *Figure 1* (described earlier), where the distribution for RBA sequences is shifted upward relative to linear sequences. A two-proportion z-test confirmed the gap was highly significant (Cohen's $d \approx 0.35$, $z = 4.9$, $p < .001$), reinforcing the practical importance of the 10-percentage-point advantage. It is important to note that RBA sessions also tended to involve slightly more reading activity -

for example, they had a longer median sequence (the median number of page transitions was 12 for RBA vs. 9 for linear) and on average a bit more time spent. This raised the question addressed in RQ2: is the performance gain simply due to spending more time or other factors, or is there an independent contribution of the RBA strategy?

3.2 RQ2: Effect of RBA after controlling for time, sequence, and ability

To address RQ2, we utilized mixed-effects logistic regression models, with results displayed in Figure 3 (L and R) as forest plots of effect sizes (log-odds coefficients with 95 % confidence intervals) for RBA and other predictors. Figure 3 (L) excludes prior ability, while Figure 3 (R) includes it via multiple imputation. Both models assess whether students answered MCQs correctly, with fixed effects for RBA presence, total time, and sequence length. Significant random effects for student and unit highlight variability across individuals and content. In Figure 3 (L), without prior ability, the RBA Loop Present coefficient is approximately 0.208 log-odds (odds ratio ~1.23), indicating a 23 % higher likelihood of answering correctly in RBA sessions compared to linear sessions.

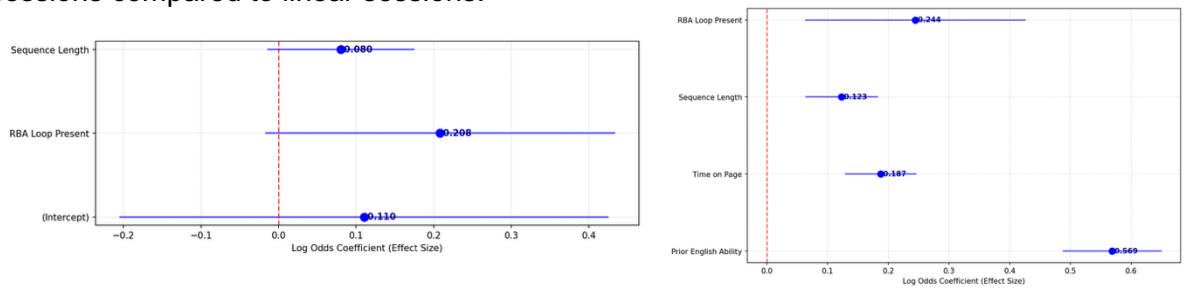


Figure 3. Effect size of predictors L: Without prior ability | R: With Multiple Imputation 95% CIs

Sequence length shows a modest but reliable benefit (*β = 0.123, OR = 1.13, 95 % CI [1.07, 1.20], p < 0.001*), indicating that visiting more pages is associated with slightly higher success. Time-on-page has a comparable, significant effect (*β = 0.188, OR = 1.21, 95 % CI [1.14, 1.28], p < 0.001*). When prior ability is included (*Figure 3 R*) and after controlling for time-on-page and sequence length, the RBA loop remains statistically significant (*β = 0.244, SE = 0.092, p = 0.008*), corresponding to an odds ratio of 1.28 (*95 % CI [1.07, 1.53]*). This translates to a roughly 6-percentage-point increase in predicted success probability, for example, a typical linear session's 46 % chance of a correct answer rises to about 52 % with an RBA loop. Prior English ability is the strongest predictor (*β = 0.569, OR = 1.77 per SD, 95 % CI [1.63, 1.91], p < 0.001*), reinforcing the importance of baseline proficiency. Random effects still indicate meaningful variance among students and content difficulty.

In summary, after adjusting for confounding factors, RBA loops offer about a 28% increase in quiz success odds (5-6% absolute gain), supporting our hypothesis that reviewing material before answering enhances immediate learning outcomes.

3.3 Robustness Checks: Session Definitions and Loop Size

We conducted several checks to confirm the robustness of our results: **Varying Session Gap Threshold:** We analyzed different session break definitions (5 min, 15 min, and 45 min). The prevalence of RBA patterns and their accuracy effects were consistent (*Table 1*). At 15-minute gaps, RBA loops appeared in ~38% of sequences with a ~9% accuracy difference; at 45 minutes, ~37% prevalence and ~10.5% difference. The logistic regression odds ratio for RBA remained around 1.25-1.30, indicating our conclusions are stable regardless of session definitions. **Loop Size (k) Sensitivity:** We examined how backtrack distance affects outcomes. At k = 1, RBA prevalence was nearly 0%, as most users moved forward. At k = 2, about 23% had back-jumps; at k = 3, 26%; k = 4, ~37.5%; and k = 5, ~38.6%. This indicates RBA loops tend to be around 4 pages long, capturing most instances. The accuracy benefit was ~5% points for k = 2 or 3, increasing to 9.1% at k = 4, with no further gains at k = 5. This suggests reviewing 4 pages yields optimal benefits. *Figure 4 (R)* shows that RBA advantages are consistent across session-gap definitions, with medium- and high-ability groups

outperforming linear sessions by 8-12 percentage points, and low-ability learners seeing a 3-5 percentage points (pp) lift. Short backtracks (1-2 pages) had less impact, while longer backtracks (beyond 4 pages) showed no added benefit, indicating that brief reviews before quizzes are most effective (see *Figure 5*). **Model Variants**: Alternative modeling approaches were explored, showing that students who frequently used RBA had higher mean scores. A logistic model excluding sequence length or time maintained a significant RBA effect, indicating it is not confounded by these variables. A mixed-effects model on complete cases also produced a similar odds ratio for RBA (~1.26), suggesting that imputation did not introduce bias.

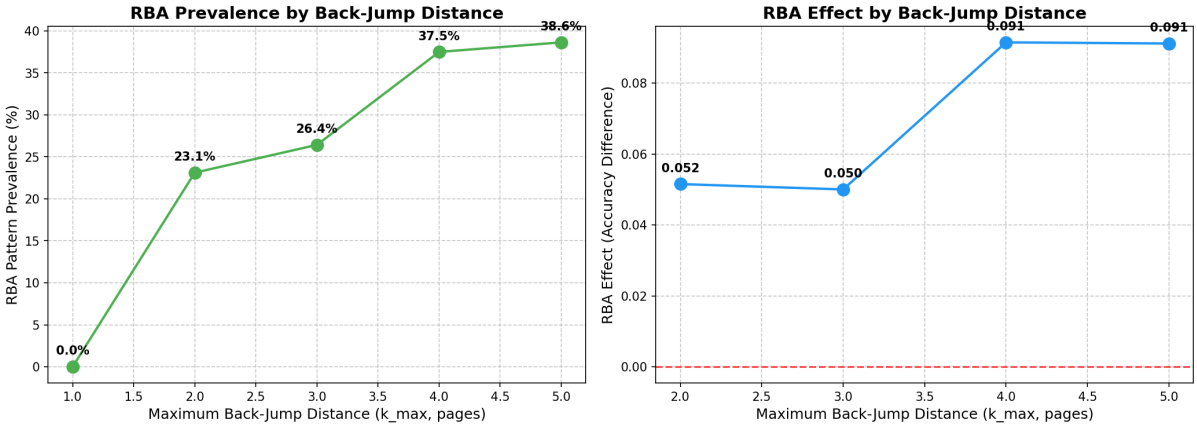


Figure 4. K-Sensitivity | Prevalence (L) & Impact (R) of RBA Loops by Back-Jump Distance

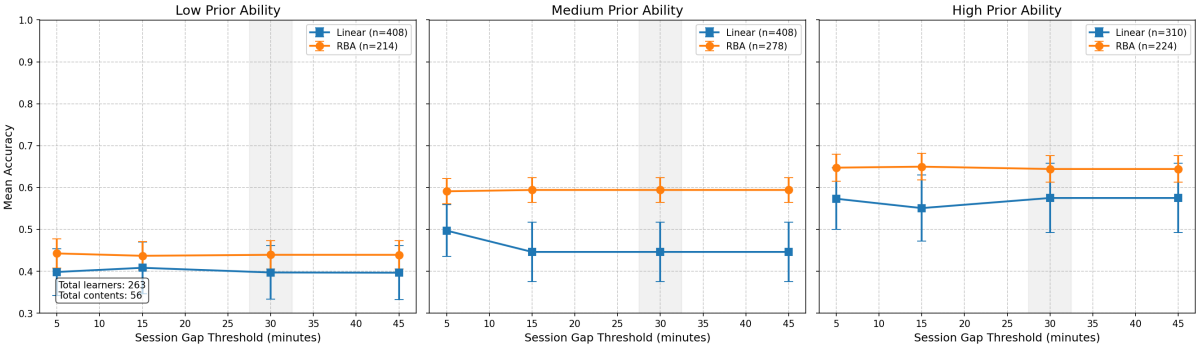


Figure 5. Accuracy by Session Gap Threshold, Navigation Type, and Prior Ability

In conclusion, students who reread content non-linearly before answering open-book quizzes performed 5-10 percentage points better in accuracy compared to those who read straight through, supporting the hypothesis that RBA loops are an effective learning strategy in this e-book context.

3.4 RQ3: Common Navigation Motifs in Top-Quartile MCQ-Accuracy Sessions

What navigation patterns characterize the most successful learners beyond RBA loops? We analyzed log sequences for frequent motifs, comparing high vs. low performing sessions. *Figure 6* shows a heatmap of the top 10 navigation patterns, with each row representing a specific sequence of page transitions (2 or 3 steps). The count at the start of each row indicates how often that pattern occurred in 500 sampled sequences. For example, "438 (87.6%): *2 → 3*" indicates this transition occurred in 438 sequences (87.6%). The second row "449 (89.8%): *1 → 2*" is even more frequent, as many sessions start at page 1. These linear patterns are common across all sessions and do not distinguish high vs. low performers. More notable are the backward transitions like "*3 → 2*," found in 326 sequences (65.2%), which indicate RBA-type loops. These backtracks, such as "*4 → 3*" and "*2 → 1*," were significantly more common in high-scoring sessions. The heatmap shows that high-accuracy sessions (left)

have more purple marks in the backtrack rows, indicating these patterns are present, while lower performers (right) show more yellow, indicating their absence.

Over two-thirds of top-quartile sessions involved backward navigation (e.g., from page 3 to 2), while less than half of bottom-quartile sessions did. High performers frequently revisited pages, with notable motifs such as "*3 → 3*" (self-loop on page 3) appearing in 252 sequences (50.4%). This indicates students spent extra time on the same page, likely to re-read or take notes. Self-loops like "*1 → 1*" (92.8% of sequences) suggest many sessions involved repeated access to page 1, indicating students returned to summaries or key points. Another common motif was "*2 → 1*" (309 sequences, 61.8%), where top students often returned to page 1 for reviews, a behaviours less common among lower performers. We also examined 3-step motifs, such as "*4 → 3 → 4*" (140 sequences, 28.0%), showing thorough review loops between pages 3 and 4. These patterns were predominantly found in higher quartile sessions, indicating that successful learners engaged in back-and-forth navigation rather than linear reading.

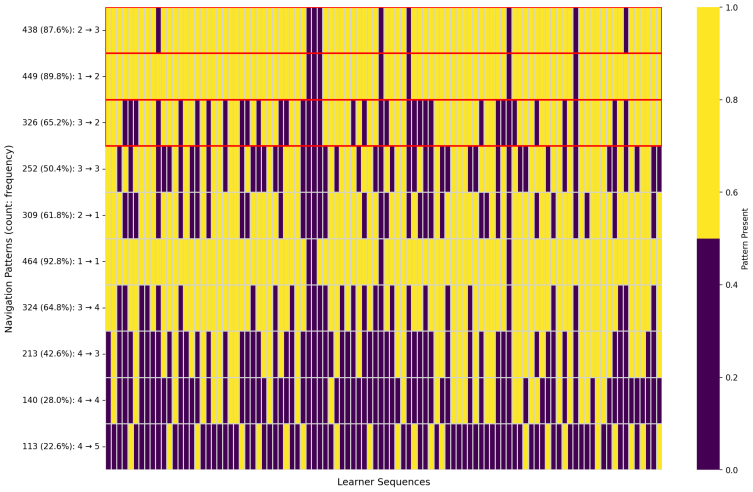


Figure 6. Top Navigation Patterns (Motifs)

In summary, the most successful sessions exhibited higher instances of review-oriented navigation motifs. While simple forward-reading (*1→2→3…*) was common, top quartile students distinguished themselves with strategic backtracking. Patterns of revisiting pages were linked to deeper engagement with content, as illustrated by an example involving English grammar, where high-performing students effectively navigated between pages to reinforce understanding, contrasting with lower performers who often missed critical details by not revisiting earlier content.

4. Discussion and Limitations

This research elucidates the relationship between digital textbook navigation and educational outcomes, with a particular focus on rereading before answering (RBA) as an advantageous pedagogical strategy. Participants who engaged in RBA cycles attained superior quiz scores, irrespective of the total duration of study or their prior capabilities (as visualized in *Figure 5*). These findings are consistent with cognitive theories that prioritize self-regulated learning and reading comprehension, which are especially pertinent in contexts of EFL (Jafarigohar & Morshedian, 2014), where learners frequently revisit previous pages to enhance their understanding of linguistic elements. Our results reinforce earlier investigations, such as those conducted by Flanagan et al. (2022), which observed enhanced performance linked to page revisits. Notably, our study emphasizes that the sequence of study, not merely the duration of study, exerts a significant impact on learning, thereby advocating for a temporal analysis within the domain of learning analytics. However, several limitations necessitate consideration:

- **Causality**: The correlational design of our study constrains the ability to make definitive assertions regarding causation. Although prior ability was controlled for, extraneous variables such as student motivation or metacognitive skills may affect the utilization of RBA. Experimental investigations that explicitly prompt RBA could yield clearer causal relationships.
- **Immediate vs. Long-term Learning**: The results of our study primarily address immediate quiz performance. The implications for long-term retention attributed to RBA remain ambiguous; subsequent research should investigate prolonged assessments to substantiate these findings.
- **Generalizability**: The specific context of this study, a high school EFL curriculum in Japan, may constrain the generalizability of the results. Outcomes could vary across different subjects or in contexts requiring closed-book assessments. Nonetheless, the efficacy of active content review is likely to persist across various settings. Future research should examine analogous patterns within diverse educational environments.
- **Data Interpretation**: Deducing strategies from log data bears the risk of mischaracterizing certain page revisits as constructive, when they may instead signify confusion or distraction. The incorporation of qualitative methodologies such as screen capture or eye-tracking could elucidate the intentions of learners more clearly.
- **Highlighting and Note-taking**: Although the act of highlighting was captured indirectly, the explicit analysis of this behaviour was limited. Given anecdotal support linking highlighting to academic success, further motif analysis encompassing annotation practices could enhance comprehension.
- **Motif Analysis Scope**: Our motif analysis was predominantly centered on prominent navigation behaviours. More nuanced actions merit comprehensive statistical scrutiny (e.g., motif enrichment testing) for a more profound understanding of navigation strategies related to performance.
- **Platform Limitations**: The reliance on BookRoll logs necessitates continuous online engagement, which may overlook offline study activities. Additionally, the imposition of a 30-minute session threshold could result in an underestimation of session continuity. Future inquiries might implement varied definitions of sessions to affirm the robustness of findings.

Despite these limitations, the persistent correlation between RBA cycles and improved educational outcomes underscores the potential significance of promoting strategic review behaviours among students.

5. Implications and Future Work

Our findings offer practical implications for students, educators, and e-learning systems. Students, particularly those who tend to rush quiz attempts, may benefit from explicit guidance on adopting a "Rereading-Before-Answer" (RBA) strategy. Our data indicate that brief review loops significantly increase quiz accuracy, notably improving lower-ability learners' performance from approximately 40% to 45%. E-book platforms could facilitate this by incorporating adaptive prompts encouraging students to review critical pages before answering quiz questions. Additionally, platforms can enhance navigation by highlighting pages frequently revisited by high-performing students, thereby implicitly sharing effective study strategies. Teacher dashboards displaying students' navigation patterns could provide early indicators of struggling learners, enabling timely pedagogical interventions. Future research can expand these findings by applying similar analyses across varied datasets, including different educational contexts or disciplines, to test RBA's broader applicability. Longitudinal studies could further examine the impact of systematically training students to adopt RBA, assessing long-term improvements and the efficacy of adaptive interventions. Additionally, employing advanced machine-learning methods, such as sequence-to-sequence modeling, could reveal nuanced navigation profiles and their correlation with learning outcomes. Integrating semantic analyses via NLP with log data might enable predictive

recommendations, thereby enhancing personalized tutoring functionalities. Extending beyond e-books, exploring analogous rereading behaviours in video-based learning or interactive simulations could provide cross-media validation and further strengthen the generalizability of these insights.

In conclusion, reinforcing timely content review substantially improves student performance. Leveraging digital learning environments' analytic capabilities allows transforming passive data collection into proactive instructional support.

Acknowledgements

This work was partly supported by JSPS Grant-in-Aid for Scientific Research (B) JP20H01722 and JP23H01001, (Exploratory) JP21K19824, (A) JP23H00505, and NEDO JPNP20006.

References

Bermudez, R. S., Sison, A. M., & Medina, R. P. (2020, January). Extended PrefixSpan for efficient sequential pattern mining in a game-based learning environment. In Proceedings of the 2020 2nd Asia Pacific Information Technology Conference (pp. 118-122).

Chen, C.-H., & Su, C.-Y. (2019). Using the BookRoll e-book system to promote self-regulated learning, self-efficacy and academic achievement for university students. *Journal of Educational Technology & Society*, *22*(4), 33-46.

Flanagan, B., Majumdar, R., & Ogata, H. (2022). Early-warning prediction of student performance and engagement in open book assessment by reading behavior analysis. *International Journal of Educational Technology in Higher Education*, *19*(1), 41.

Flanagan, B., & Ogata, H. (2018). Learning analytics platform in higher education in Japan. *Knowledge Management & E-Learning: An International Journal*, *10*(4), 469-484.

Jacquemont, S., Jacquenet, F., & Sebban, M. (2009). Mining probabilistic automata: a statistical view of sequential pattern mining. *Machine Learning*, *75*, 91-127.

Jafarigohar, M., & Morshedian, M. (2014). The effect of self-regulation on improving EFL readers' ability to make within-text inferences. *Issues in Language Teaching*, *3*(2), 286-263.

Kannan, V., Warriem, J. M., Majumdar, R., & Ogata, H. (2022). Learning dialogs orchestrated with BookRoll: effects on engagement and learning in an undergraduate physics course. *Research and practice in technology enhanced learning*, *17*(1), 28.

Kim, J., Guo, P. J., Seaton, D. T., Mitros, P., Gajos, K. Z., & Miller, R. C. (2014, March). Understanding in-video dropouts and interaction peaks in online lecture videos. In *Proceedings of the first ACM conference on Learning@ scale conference* (pp. 31-40).

Kovanovic, V., Gašević, D., Dawson, S., Joksimovic, S., & Baker, R. (2015). Does time-on-task estimation matter? Implications on validity of learning analytics findings. *Journal of Learning Analytics*, *2*(3), 81-110.

Ming, G. A. O., ZHANG, J., Di, S. U. N., & ZHANG, J. (2019, December). Using Clickstream to Understand Learning Paths and the Network Structure of Learning Resources: Using MOOC as an Example. In International Conference on Computers in Education.

Ozechowski, T. J., Turner, C. W., & Hops, H. (2007). Mixed-effects logistic regression for estimating transitional probabilities in sequentially coded observational data. *Psychological methods*, *12*(3), 317.

Royston, P. (2004). Multiple imputation of missing values. *The Stata Journal*, *4*(3), 227-241.

Thomas, H. K., & Healy, A. F. (2012). A comparison of rereading benefits in first and second language reading. *Language Learning*, *62*(1), 198-235.

Valle Torre, M., Oertel, C., & Specht, M. (2024, March). The sequence matters in learning-a systematic literature review. In Proceedings of the 14th Learning Analytics and Knowledge Conference (pp. 263-272).

Wijerathne, I., Flanagan, B., & Ogata, H. (2024). Empowering Educational Researchers with a Privacy-Centric Data Platform: Design, Implementation, and Implications. In International Conference on Computers in Education, (pp. 434-436).

Woollaston, S., Flanagan, B., Ocheja, P., Toyokawa, Y., & Ogata, H. (2025, March). ARCHIE: Exploring Language Learner Behaviors in LLM Chatbot-Supported Active Reading Log Data with Epistemic Network Analysis. In Proceedings of the 15th International Learning Analytics and Knowledge Conference (pp. 642-654).

Investigating the Impact of Time Constraints on Sudoku Solving: A Multimodal Analysis of Cognitive and Affective States

Deepika YADAV, Tushar PATLE *, Sonika PAL, Amit PAIKRAO, and Ritayan MITRA
Centre for Educational Technology, Indian Institute of Technology, Bombay
*95tusharpatle@gmail.com

Abstract: Time constraint during problem solving is an active area of research and has been found to affect performance, emotion, and cognition in students. In this study, we use galvanic skin response (GSR), eye tracking, and retrospective interviews to investigate such effects during Sudoku problem-solving. A within-subjects experimental design was employed, where participants solved 4x4 Sudoku puzzles with and without time constraints. The results indicated a significant increase in skin conductance and pupil diameters, with large effect sizes (GSR and pupil diameter values, dz = 3.09 and 1.98, respectively), reflecting substantial rises in stress and cognitive load under time constraints. Analysis of gaze patterns and interview responses revealed no notable change in problem-solving strategies between conditions, indicating stability in individual approaches. Together, these findings contribute to the understanding of the impact of time constraints on problem-solving behavior with implications for designing adaptive learning systems and educational practices that optimize cognitive load management and enhance learner engagement.

Keywords: Time-constrained problem-solving, 4X4 Sudoku, Eye Tracking, Galvanic Skin Response (GSR), Cognitive load, Stress

1. Introduction

In the realm of education, time constraints play a pivotal role in shaping the learning experience and outcomes. The imposition of time limits can significantly influence cognitive processes and emotional responses, impacting how students engage with and approach problem-solving tasks (Sweller, 1988). Time pressure can enhance cognitive load and induce stress, potentially hindering performance and learning (Beilock & Carr, 2005; Sinha et al., 2016; Maule & Hockey, 1993). Some research has suggested that the introduction of time constraints can potentially heighten cognitive load and stress, consequently affecting the individual's ability to employ effective problem-solving strategies (Mann & Tan, 1993; Maule & Hockey, 1993; Sweller, 1988). For instance, Speier et al. (1999) examined the influence of task interruptions on individual decision-making, concluding that time constraints can lead to information overload and impaired decision quality. Conversely, appropriate time constraints may stimulate efficient problem-solving and improve focus (Mann & Tan, 1993). Investigating such effects in more detail can provide valuable insights into designing learning environments that optimize performance while avoiding stress and cognitive overload.

Existing research has explored the impact of time constraints on learning and problem-solving in various contexts, highlighting their significant role in shaping cognitive processes and emotional responses (Gonzalez, 2004; Maule & Hockey, 1993). Rothstein (1986) found that time pressure increases cognitive load and negatively impacts judgment/decision quality under time pressure, while Sinha et al. (2016) showed that time constraints heightened emotional arousal, evident in increased skin conductance.

Time-constrained tasks like timed tests, rapid decision-making tasks, and high-pressure simulations help understand their impact on problem-solving. Sudoku puzzles, which require rapid logical reasoning, are widely recognized as cognitively demanding and suitable for such research (Patil et al., 2020). The cognitive processes involved in solving Sudoku puzzles have drawn interest from researchers aiming to understand the intricacies of human problem-solving and decision-making abilities (Chandra Prakash et al., 2017; Kalia et al., 2019; Leu et al., 2014). Sudoku offers a controlled setting where variables can be manipulated to examine their effects on cognitive load, emotional arousal, and strategy use. The theoretical underpinnings of this study are rooted in cognitive load theory (Sweller, 1988) and the well-established relationship between cognition and affect (D'Mello & Graesser, 2011).

Several studies have explored cognitive processes in Sudoku solving (Shakti et al., 2019), shedding light on its cognitive demands and strategies. Tran (2023) identified strategies like scanning rows/columns, box-checking, and logical deduction. However, the impact of time constraints on these processes remains largely unexplored. Beyond research, Sudoku has also been used pedagogically to engage students in developing logical deduction skills in mathematics classrooms (Kwan, 2010).

This study aims to bridge this gap by employing a multi-modal approach that combines eye-tracking, GSR, and interviews to investigate the effects of time constraints on cognitive load, stress, and problem-solving strategies in Sudoku puzzle-solving, where a within-subjects experimental design was employed (Charness et al., 2012; Paikrao & Mitra, 2023), wherein participants solved Sudoku puzzles under both conditions (with and without time constraints). The research questions are as follows:

RQ1. How do time constraints affect cognitive load and stress during Sudoku puzzle-solving?
RQ2a. What are the different strategies employed by the participant for solving Sudoku?
RQ2b. Do participants employ different problem-solving strategies when solving Sudoku puzzles with and without time constraints?

2. Study design and Methodology

2.1 *Procedure*

The study sample comprised individuals from the engineering student population, aged between 23 and 32 years. The sample consisted of 6 male and 2 female participants. We have selected participants based on their prior experience (Basic level experience) in solving Sudoku puzzles. A convenience sampling approach was used; although non-probabilistic, it yielded a mix of participants, adhering to ethical and privacy guidelines. Participants provided informed consent before the commencement of the study, and their identities were anonymized. To minimize arousal from external factors such as temperature and noise, the study took place in a controlled environment (air-conditioned, noise-canceling/absorbing rooms). Two stimuli were presented: Sudoku-1 without time constraints and Sudoku-2 with time constraints. Before starting, the participants were instructed to remain calm and composed while their eye movements were monitored using an eye-tracker and a GSR sensor was afixed to the non-dominant hand. The study activities were recorded with 'OBS Studio' screen-capture software. After completing the puzzles, participants participated in semi-structured retrospective interviews. During these interviews, they watched video recordings of their puzzle-solving sessions, which helped elicit insights into their problem-solving strategies, emotional responses, and cognitive processes, as shown in Figure 1.

2.2 *Presentation of Stimulus (Sudoku)*

The stimuli used in this study consisted of a set of 4x4 Sudoku puzzles, presented using Tobii Pro Lab software. There are two categories of Sudoku puzzles: Sudoku-1 (S1) without time constraints and Sudoku-2 (S2) with time constraints. Both categories are of easy difficulty level. The participants interacted with the Sudoku puzzles by selecting numbers (1 to 4) from a virtual keyboard displayed on the screen, positioned centrally to the right side of the Sudoku grid, as depicted in Figure 2. Upon completion of S1, participants would press the "fn + F10" button to advance to S2, which automatically progressed after a gap of 45 seconds. Before each puzzle, the participants would see a slide containing information about the activity, ensuring clarity regarding the task requirements. The sequence of stimuli presented to participants is outlined in Figure 3.

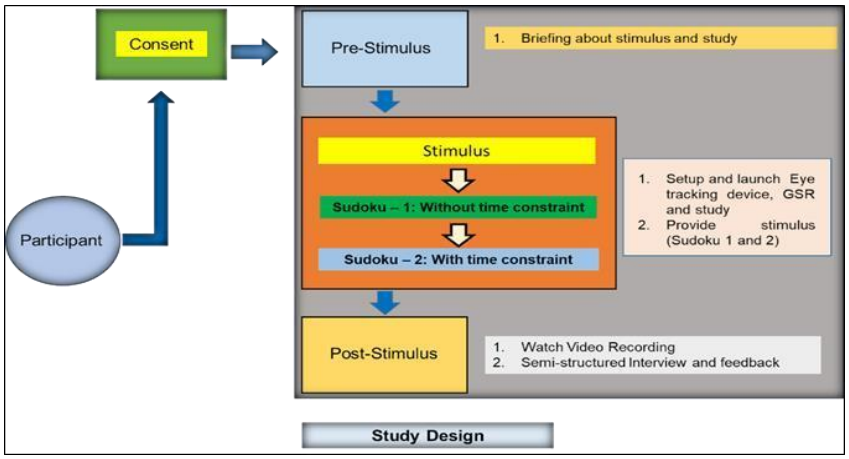


Figure 1. Study Design.

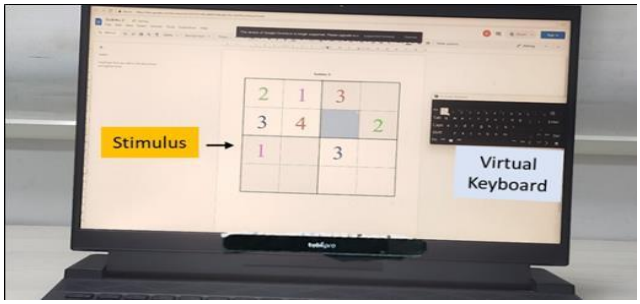


Figure 2. Presentation of the Stimulus on Tobii Pro Lab software.

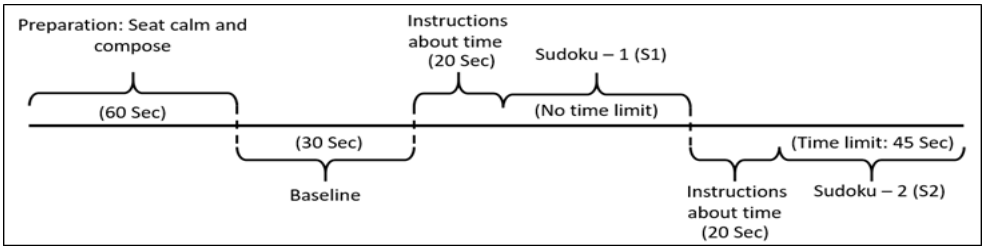


Figure 3. Sequence of tasks.

2.3 *Data Collection and Analysis*

A Shimmer3 GSR+ sensor (120 Hz) was used to measure arousal, which has been shown to indicate stress, engagement and cognitive load. Baseline GSR values were measured before each stimulus to establish a reference point. A Tobii Pro Nano (60 Hz) eye-tracking device was used to measure pupil diameter, gaze patterns, and fixation durations. These metrics provided insights into visual attention, cognitive load, and problem-solving strategies. Semi-structured

retrospective interviews were conducted after the completion of study to understand strategy selection and emotional arousal, such as stress or increased cognitive load due to time constraints.

To ensure the data was ready for analysis, we performed basic pre-processing on the GSR and eye-tracking data. For GSR, baseline values were recorded before each task and used to standardize the data, allowing us to compare participants' responses across the timed and untimed conditions. For the eye-tracking data, we used Tobii Pro Lab's built-in filters to automatically remove blinks and missing gaze points. Pupil diameter values were standardized for each participant to allow meaningful comparisons. Only those portions of data where the tracking accuracy was high (over 90%) were included in the analysis. The integration of multimodal data and their role in answering our research questions is depicted in Figure 4.

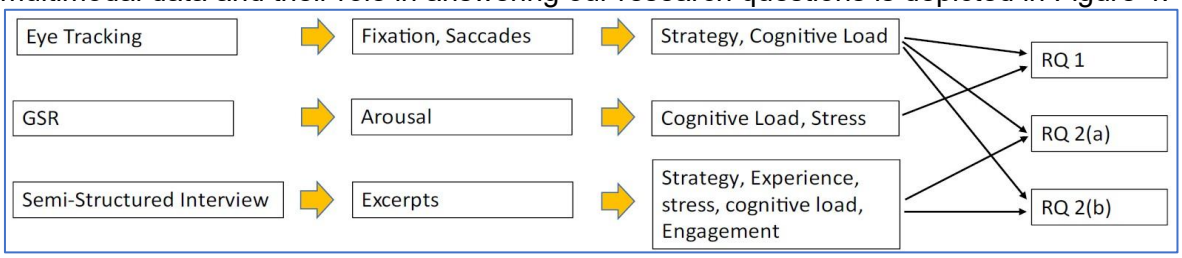


Figure 4. Mapping of Data analysis and tools with RQs.

3. Results

3.1 *Analysis of GSR Data*

Standardized GSR increased continuously during play and was higher in S2 than S1. A paired-samples t-test showed significantly higher mean standardized GSR under time constraints, $t(7)=8.74$, $p<.001$, $dz=3.09$ (Figure 5, 6). This robust effect size underscores the increased stress induced by temporal pressure (Sinha et al., 2016).

3.2 *Analysis of Pupil Diameter and Gaze Pattern*

Standardized pupil diameter was also higher in S2 than S1. A paired-samples t-test indicated larger mean standardized pupil diameter under time constraints, $t(7)=5.60$, $p<.001$, $dz=1.98$ (Figure 7, 8). , highlighting a significant influence of temporal pressure on pupil diameter (van der Wel & van Steenbergen, 2018). This substantial effect size underscores the heightened cognitive load caused by the time constraints.

Fixation sequences can establish the order of attending to important steps during problem solving (Mitra et al., 2017; Negi & Mitra, 2022). A comprehensive manual analysis of eye gaze patterns was performed to elucidate the strategies employed by the participants. We present here two scan paths illustrating the strategies employed by the participants (Figure 9), where their eye movements, indicated by red arrows (saccades), and green circles (fixation) followed a pattern of scanning along rows, columns, and boxes to locate possible cell placements. The yellow icons represent fixations, indicating the cells where participants filled in numbers based on their deductions.

3.3.1 *Rows and Columns Strategy*

In the Row and Column strategy, participants systematically scan the rows and columns of the Sudoku grid to identify given numbers and deduce placements. As illustrated in Figure 9, participants follow a sequence of fixations and saccades along rows and columns. Initially, they scan the rows to identify possible placements (as shown by the arrows indicating the sequence of fixations from one cell to another). Subsequently, they shift their focus to the columns, continuing this methodical scanning and deducing process until the entire Sudoku grid is filled.

This strategy relies heavily on a structured approach, where participants use logical deductions based on row and column intersections to place numbers accurately.

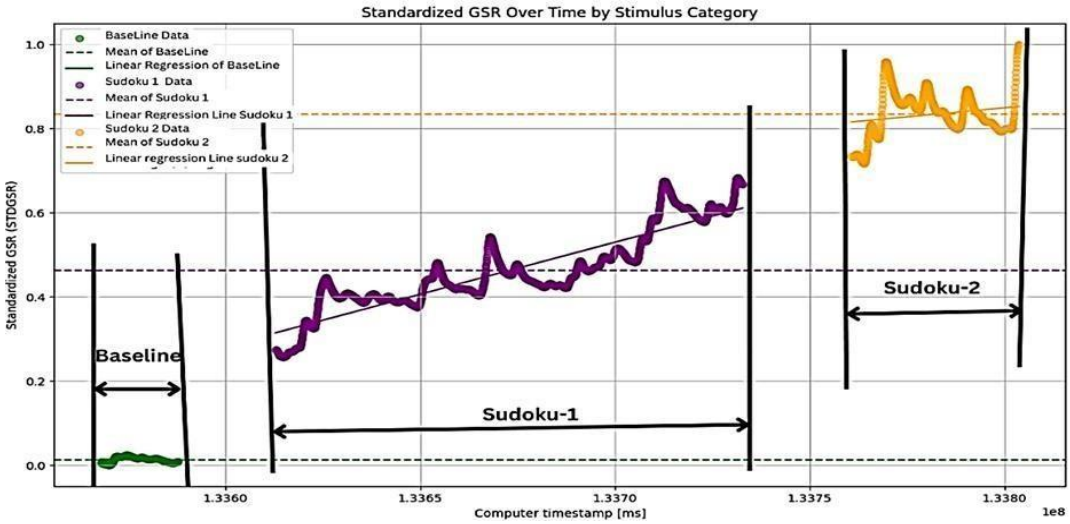


Figure 5. Plot for the standardized mean GSR values for Baseline, S1 and S2.

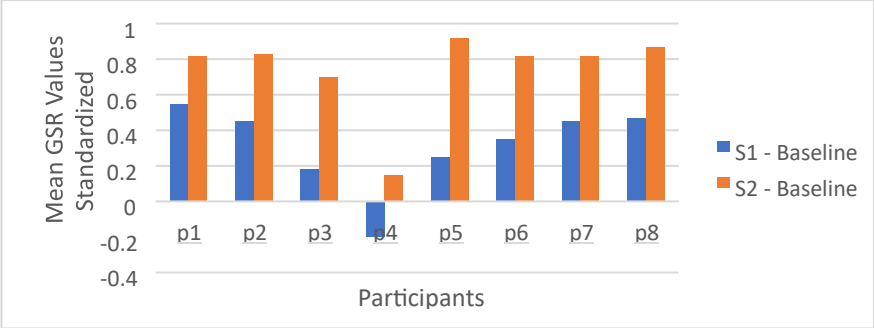


Figure 6. Bar plot of the standardized mean GSR values for S1-baseline and S2-baseline across all participants

3.3.2 *Rows, Columns and Box Strategy*

The Row, Column, and Box strategy adds another layer of scanning by incorporating the examination of 2x2 boxes within the Sudoku grid. As depicted in Figure 10, participants not only scan rows and columns but also focus on specific boxes. They start by examining the 2x2 boxes, using a combination of rows and columns scans to identify potential placements within these smaller grids. The figure shows the sequence of fixations within the boxes, highlighting the participants' methodical examination and deduction process. Once a box is solved, participants apply the same strategy to subsequent boxes, ultimately solving the entire Sudoku grid.

3.4 Analysis of Saccades

The examination of eye-tracking patterns provided valuable insights into visual attention and information processing during Sudoku solving (Krebs et al., 2021). Participants exhibited distinct gaze patterns associated with different problem-solving strategies, as evidenced by variations in fixation durations, saccade velocities, and gaze transitions. While there were variations in gaze behavior across individuals, common patterns emerged, particularly in response to time constraints. Specifically, in Sudoku-1 (S1), the average peak saccade velocity was 119.22°/s with a standard deviation of 118.62°/s, and the average saccade amplitude was 3.585 degrees. In Sudoku-2 (S2), these metrics increased, with an average peak saccade

velocity of 146.70°/s and a standard deviation of 145.37°/s, along with an average saccade amplitude of 4.19 degrees.

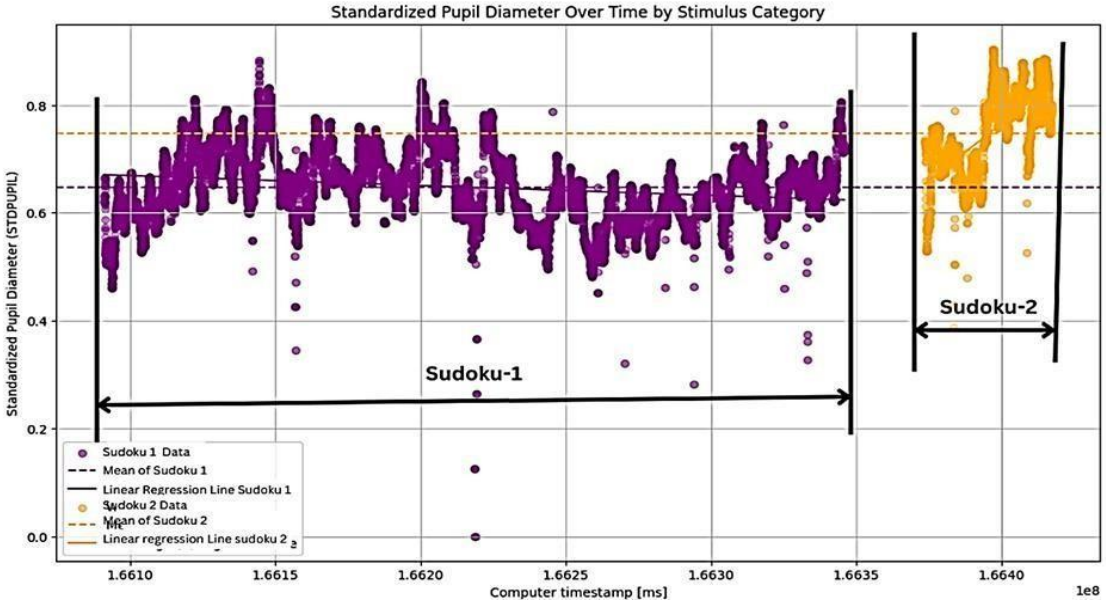


Figure 7. Plot for the mean pupil diameter values (Standardized) for S1 and S2.

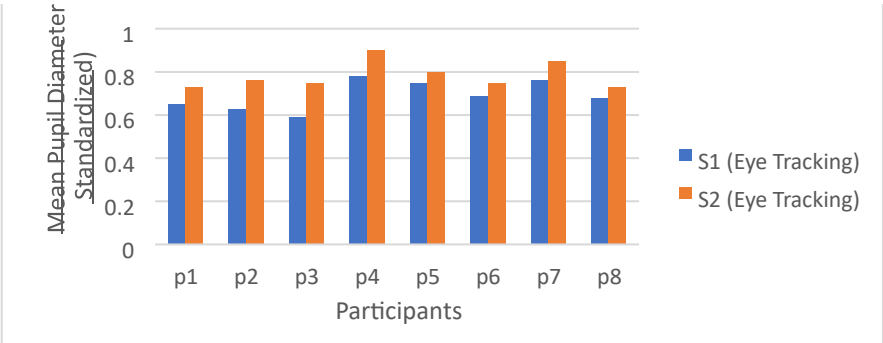


Figure 8. Bar plot for the mean pupil diameter values (Standardized) for S1 and S2

3.5 Analysis of Semi-structured Retrospective Interview

To validate the eye-tracking findings, a semi-structured retrospective interview was conducted with participants. The interviews revealed consistent problem-solving strategies employed by participants during Sudoku solving, irrespective of time constraints (Behrens et al., 2023). The interview analysis corroborated the strategies observed in Figure 9. For instance, Participant 1 stated, "*I was focusing on box and rows and followed the same strategy*," while Participant 6 mentioned, "*I was just looking at numbers and then started solving Rows and Columns. I started applying the same strategy in the time constraint problem.*" Participant 8 similarly noted, "*I started looking at the box and associated rows and columns to fill the box first and applied the same strategy.*" Other participants also reported using the same strategies for both puzzles, suggesting a tendency for individuals to rely on stable problem-solving methods regardless of temporal constraints.

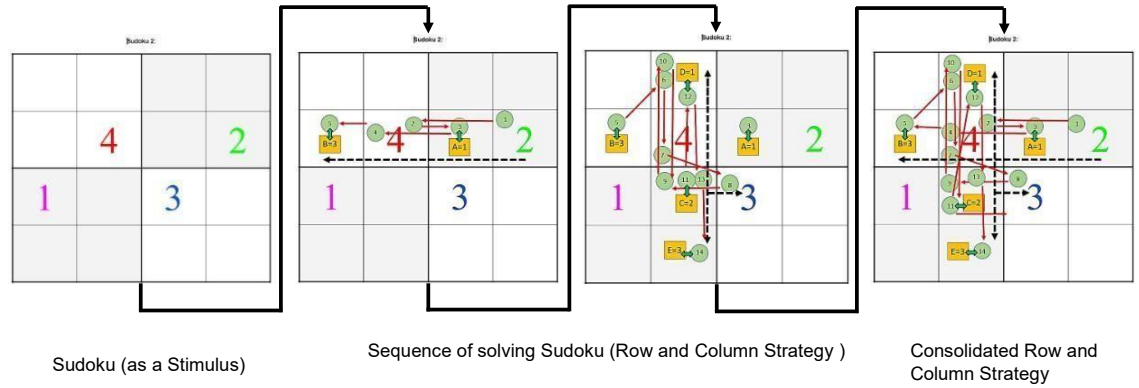


Figure 9. Problem-solving using rows and columns strategy: Participants scan the Sudoku grid by alternating between rows (horizontal arrows (ii)) and columns (vertical arrows (iii)) to identify and place numbers. The arrows indicate the sequence of fixations and saccades, showing a systematic approach to solving the entire grid.

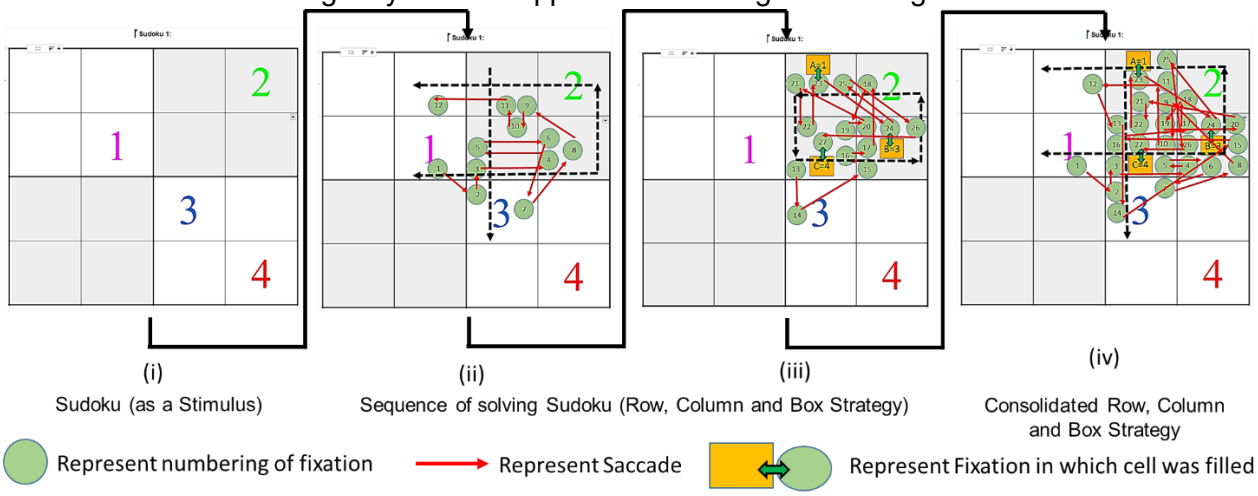


Figure 10. Problem-solving using rows, columns and box strategy: Participants start by scanning rows, and columns (ii), and examining specific 2x2 boxes (iii). Arrows show the sequence of fixations and saccades. Initially, focus on rows and columns associated with a box for deductions. The consolidated strategy (iv) shows solving the entire grid by methodically scanning and solving each 2x2 box and integrating these solutions.

Also, the retrospective interviews revealed notable affective components associated with time constraints during Sudoku puzzle-solving. Participants reported experiencing stress and nervousness due to the imposed time limit. For instance, Participant 1 mentioned, "*As there was a time limit, I was thinking about completing it fast*" reflecting a sense of urgency. Participant 4 remarked, "*Earlier I was thinking, 45 seconds are less to complete it….As time was passing, I was getting stressed and trying to solve it fast*". Similarly, Participant 6 noted, "*Initially I was nervous…then I thought, let's solve it and not think about time*" indicating initial stress followed by an adaptive coping strategy. Similarly, highlights the pressure felt due to the time constraints. These excerpts underscore the heightened stress levels, corroborating the physiological data obtained from GSR measurements. Although each modality, GSR, pupil dilation, gaze pattern, and interviews, was analyzed separately, the convergent findings across these measures support a coherent interpretation of heightened cognitive and emotional responses under time constraints.

4. Discussion

The findings of this study present a multimodal perspective on how time constraints influence cognitive load, stress levels, and problem-solving strategies during Sudoku puzzle-solving (Behrens et al., 2023). Addressing RQ1, the data presented compelling evidence that time constraints exert a significant influence on both cognitive load and stress levels. Pupil diameter

measurements, a reliable indicator of cognitive load (Krebs et al., 2021), revealed larger pupil diameters during Sudoku puzzle-solving under time constraints (Sudoku-2) compared to scenarios without time constraints (Sudoku-1). This suggests an increased demand for cognitive resources when participants were under temporal pressure. Concurrently, Galvanic Skin Response (GSR) measurements, known to reflect changes in emotional arousal (stress), exhibited higher mean GSR values in Sudoku-2, indicating heightened stress levels. The robust effect sizes further underscored the profound physiological reactions elicited by temporal constraints, emphasizing the substantial impact of time pressure on the cognitive and affective domains (D'Mello & Graesser, 2011). Although GSR data can be indicative of engagement (McNeal et. al., 2014) or other constructs as well, participants categorically noted increased stress in the time-constrained conditions during the interviews. This work's novelty lies in integrating GSR, eye-tracking, and retrospective interviews within a single study. Unlike prior work examining these modalities separately (e.g., Shakti et al., 2019; Krebs et al., 2021), our triangulated approach enhances validity and offers richer insights into cognitive and affective states under time-constrained problem-solving. Additionally, we feel compelled to note here the importance of multimodal data in coming to such conclusions.

Addressing RQ2(a), participants predominantly followed strategies such as the Rows, Columns, and Box, rows and box, or sequential solving of Rows and Columns. (Simonis, 2005; Patil et al., 2020). Triangulating retrospective interviews with eye-tracking revealed consistent strategy use, with gaze patterns aligning with self-reports and supporting prior work on Sudoku problem solving (Simonis, 2005). This multimodal approach demonstrates how combining qualitative and quantitative data yields deeper insight into cognitive processes. Importantly, it underscores the value of such methods in education: by monitoring indicators like GSR or gaze behavior, instructors can better calibrate task difficulty and timing, keeping learners challenged but not overwhelmed.

Addressing RQ2 (b), the study examined how time constraints influence Sudoku strategies. Interviews and eye-tracking revealed that participants largely relied on established methods, rows, columns, and boxes (Simonis, 2005), regardless of time pressure (Behrens et al., 2023). While core strategies remained stable, eye-tracking showed adaptations under pressure: shorter fixations, faster saccades, and more frequent gaze shifts (Krebs et al., 2021, Paikrao et al., 202). In particular, increased saccade velocity and amplitude in the time-constrained puzzle indicated heightened visual scanning to sustain the same approach within stricter limits. This aligns with findings that time pressure accelerates cognitive processing (Mahanama et al., 2022). Thus, although strategies were not fundamentally altered, participants modulated attention and search speed to manage temporal demands, illustrating adaptive cognitive control under pressure.

The integration of qualitative and quantitative data provides a comprehensive view of the cognitive processes involved in Sudoku problem-solving. The consistency in strategies, as reported by participants and supported by eye-tracking data, reflects the stability of preferred approaches across conditions. While we did not conduct integrated statistical modelling, the convergence of trends in GSR, pupil dilation, and participant reflections strengthens our interpretation. This triangulation underscores the value of a multimodal approach in distinguishing stress from engagement and capturing the complexity of problem-solving behaviour.

5. Conclusion

Our study contributes significantly to the literature by illuminating the effects of time constraints on cognitive and affective processes in task performance. First, we observed that time constraints led to heightened cognitive load and stress during Sudoku solving, as reflected in pupil diameter, GSR data, and participant interviews, aligning with prior research on the effects of temporal pressure on cognition and emotion.

Second, participants maintained consistent problem-solving strategies regardless of time constraints. However, we noted adaptive adjustments in visual attention, evidenced by subtle differences in eye movements. This suggests that individuals dynamically allocate cognitive resources in response to task demands, supporting existing literature on the flexible nature of cognitive processes.

Overall, our findings contribute valuable insights to various domains involving tasks under time pressure. By understanding how individuals respond to temporal constraints, researchers and practitioners can develop interventions to optimize cognitive and emotional regulation in educational, assessment, and decision-making contexts.

6. Limitations and Future Work

This study involved a small sample size, which may limit the range of strategy types and their variability captured, especially in the qualitative data. However, the statistically significant results, despite the limited sample, suggest a strong effect of time constraints. The use of a 4x4 Sudoku grid may also reduce task complexity and affect generalizability to more demanding problem-solving contexts. Nonetheless, the controlled structure allowed us to focus on physiological and strategic consistency. Future studies could explore more complex or open-ended tasks where strategy adaptation is more flexible. Additionally, integrating multimodal data using machine learning or latent modeling approaches could reveal deeper patterns across physiological and behavioral responses under time pressure.

References

Behrens, T., Räuker, M., Kalbfleisch, M., & Jäkel, F. (2023). Flexible use of tactics in Sudoku. *Thinking & Reasoning*, *29*(4), 488–530. https://doi.org/10.1080/13546783.2022.2091040

Beilock, S. L., & Carr, T. H. (2005). When high-powered people fail: Working memory and “choking under pressure” in math. *Psychological Science*, 16(2), 101–105. https://doi.org/10.1111/j.0956-7976.2005.00789.x

Chandra Prakash, V., Sastry, J. K. R., Anusha, K. B., Spandana, A. B., Dhatrija, N., & Nikhil, V. (2017). Applicability of Sudoku game for building the cognitive model of a student for career assessment - An analytical study. *International Journal of Engineering & Technology*, *7*(1.1), 246. https://doi.org/10.14419/ijet.v7i1.1.9479

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, *81*(1), 1–8. https://doi.org/10.1016/j.jebo.2011.08.009

D’Mello, S., & Graesser, A. (2011). The half-life of cognitive-affective states during complex learning. *Cognition and Emotion*, *25*(7), 1299–1308. https://doi.org/10.1080/02699931.2011.613668

Gonzalez, C. (2004). Learning to Make Decisions in Dynamic Environments: Effects of Time Constraints and Cognitive Abilities. *Human Factors*.

Kalia, V., Fuesting, M., & Cody, M. (2019). Perseverance in solving Sudoku: Role of grit and cognitive flexibility in problem solving. *Journal of Cognitive Psychology*, *31*(3), 370–378. https://doi.org/10.1080/20445911.2019.1604527

Krebs, C., Falkner, M., Niklaus, J., Persello, L., Klöppel, S., Nef, T., & Urwyler, P. (2021). Application of Eye Tracking in Puzzle Games for Adjunct Cognitive Markers: Pilot Observational Study in Older Adults. *JMIR Serious Games*, *9*(1), e24151. https://doi.org/10.2196/24151

Kwan, C. C. Y. (2010). Spreadsheet-based Sudoku as a tool for teaching logical deduction. Spreadsheets in Education (*eJSiE*), 4(1).

Leu, G., Tang, J., & Abbass, H. (2014). On the Role of Working Memory in Trading-Off Skills and Situation Awareness in Sudoku. In C. K. Loo, K. S. Yap, K. W. Wong, A. T. Beng Jin, & K. Huang (Eds.), Neural Information Processing (Vol. 8836, pp. 571–578). *Springer International* Publishing. https://doi.org/10.1007/978-3-319-12643-2_69

Mahanama, B., Jayawardana, Y., Rengarajan, S., Jayawardena, G., Chukoskie, L., Snider, J., & Jayarathna, S. (2022). Eye Movement and Pupil Measures: A Review. *Frontiers in Computer Science,* 3, 733531. https://doi.org/10.3389/fcomp.2021.733531

Mann, L., & Tan, C. (1993). The Hassled Decision Maker: The Effects of Perceived Time Pressure on Information Processing in Decision Making. *Australian Journal of Management*, 18(2), 197– 209. https://doi.org/10.1177/031289629301800204

Maule, A.J., Hockey, G.R.J. (1993). State, Stress, and Time Pressure. In: Svenson, O., Maule, A.J. (eds) Time Pressure and Stress in Human Judgment and Decision Making. Springer, Boston, MA. https://doi.org/10.1007/978-1-4757-6846-6_6

Mitra, R., McNeal, K. S., & Bondell, H. D. (2017). Pupillary response to complex interdependent tasks: A cognitive-load theory perspective. *Behavior Research Methods*, *49*(5), 1905–1919. https://doi.org/10.3758/s13428-016-0833-y

McNeal, K. S., Spry, J. M., Mitra, R., & Tipton, J. L. (2014). Measuring student engagement, knowledge, and perceptions of climate change in an introductory environmental geology course. *Journal of Geoscience Education*, *62*(4), 655-667.

Negi, S., & Mitra, R. (2022). Native language subtitling of educational videos: A multimodal analysis with eye tracking, EEG and self-reports. *British Journal of Educational Technology*, *53*(6), 1793– 1816. https://doi.org/10.1111/bjet.13214

Paikrao, A. M., & Mitra, R. (2023). Expert-Novice Differences In Electrical Circuit Analysis Based On The Order Of Attention. *European Society for Engineering Education (SEFI).* https://doi.org/10.21427/G9YB-5J15

Paikrao, A., Nath, D., Rajendran, R., & Mitra, R., (2025). Enhancing Retrospective Think-Aloud Protocols with Eye-Gaze data. In B. Jiang et al. (Eds.), Proceedings of the *33rd International Conference on Computers in Education.* Asia-Pacific Society for Computers in Education.

Patil, A. L., Deepak, K. K., & Kochhar, K. P. (2020). Role of prefrontal cortex during Sudoku task: fNIRS study. Translational Neuroscience, 11(1), 419–427. https://doi.org/10.1515/tnsci-2020-0147

Rothstein, H. G. (1986). The effects of time pressure on judgment in multiple cue probability learning. Organizational Behavior and Human Decision Processes, 37(1), 83-92.

Shakti, D., Saha, S., Sardana, V., Akula, P., Ananda, C. M., & Tewary, S. (2019). EEG as a tool to measure cognitive load while playing Sudoku: A preliminary study. 2019 3rd International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech), 1–5. https://doi.org/10.1109/IEMENTech48150.2019.

Simonis, H. (2005). *Sudoku as a Constraint Problem*.

Sinha, A., Das, P., Gavas, R., Chatterjee, D., & Saha, S. K. (2016). Physiological sensing based stress analysis during assessment. *2016 IEEE Frontiers in Education Conference (FIE)*, 1–8. https://doi.org/10.1109/FIE.2016.7757587

Speier, C., Valacich, J. S., & Vessey, I. (1999). The Influence of Task Interruption on Individual Decision Making: An Information Overload Perspective. *Decision Sciences,* 30(2), 337–360. https://doi.org/10.1111/j.1540-5915.1999.tb01613.x

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285.

Tran, A. (2023). Enhancing GNNs: An Exploration of Iterative Solving and Augmentation Techniques.

Van der Wel, R. P. R. D., & van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin & Review*, 25(6), 2005–2015. https://doi.org/10.3758/s13423-018-1432-y

Unpacking student engagement with everyday ethics in GenAI mediated dialogue

Lakshmi T G[a*], Gurudeo THAKARE[a], Alekh VELAYUDHAN[a] & Karishma SHANGHVI[a]
[a]*Shikha Institute of Education, Shikha Academy, Mumbai, India*
*lakshmi.ganesh@shikhaacademy.org

Abstract: With the rise of digital technologies, there is an increasing need for ethics education tailored to digital natives—children who have grown up in a technology-driven world. Traditional approaches to teaching ethics often fail to resonate with this generation, highlighting the need for innovative pedagogical methods. To address this, we designed a Generative AI (GenAI) chatbot that serves as a reflective agent, providing a platform for students to engage with ethical dilemmas and reflect on their values. Utilizing the Dialogue on Ethics (DoE) framework, this study analyzes the ethical thinking aspects during students' interactions with the chatbot. The findings suggest that students engage more actively with values through Interpretation and Judgement, moving beyond theoretical understanding to real-world ethical decision-making. The chatbot's role as a reflective agent, promotes a back-and-forth engagement with values. This research underscores the potential of AI-driven tools in supporting ethics education by encouraging dialogic pedagogy and empowering students to articulate and refine their personal ethical frameworks.

Keywords: ethics education, genAI ethics, dialogue on ethics framework, dialogic pedagogy

1. Introduction

The National Education Policy (NEP) 2020 of India, along with organizations like UNESCO, the World Bank, and the OECD, emphasizes the integration of values-based education as a central pillar of holistic development in schools (UNESCO, 2015; World Bank, 2018; OECD, 2019). However, teaching ethics to digital-native students, who are immersed in technology from an early age, presents new challenges. Traditional approaches to moral education often lack relevance or resonance with students accustomed to interactive, on-demand digital environments (Berkowitz & Bier, 2005).

Traditional ethics education often emphasizes abstract moral principles or adult-defined dilemmas that may lack relevance to children's lived experiences, particularly for digital-native learners who engage with technology from an early age (Bers, 2018). With rapid advancements in artificial intelligence (AI) and its increasing presence in daily life, new pedagogical challenges arise in teaching ethics that resonate with children's realities and fosters authentic moral reflection (Shaw & Brindle, 2020). Although digital tools and AI applications have been introduced to support ethics education, many remain limited in their capacity to facilitate dynamic, dialogic learning or to personalize engagement based on students' own values (Patry et al., 2008; Hollingsworth et al., 2003). This gap highlights the need for innovative approaches that integrate dialogic pedagogy (Wegerif, 2007) with AI technologies to nurture active and reflective ethical reasoning.

The present study applies this perspective to explore how 10- to 11-year-old students articulate and evolve their ethical reasoning through interaction with a generative AI (GenAI) chatbot. The learning experience was designed around dialogic pedagogy, where students act as co-constructors of ethical understanding rather than passive recipients of fixed moral content (Alexander, 2006; Wegerif, 2013). The chatbot serves as a reflective agent that

responds according to each student's personal value framework, enabling dialogues. Such a context offers a unique opportunity to explore the hypothesis—well-supported by moral development research—that *ethical growth emerges through iterative, situated dialogue, including interactions with non-human agents designed for pedagogical mirroring* (Brandel et al., 2024).

This study investigates student engagement with ethical values via GenAI chatbot conversations using the Dialogue on Ethics (DoE) framework (Cedar et al., 2021). It aims to unpack how students express and reflect on their chosen values and how the sequences of their moral reasoning unfold in dialogue with the chatbot. This approach advances the field of ethics education by situating ethical reflection within technology-enhanced dialogic contexts that align with the realities of digital-native learners. The findings have implications for the design of genAI-mediated pedagogies that foster authentic value-based learning and support the development of personal ethical frameworks in contemporary classrooms.

2. Theoretical Basis

2.1 *Dialogue for Ethical Development*

Ethical development in children has often been studied through frameworks emphasizing individual cognitive stages or externally imposed moral rules (Kohlberg, 1984). However, a growing body of research argues that moral understanding is socially constructed and dynamically shaped through interaction. From this perspective, dialogue becomes not just a method of communication but a core medium for ethical thinking and identity formation (Wegerif, 2013; Mercer et al., 2019).

Although ethics education has traditionally occurred in face-to-face contexts, digital platforms increasingly mediate values learning. Previous programs like VaKE (Values and Knowledge Education) and REACH Beyond Tolerance have incorporated structured digital or dialogic components to foster reflection and reasoning (Patry et al., 2008; Hollingsworth et al., 2003). However, Cedar et al. (2021) emphasize that many existing programs frame ethical development through adult-defined questions or moral exemplars, limiting opportunities for emergent ethical reflection. The potential of genAI and dialogic agents in this space remains underexplored and under-theorized.

2.2 *Dialogue on Ethics (DoE) Framework*

The Dialogue on Ethics (DoE) framework, as proposed by Cedar et al. (2021), offers a structured approach to analysing the ethical reasoning that emerges in student dialogue. The framework is grounded in micro genetic analysis of student talk, providing a fine-grained method to identify moments of ethical thinking within naturalistic dialogue.

DoE categorizes ethical engagement into three primary aspects:

- Conceptualization – when a value is explicitly or implicitly defined or discussed
- Interpretation – when a value is attributed to others, either ascriptions of motivation or situational analysis
- Judgement – when the speaker takes a personal moral position, often marked by evaluative language or commitments to action.

The strength of the DoE framework lies in its capacity to trace how ethical understanding develops within and across episodes of dialogue, rather than evaluating morality through externally imposed standards.

In the present study, the DoE framework was adapted to analyse student–chatbot interactions. Although the interaction partner is artificial, the structure of dialogue still allows for identification of ethical reasoning patterns. Each conversational turn was treated as a potential site of ethical engagement, and the three DoE dimensions were used to code and trace the development of ethical thinking.

3. Course Design and Implementation

3.1 *Course Structure*

The Personal Code of Ethics (PCOE) course was a 20-hour classroom-based program designed for Grade 5–6 learners (ages 10–11) to construct, test, and refine their personal ethical frameworks using genAI-enabled reflective tools. The course followed a scaffolded learning path, beginning with the real-world implications of ethics in an AI-driven world, progressing through the conceptual exploration of ethics, historical analysis, and applied moral dilemmas, and culminating in the creation of student-authored "digital twin" chatbots that embodied their chosen values. The capstone phase centered on a project where students designed a digital twin based on their PCOE for a selected value, created in Class 10, and built into a chatbot in Class 15. The chatbot acted as a reflective tool, allowing students to test and revise their ethical reasoning in response to new dilemmas, culminating in a reflective dialogue about their values. This paper focuses specifically on analyzing the interactions from Class 15, emphasizing the chatbot's role in facilitating ethical reflection and value articulation.

3.2 *Course Design Principles*

The course was explicitly built on constructivist foundations. Students were not given a fixed ethical doctrine; rather, they were guided to build meaning through exploration, discussion, and creation. For example, students developed their own value frameworks through mind maps, story-based analyses, and real-life reflection exercises—constructing a personalized "moral GPS" rather than adopting pre-given rules (Shah, 2019).

The approach was further rooted in dialogic pedagogies, drawing on traditions of ethical discussion, mutual respect, and value clarification. Classes incorporated structured group conversations, Socratic questioning, and peer feedback. Students were encouraged to confront contradictions, consider counterpoints, and revise their views, mirroring the iterative, social nature of ethical development (Alexander, 2005; Wegerif, 2013).

Significantly, the course departed from traditional moral education in its focus on everyday ethics. Instead of abstract moral dilemmas detached from students' lives, the curriculum prioritized interpersonal and culturally situated situations. Scenarios such as loyalty to a friend who has committed a mistake, or family practices conflicting with personal values, made ethics tangible and relatable. This emphasis on personalization also aligned with the emerging consensus in values education that encourages learner agency, relevance, and application (UNESCO, 2021).

3.3 *Technical Implementation*

To realize the pedagogical vision, a custom GenAI chatbot system was implemented, designed with the principle of value mirroring (Wegerif, 2013), the ability to reflect and respond according to student-authored ethical blueprints. The system's architecture used lightweight prompt-engineered LLMs (OpenAI's GPT-3 model) connected to a student interface that allowed learners to input, test, and revise their chatbot personas. The link to the chatbot is provided here for reference.

The value mirroring functionality was designed to constrain chatbot outputs to the student's ethical framework. For example, if a student prioritized honesty and empathy, the digital twin would be prompted to respond in dilemmas using only those values. The system avoided general ethical reasoning unless specified in the student's prompt, preserving the authenticity of student-authored ethics.

To support ethical engagement, dialogue structuring was built into the system: every chatbot interaction followed a turn-taking model prompting students to first input a dilemma, then receive a mirrored response, and finally reflect on their own feelings and possible divergence from the chatbot's decision. This enabled an internal dialogic loop, what we call "mirror as pedagogy" (Wegerif, 2013), where students encountered their values embodied

externally and reflected on alignment. The system thus served as both mirror and scaffold, a pedagogical agent that responds, provokes, and reflects alongside the learner.

4. Research Methodology

This is an exploratory study to investigate how Grade 5 students engage with values through AI-mediated dialogues using the Dialogue on Ethics (DoE) framework (Cedar et al, 2021). The primary research question (RQ) is- *How do the students engage with values using the genAI chatbot?* This study used a qualitative research approach to understand how Grade 5 students engage with everyday ethics while interacting with a genAI chatbot.

4.1 *Participants*

The study involved 23 Grade 5 students (ages 10–11) from an urban English-medium school, with 10 females and 13 males. The students come from an economically weaker section of the society with the parents' yearly income less than 7000 USD. Participants were selected through convenience sampling (Creswell & Creswell, 2017) from grade 5 class participating in a broader Skills Lab program on values and ethics. While students had no prior structured instruction in ethics, they navigated value-laden scenarios informally in daily life. Both parental consent and child assent were secured in accordance with ethical research protocols involving minors.

4.2 *Instruments and Data Collection*

The primary data source was chat logs generated from structured student–chatbot interactions. Students interacted with GenAI bots customized to reflect their own value systems, developed in earlier course sessions. Chatbots were hosted in a closed, supervised environment, and interactions were logged digitally. In total, students engage in a class session (~120 minutes) every week. To address the above mentioned RQ, this particular study focuses on the analysis of class 15 with an approximate duration of 60 minutes.

4.3 *Data Analysis*

For this study, chat logs constituted the data source, and were analysed using the Dialogue on Ethics (DoE) (Cedar et al, 2021) framework following the principles of content analysis and deductive coding. The analysis was majorly informed by the deductive codes: *Conceptualization*, *Interpretation*, *Judgement*, and any chat-based interactions other than these three codes were grouped under- *Other*. Three researchers (the first three authors of this paper) collaboratively coded the transcripts in iterative rounds. The first round of coding resulted in a percentage agreement of 45%, with discussions on how the definitions from DoE framework (Cedar et al, 2021) can be refined as per the study context (see Table 1). For example, the researchers arrived at a consensus to code for chatbot responses along with student responses where the chatbot asks or answers more than the predefined prompt. The coding cycles along with refinement of definitions continued with discussions on disagreements, and resolved through consensus resulting in a second-round percentage agreement of 70%, followed by a third-round percentage agreement of 91%.

As mentioned above, both the student and chatbot responses were coded for analysis to capture the full scope of the ethical dialogue. Each turn of conversation, whether initiated by the student or the chatbot, was treated as a distinct unit of analysis. This approach was done to understand the process of ethical thinking through back-and-forth exchanges. However, we noted that some turns, particularly those written by students, were lengthy and contained multiple ideas. In such cases, it was possible for a single turn to include more than one dimension of ethical engagement (e.g., a *conceptualization* followed by a *judgement*), and we applied multiple codes accordingly. This ensured that the coding remained sensitive to the richness and complexity of the students' responses.

Table 1. *DoE Coding Framework (Cedar et al., 2021) with adapted definitions*

Code	Original Definition (DoE)	Adapted Definition (this study & context)	Example
Conceptualization	concerns explicit/implicit discussion of ethical concepts	Definition, explicit or implicit discussion of a value	*Forgiveness means being generous and choosing to let go of hurt when someone has wronged you.*
Interpretation	concerns attributing beliefs, desires, and intentions to the characters and reconstructing causality between events	'You' or character-based assignments, attributions to characters	*You are too generous, you make your friend a king,but he is too clever (and) takes whole leadership (and) kills you*
Judgement	involves a move from the narrative towards ethical personal positioning	'I' and a personal positioning	*I use (this) with (my) friend when he blames me but I forgive him.*
Other	-	Greetings & pleasantries	*Great, nice*
Other	-	Gibberish text which had no context and incomprehensible	*I don't drive to understand*

Following the coding process, a frequency analysis was initially performed to quantify the occurrence of each code across all dialogues separately. This helped determine the dominant modes of ethical engagement for the student as well as the chatbot. Additionally, for selected cases, sample sequences of student dialogue codes were visualized as graphs to illustrate the transitions between different types of ethical reasoning across time. These visual representations provided insight into how individual students moved through ethical concepts within a session. Frequency analysis highlights dominant ethical reasoning modes, while time series analysis uncovers the temporal evolution of moral reasoning over the course of an interaction. Together, these methods allow us to understand not only the content of student ethical reasoning but also the process by which values are refined and internalized through dialogue. For each student, the full sequence of coded turns was preserved, maintaining the order in which ethical aspects emerged during the dialogue.

5. Findings

Research Question (RQ): How do the students engage with values using the genAI chatbot?

5.1 *Frequency of 'DoE' Aspects for Student and Chatbot Interactions*

To answer the RQ, the analysis of student and chatbot responses were done using the Dialogue on Ethics (DoE) framework. The most frequently observed DoE aspect in the students' responses (figure 1) was Other, accounting for the majority of coded turns (173 out of 431). This category includes greetings, procedural language, and conversational exchanges that do not directly address ethical aspects of the framework. For example, some of the user responses such as" *HI, it can change on View point on at online, okay*". While these interactions set the stage for more meaningful dialogue, they do not contribute directly to the ethical reasoning process.

Following Other, the second most frequent code was Judgement, with 96 instances. This reflects students taking a personal moral stance during the conversation, where they

actively evaluated scenarios and made value-based decisions centering around themselves. For example, "*With friend(s) I (will) be honest…..and respectful I (will) keep secrets and (stay true) to my promise and support their feelings*". These judgements were often related to personal experiences or hypothetical situations presented by the chatbot, and represented students' abilities to apply their personal value frameworks to real-world scenarios. In this example, the student is applying the personal code of ethics framework on the chosen value 'responsibility'.

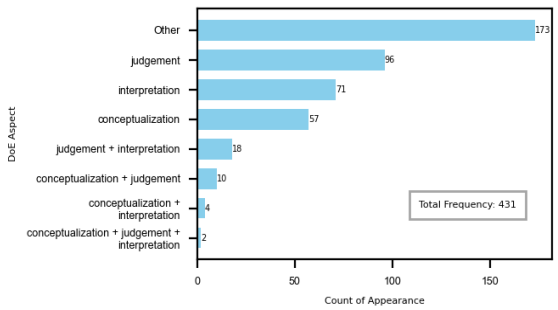


Figure 1. Frequency of 'DoE' Aspects for Student Interactions

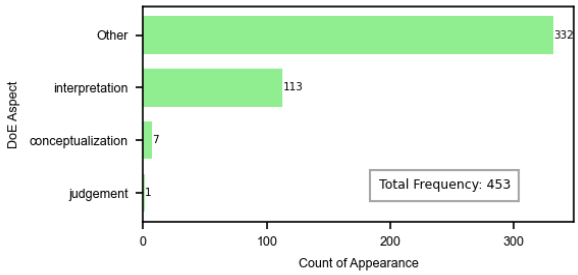


Figure 2. Frequency of 'DoE' Aspects for Chatbot Interactions

The third most frequent code was *Interpretation*, occurring 71 times. Students frequently engaged in interpreting ethical situations or attributions of values to characters, actions, or outcomes. For example, "*Your parents ask you to look after your younger sibling for an hour but your friends just invited you to play outside. Do you stay and watch your sibling like you were asked or do you leave without telling anyone*". This indicates that students often reflected on how values applied to others or how they were represented within the dialogue, without immediately asserting their own moral judgement.

Lastly, *Conceptualization* was the least frequent code, appearing 57 times. This suggests that while students did engage in defining or explaining values, much of the dialogue centered around applying and interpreting these values, rather than discussing or debating their theoretical meaning. For example, "*Generosity is not just about sharing material things; it's also about offering time, kindness, and support*". The relatively low frequency of *Conceptualization* points to a possible preference for action-oriented and contextual engagement with values over abstract discussions of their definitions.

Apart from the DoE codes, since the coding was performed for every turn of the interaction, some of the interactions were long (about 2-3 sentences). In a large chunk of dialogue, we often saw that the turns began with a particular code and then went on to another code. This is represented as code 1 + code 2. For example, *conceptualisation + judgement* code refers to this example, "*Online means using internet safely and respectfully thinking before posting protecting privacy and avoiding harmful behaviour example I don't share personal information avoid spreading rumours and report or block hurtful content instead of join again I also use the time wisely and don't let online distraction get in the way of the real-world duties.*"

The combination of codes demonstrates the complexity of ethical reasoning in longer dialogues. These combinations reflect the nuanced way in which students engage with values. Notably, the *judgement + interpretation* code appeared 18 times, indicating moments where students first made a personal moral judgment and then interpreted an ethical situation. This sequence demonstrates the students' ability to apply their own values and evaluate actions. The *conceptualization + judgement* code occurred 10 times, suggesting instances where students not only conceptualized ethics but also expressed a personal decision. A smaller number of interactions involved *conceptualization + interpretation* (4 occurrences) and *conceptualization + judgement + interpretation* (2 occurrences), pointing to rare instances where students simultaneously engaged in defining, interpreting, and judging ethical situations. These more complex combinations reflect the interconnected nature of moral reasoning, where students draw upon multiple ethical dimensions in their decision-making

processes.

Similar to the student interactions, the chatbot's responses were categorized using the Dialogue on Ethics (DoE) framework into *Conceptualization*, *Interpretation*, *Judgement*, and *Other* categories. The *Other* category again accounted for the highest frequency of responses, with 330 occurrences. This category primarily includes procedural language, technical instructions, and general conversational prompts by the chatbot, ensuring that the dialogue flowed smoothly without directly engaging in ethical reasoning. For example," Thank you for providing how someone would act responsibly online! Now, please share how your understanding of responsibility might change as you grow."

Figure 2 presents the frequency distribution of DoE aspects in chatbot interactions. *Interpretation*, with 113 occurrences, was the second most common, reflecting the chatbot's role in encouraging students to think about and interpret ethical dilemmas, often prompting them to reflect on how a value might apply in different contexts. For example," *Sure! Here's a scenario: You are working on a group project at school. One of your teammates isn't contributing and relies on you to do their part. What would a responsible person do in this situation according to your code?*"

Conceptualization, which involves the explicit or implicit definition of values, appeared only 7 times, suggesting that while the chatbot occasionally introduced definitions or contextual explanations of values, its primary role was to prompt ethical reasoning and decision-making rather than to provide fixed definitions. For example, "*PCOE stands for Personal Code of Ethics. It is a guide for how you want to act based on your values and beliefs. It helps you make choices that align with what is important to you. Now, please tell me about your chosen value. What is it?*"

The chatbot's dominant role in *Interpretation* aligns with its function as a reflective agent in the dialogue. The chatbot was designed to mirror the student's value framework, promoting reflection without offering its own moral judgments. By encouraging students to interpret ethical situations and make judgments based on their personal values, the chatbot facilitated an interactive learning experience in which the student's own ethical reasoning was actively tested, refined, and articulated.

5.2 *Case study of sample students' sequence analysis*

While frequency analysis provides an overview of common trends across all students, individual case studies allow for a more nuanced exploration of how ethical reasoning evolves in sample learners. This is critical for answering RQ, which seeks to explore how students engage with values using the chatbot. By closely analyzing sample sequences of DoE codes in a student's dialogue, we have attempted to uncover the processes that lead to ethical thinking.

5.2.1 *Sample Student 1 (S5)*

Student S5's interaction on the chosen value 'generosity' over 10 turns reveals a progression of ethical reasoning, illustrated by the line graph in Figure 3. The graph maps the frequency and sequence of different DoE aspects in the student's dialogue with the chatbot. Below is an analysis of the key stages and transitions observed in S5's engagement. The sequence shows a progression from initial conceptualization of values, through judgement and interpretation, followed by a back-and-forth engagement between interpretation and judgement. The analysis reveals several distinct phases in the student's moral reasoning, which are discussed below.

Initial Engagement (Turn 1–3): At the beginning of the interaction, the Other code is dominant, reflecting the procedural or introductory exchanges between the student and the chatbot. These early turns, which make up the majority of the dialogue initially, included greetings, setup prompts, or neutral questions. Once the student engages with the ethical content, the Conceptualization code appears. This phase is marked by the student's articulation of values or definitions. For example, "*Generosity, to me, means giving freely and*

selflessly to help others or bring joy, without expecting anything in return."

Moral Positioning - Judgement (Turns 4–6): As the conversation progresses, the student begins to engage more actively with the ethical dilemma, which is marked by the appearance of Judgement. The student moves from simply defining values to making personal ethical decisions. For example, "*At school I will share my supplies. And help my classmates. With friends I will help them when they need. And celebrate their achievements At home I will help my mom in doing the house chores. And share my things with my siblings*". This phase is critical because it demonstrates a personal commitment to a value. The student is not just reflecting on a value conceptually but is applying it to a real or hypothetical scenario.

Interpretation and Reflection - Interpretation (Turns 7–9): After establishing a judgement, the student begins to engage more deeply with the scenario by interpreting the situation in relation to the value. For example, "*Share Knowledge: Offer helpful advice, answer questions, or share useful resources without expecting anything in return. For example, contributing to forums or discussions to help others learn or solve problems*." This phase shows how students rethink or adjust their ethical decisions by considering the broader implications or complexities of a dilemma. The Interpretation phase appears more frequently in the latter half of the conversation, indicating that the student reflects on and revisits their earlier judgements.

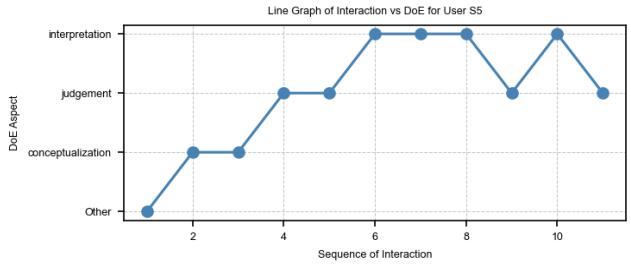


Figure 3. Time series analysis of a student's (S5) engagement

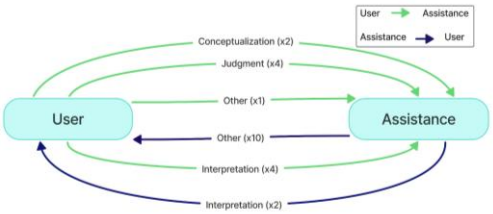


Figure 4. Interaction pattern between the student S5 (user) and genAI chatbot (assistance)

Back-and-Forth between Judgement and Interpretation (Turns 10+): The final phase of the interaction reveals a dynamic back-and-forth between Interpretation and Judgement, where the student revises or solidifies their moral reasoning through reflection. This phase is indicative of ongoing moral development, where the student continuously adjusts their stance, weighs alternative perspectives, and integrates emotional and rational responses into their ethical decision-making. The interaction pattern of the student (user) and genAI chatbot (assistant) is provided below in figure 4. This sequence reveals a dialogue pattern where the student progresses through various aspects of ethical thinking while the AI assistant primarily maintains a facilitative role.

5.2.2 *Sample Student 2 (S22)*

The student's interactions on the chosen value 'resilience' moves through phases of defining values, making moral decisions, reflecting on those decisions, and re-assessing them in light of the chatbot's responses. The time series graph (Figure 5) illustrates how the student takes 19 turns and shifts between Conceptualization, Judgement, Interpretation, and Other throughout the conversation. The dialogue follows a non-linear pattern.

Initial Conceptualization and Judgement (Turns 1–4): The student begins with Conceptualization, stating their understanding of a value. For example, "*Resilience is adopting Difficult Situation and face the struggle and win or try till he can*" This is followed by Judgement, where the student makes a moral decision about how they would act in a given scenario "*At school: In teamwork I will face the problem and will Adopt the difficult situation*".

Revisiting Conceptualization and Transition to Interpretation (Turns 5–8): The student re-engages with Conceptualization (Turn 5), refining or elaborating on their earlier

understanding of the value. For example, “sharing personal struggles, offering empathetic listening, and celebrating each other's achievements.” This is followed by a shift to Interpretation (Turn 8), where the student begins to apply the value to a more specific ethical situation or scenario. For example, *“My family members are facing problems in time management. What will you do*”.

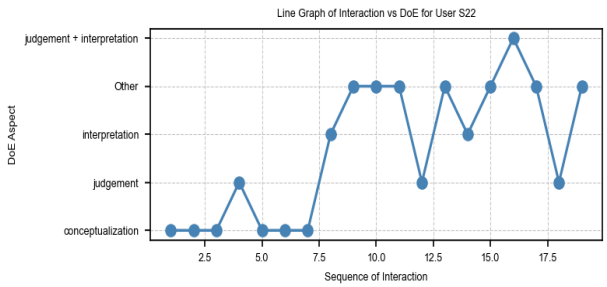


Figure 5. Time series analysis of a student’s (S22) engagement

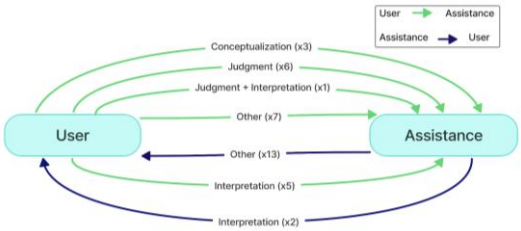


Figure 6. Interaction pattern between the student S22 (user) and genAI chatbot (assistance)

Intermittent Interactions with 'Other' and Reflection (Turns 9-11): The sequence moves into a phase marked by the Other code, reflecting procedural or contextual exchanges rather than ethical reasoning. This may include prompts from the chatbot that lead to new scenarios or brief responses to prior statements. This also marks a phase where the chatbot directs the students towards certain expected phases such as judgement & interpretation.

Iteration of Judgement and interpretation (12+): The student’s interaction transitions back into *Judgement* (Turn 12), where they reaffirm or reconsider their earlier stance. For example, “*I will not tell them to adopt, I will give strategies I will adopt when my group is struggling or in a difficult situation*”. This is followed by a return to *Interpretation* (Turn 14), where the student applies their moral stance to a new situation. For example, “*My group was stuck on a Slide. What will you do?*” This back-and-forth movement reveals the fluidity of ethical reasoning in the dialogue, where the student both asserts and reflects on their values in a dynamic, ongoing process. Additionally, as we see a combination code of interpretation + judgement we see this process to be complex and iterative.

An interesting feature of Student S22’s dialogue (Figure 6) is the frequent engagement with *Interpretation*, particularly in later turns. *Interpretation* allows the student to rethink and reframe their moral decisions in light of the chatbot’s reflective prompts, indicating a deeper level of ethical engagement. This aligns with the pedagogical goal of using the chatbot as a reflective agent, guiding students to engage in self-reflection and critical thinking about their values. Notably, the student’s dialogue alternates between *Judgement* and *Interpretation*, demonstrating the iterative process of refining ethical decisions through interaction with the chatbot.

6. Discussion and Conclusion

The findings of this study offer valuable insights into how students engage with values through GenAI-mediated dialogue, emphasizing the potential of technology in ethics education. The high frequency of Interpretation and Judgement codes in student responses reveals that students were deeply engaged in applying and evaluating values in real-world scenarios, rather than merely defining them. This shows that students moved beyond theoretical understanding to practical ethical decision-making. The chatbot’s role as a reflective agent is evident through its frequent use of Interpretation to prompt students to critically examine their decisions and consider multiple perspectives, encouraging a deeper level of reflection. While the Other codes were largely procedural, they highlight the chatbot’s role in maintaining an ongoing dialogue and facilitating the learning process.

The findings of this study extend existing theories of dialogic ethics education by demonstrating that AI, when designed as a reflective agent, can meaningfully support moral reasoning in young learners. Consistent with Wegerif’s (2013) and Mercer et al.’s (2019) view

that ethical development is socially constructed through dialogue, our results show that generative AI can facilitate Interpretation and Judgement, two markers of deeper ethical engagement, by prompting learners to reflect on value-laden scenarios. This supports Brandel et al.'s (2024) suggestion that dialogic exchanges, even with non-human agents, can serve as legitimate contexts for ethical growth. Furthermore, the chatbot's ability to mirror student-authored ethical frameworks offers a scalable pedagogical mechanism that aligns with and extends Cedar et al.'s (2021) Dialogue on Ethics framework, suggesting that dialogic AI agents can serve not only as facilitators but also as ethical mirrors in value-sensitive learning environments.

The study's findings, while promising, are limited by the small and homogeneous sample, potential interpretative bias, and the constraints of a single-session design with pre-programmed chatbot responses. Future research across diverse contexts and longitudinal studies will be essential, as advancing generative AI capabilities hold considerable promise for transforming values-based education into more interactive and personalized learning experiences.

References

Alexander, R. (2006). *Towards dialogic teaching: Rethinking classroom talk* (3rd ed.). Dialogos.

Berkowitz, M. W., & Bier, M. C. (2005). What works in character education: A research-driven guide for educators.

Bers, M. U. (2018). Coding as a playground: Programming and computational thinking in the early childhood classroom. *Routledge*.

Brandel, N., Schwarz, B. B., Cedar, T., Baker, M. J., Bietti, L. M., Pallarès, G., & Détienne, F. (2024). Dialogue on ethics and ethics of dialogue: An exploratory study. *European Journal of Psychology of Education, 39*, 2619–2654. https://doi.org/10.1007/s10212-024-00856-z

Cedar, T., Baker, M. J., Bietti, L. M., Détienne, F., Nir, E., Pallarès, G., & Schwarz, B. B. (2021). Dialogue on ethics, ethics of dialogue: Microgenetic analysis of students' moral thinking. In F. Maine & M. Vrikki (Eds.), *Dialogue for intercultural understanding* (pp. 103–117). Springer. https://doi.org/10.1007/978-3-030-71778-0_8

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

Hollingsworth, L. A., Didelot, M. J., & Smith, J. O. (2003). REACH beyond tolerance: A framework for teaching children empathy and responsibility. *Journal of Humanistic Counseling, Education and Development, 42*, 139–151.

Kohlberg, L. (1984). *Essays on Moral Development, Volume 2: The Psychology of Moral Development*. Harper & Row.

Mercer, N., Wegerif, R., & Major, L. (Eds.). (2019). *The Routledge International Handbook of Research on Dialogic Education*. Routledge.

National Education Policy (NEP) 2020. (2020). Ministry of Education, Government of India. Retrieved from: https://www.education.gov.in/sites/upload_files/mhrd/files/NEP_Final_English_0.pdf

OECD. (2019). *The future of education and skills: Education 2030*. Organisation for Economic Co-operation and Development. Retrieved from: https://www.oecd.org/

Patry, J. L., Weyringer, S., & Weinberger, A. (2008). Interaction of science and values in schools: VaKE—A method to nurture moral sensibilities. In K. Tirri (Ed.), *Educating moral sensibilities in urban schools* (pp. 157–170). Sense Publishers.

Shah, R. K. (2019). Effective Constructivist Teaching Learning in the Classroom. *Shanlax International Journal of Education*, 7(4), 1–13.

Shaw, T., & Brindle, M. (2020). Ethics education for a digital society. *Philosophy & Technology, 33*(2), 289–313.

UNESCO. (2015). *Global citizenship education: Preparing learners for the challenges of the 21st century*. United Nations Educational, Scientific and Cultural Organization. Retrieved from: https://unesdoc.unesco.org/

Wegerif, R. (2013). *Dialogic: Education for the Internet Age*. Routledge.

World Bank. (2018). *The World Development Report 2018: Learning to realize education's promise*. World Bank Group. Retrieved from: https://www.worldbank.org/

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Dialogic Expertise in Practice: How Expert Teachers Orchestrate Classroom Talk in Distinct Mathematics Lesson Types

Siyu WANG[a], Yaqian ZHENG[a], Mingze SUN[a], Yaping XU[b] & Yanyan LI[a*]
[a]*School of Educational Technology, Beijing Normal University, China*
[b]*School of Educational Technology, Northwest Normal University, China*
*liyy@bnu.edu.cn

Abstract: Dialogic teaching, which emphasizes the strategic use of classroom talk to support student thinking, has been widely recognized for its role in promoting deep learning. While previous research has affirmed the benefits of dialogic interaction, little is known about how expert teachers organize and adapt their discourse across different lesson types. This study addresses this gap by examining the distributional and sequential patterns of discourse moves used by expert mathematics teachers in Chinese middle schools during New lessons and Exercise lessons. Drawing on 40 video-recorded lessons, the study combined statistical analysis and process mining based on first-order Markov modeling to identify discourse structures and interaction flows. Findings reveal consistent discourse routines across lesson types, characterized by structured discourse organization and progressive scaffolding strategies that support coherent dialogue and deep conceptual understanding. However, distinct differences emerged in discourse strategy selection, instructional organization, and interactional structuring, reflecting adaptations to the pedagogical goals of each lesson type. These insights enhance understanding of dialogic expertise and inform the design of teacher development initiatives that promote context-sensitive discourse strategies.

Keywords: Dialogic teaching, expert teachers, discourse moves, lesson types, process mining

1. Introduction

Effective classroom talk is essential for promoting student thinking, engagement, and conceptual understanding (Mercer, 2019). As interest in dialogic teaching continues to grow, increasing attention has been paid to how teachers structure discourse to support learning (Muhonen et al., 2024). While many studies have confirmed the value of dialogic interaction, fewer have examined how specific discourse strategies unfold in real classrooms or how these strategies vary under different instructional conditions.

Among the factors influencing classroom discourse, teacher expertise has been shown to play a decisive role (Derakhshan et al., 2024). Expert teachers are more likely to use questions, prompts, and feedback in ways that build coherence across interactions and sustain student reasoning (Yu & Huang, 2025). Empirical evidence suggests that such teachers engage students more effectively through carefully sequenced talk and responsive adaptation to emerging ideas (Wray & McDonald, 2025). At the same time, discourse patterns are shaped by instructional context, including subject matter, student age, and the structure of teaching tasks. Lesson type, in particular, introduces distinct communicative demands (Amodia-Bidakowska et al., 2023), yet little is known about how expert teachers adjust their talk between lessons designed to introduce new concepts and those focused on practice or review.

This study addresses that gap by examining the discourse of expert mathematics teachers in New and Exercise lessons. Drawing on classroom recordings from Chinese

middle schools, we analyze how discourse moves are distributed and how they are sequentially organized across lesson types. Using a combination of statistical analysis and process mining based on first-order Markov modeling, the study aims to uncover how expert teachers orchestrate classroom dialogue in response to differing pedagogical goals. The findings contribute to a deeper understanding of adaptive discourse practices and offer implications for teacher learning and dialogic pedagogy.

2. Literature Review

2.1 Dialogic Teaching and Discourse Moves

Dialogic teaching emphasizes the use of discourse to promote students' critical thinking and problem-solving abilities (Alexander, 2018; Zhang et al., 2025), thereby supporting knowledge construction and the development of higher-order reasoning skills (Mercer, 2019). Grounded in Vygotsky's sociocultural theory, dialogic pedagogy is founded on the premise that meaning is co-constructed through social interaction (Vygotsky et al., 1978), making classroom discourse central to the learning process (Wang et al., 2025). As the primary medium of classroom interaction, discourse externalizes cognitive processes (Webb et al., 2014), enabling students to articulate perspectives, engage in deep reasoning, and formulate critical responses through sustained dialogue (Resnick et al., 2015). Empirical research has shown that teacher-dominated monologic instruction often suppresses students' cognitive engagement, limiting their capacity for reasoning and reflection (Yu & Huang, 2025). In response, scholars advocate for the strategic use of discourse to scaffold students' thinking, promote dialogic exchange, and enhance the pedagogical impact of teacher talk (Muhonen et al., 2024).

To systematically analyze dialogic interaction, discourse moves have been widely adopted as the core unit of analysis (O'Connor & Michaels, 2019). Teacher discourse moves refer to intentional verbal actions used to elicit responses, sustain interaction, and guide classroom dialogue (Tao & Chen, 2024). Analyzing these moves provides insight into how discourse practices influence student learning. Various coding frameworks have been developed to support the quantitative analysis of discourse moves, including the Academically Productive Talk (APT) framework (Michaels & O'Connor, 2015) and the Scheme for Educational Dialogue Analysis (SEDA) (Hennessy et al., 2016). The APT framework classifies discourse moves into three dimensions: responsibility to the learning community, to knowledge, and to reasoning standards, which help teachers prompt students to articulate ideas and justify their thinking. SEDA offers a more fine-grained structure, comprising 33 discourse moves organized into eight categories, and provides a comprehensive tool for analyzing educational dialogue. Building on SEDA, Hennessy et al. (2020) introduced the Cambridge Dialogue Analysis Scheme (CDAS), which streamlines the coding structure to enhance feasibility for large-scale classroom discourse analysis (Amodia-Bidakowska et al., 2023). Designed for broad applicability across subjects and pedagogical contexts, CDAS aligns well with the analytical needs and disciplinary scope of the present study.

2.2 Teachers' Classroom Talk Across Experience Levels and Instructional Contexts

Empirical research has shown that well-structured teacher discourse can enhance students' engagement, metacognitive regulation, and conceptual understanding (Chen, 2020; Smit et al., 2023). Through dialogic interaction, students are supported in articulating their ideas, reasoning through problems, and building shared understanding (Adie et al., 2018). However, much of the existing research has concentrated on the outcomes of dialogic teaching, while offering limited insights into the structural features and adaptive strategies that make teacher discourse effective in diverse classroom settings.

Recent research highlights the role of pedagogical expertise in shaping teachers' discourse strategies. Expert teachers are commonly defined as those with stable instructional experience, strong professional judgment, and flexibility in adapting to student needs (Lachner et al., 2016). Empirical studies have demonstrated that such teachers more effectively orchestrate classroom discourse to support reasoning, engagement, and learning coherence. For example, Omland and Rødnes (2020) found that expert mathematics teachers consistently used open-ended questions and elaborative feedback to promote collaborative inquiry. Derakhshan et al. (2024) further observed that experienced language teachers' discourse was marked by greater simultaneity and immediacy, fostering stronger learner involvement than that of novice teachers. These findings indicate that dialogic expertise is both a marker of pedagogical proficiency and a key mechanism for scaffolding deep learning and sustaining instructional continuity.

Beyond teacher expertise, growing evidence suggests that classroom discourse is closely shaped by instructional context. Discourse practices differ across subject areas, educational stages, and curricular structures, reflecting how teachers align their talk with varying content demands and student needs. For example, Amodia-Bidakowska et al. (2023) found that in upper primary classrooms, reasoned dialogue appeared more frequently in mathematics, while both English and mathematics supported more elaborated contributions than science. Muhonen et al. (2024) found that secondary classrooms had longer dialogic episodes, while early primary grades had more high-quality teacher-led dialogues, with greater subject-specific variation in primary grades. Among contextual dimensions, the contrast between concept-introducing lessons and practice-oriented lessons has received relatively limited attention. Yu and Huang (2025) provided preliminary evidence that teacher discourse occurred more frequently in practice lessons than in new content lessons in Chinese mathematics classrooms. Nevertheless, systematic research on how discourse structures vary with lesson type, particularly in expert teaching contexts, remains scarce.

Despite increasing attention to classroom discourse, little is known about how expert teachers adapt their talk across lesson types or how their discourse is sequentially organized to support learning. To address these gaps, this study aims to investigate the distribution and sequential traits of discourse moves employed by expert mathematics teachers in lessons introducing new content and practice sessions, posing the following research questions:

(1) What are the differences in the distribution of discourse moves used by expert mathematics teachers in New and Exercise lessons?

(2) What are the differences in the sequential patterns of discourse moves used by expert mathematics teachers in New and Exercise lessons?

3. Methods

3.1 Data Sample

This study analyzed 40 video-recorded Grade 8 mathematics lessons sourced from the *Smart Education of China* platform (https://basic.smartedu.cn/), a national open-access repository for high-quality K-12 instructional resources developed by China's Ministry of Education. The sample included 20 New lessons and 20 Exercise lessons, all centered on the same curricular topic and based on a standardized textbook edition. All were delivered by expert teachers with intermediate or senior professional titles and officially recognized as national exemplary classes, each lasting 40-45 minutes to ensure consistency.

3.2 Coding Scheme

This study adapted the Cambridge Discourse Analysis Scheme (CDAS; Hennessy et al., 2020) to code classroom discourse moves. Two pairs of categories were merged for clarity: Simple Coordination and Reasoned Coordination into Coordination (CO), and Reference to Wider Context and Reference Back into Reference Beyond (RB), as shown in Table 1. To

differentiate speaker roles, all codes were prefixed with “T.” for teachers and “S.” for students (e.g., T.COI, S.CO).

Table 1. *The Coding Scheme for Discourse Moves*

Code	Definition	Example
Elaboration invitations (ELI)	Encourages individuals to supplement, elaborate on, evaluate, or clarify their own or others’ perspectives.	*Who can further contribute by identifying other types of special quadrilaterals?*
Elaboration (EL)	Builds on, elaborates, evaluates, or clarifies one’s own or others’ perspectives.	*In other words, a quadrilateral with diagonals that bisect each other is a parallelogram.*
Reasoning invitations (REI)	Invites others to explain or argue a viewpoint, or guide them to make inferences, predictions, or hypotheses.	*Why is segment DF equal to segment BE? Can someone explain the reasoning behind this?*
Reasoning (RE)	Provides an explanation or justification for perspectives, or engages in speculation, prediction, or hypothesis based on reasoning.	*Since AB equals CD and AE equals CF in parallelogram ABCD, it follows that DF equals BE.*
Co-ordination invitations (COI)	Invites the synthesis, summarization, comparison, evaluation, or resolution of multiple perspectives.	*Based on these figures, who can formulate a definition of line-symmetric figures?*
Co-ordination (CO)	Organizes and analyzes multiple perspectives or ideas, synthesizing collective viewpoints, comparing differing opinions, and reasoning based on theory or evidence.	*Through these exercises, we find that a plane figure, when folded along a straight line, exhibits coincident portions on either side of the line. Such a figure is referred to as a line-symmetric figure.*
Agreement (A)	Explicitly express acceptance or endorsement of a particular statement or contribution.	*Excellent, everyone has understood the definition of rational expressions.*
Querying (Q)	Express doubt, disagreement, or challenge a viewpoint.	*Are you sure it’s a rhombus? Check its defining characteristics.*
Reference beyond (RB)	Guides connections to prior knowledge, experiences, or views, and to relevant knowledge beyond the lesson topic.	*Can everyone recall the method we learned for finding the least common denominator of fractions?*
Other invitations (OI)	Invites various verbal contributions (e.g., expressions of opinions, ideas), excluding ELI, REI, COI, or RB.	*He said earlier that the diagonals of a parallelogram are opposite. Do you agree with this statement?*
Uncoded (UC)	Discourse not fitting any of the previously defined categories.	*That’s the end of our short but joyful learning today.*

3.3 Data Processing

The 40 recorded Grade 8 mathematics lessons were transcribed using the iFlyRec automatic speech recognition tool and manually reviewed for accuracy. The transcripts were segmented into 10,876 teacher discourse utterances based on meaningful instructional units. Prior to formal coding, two trained researchers jointly reviewed the coding scheme and conducted calibration to ensure a shared understanding of category definitions. A random sample of 1,200 utterances was then independently coded, yielding a Cohen’s Kappa of 0.80, indicating high inter-rater reliability. Discrepancies were resolved through discussion, and the remaining 9,676 utterances were subsequently coded by the same researchers.

3.4 Data Analysis

This study examined both the distributional and sequential characteristics of expert teachers' discourse in New and Exercise lessons. Descriptive statistics were used to calculate the frequency and distribution of discourse moves, and the Mann-Whitney U test was applied to assess the significance of differences across lesson types. For sequential analysis, process mining was conducted using first-order Markov modeling (FOMM) (Fan et al., 2022), generating transition variance diagrams to visualize and compare discourse flow patterns between the two lesson types.

4. Results

4.1 Distributional Characteristics of Expert Teachers' Discourse Moves in New and Exercise Lessons

Table 2 presents the frequency distribution and significance test results of discourse moves used by expert mathematics teachers across New and Exercise lessons. Overall, OI emerges as the most frequently used move in both lesson types, accounting for 24.73% in New lessons and 25.47% in Exercise lessons. This is followed by ELI (14.38% in New lessons; 18.73% in Exercise lessons), while A ranks third (8.08% and 5.76%, respectively). Notably, the frequencies of RE and Q remain low in both contexts, with RE accounting for only 0.76% in New lessons and 0.48% in Exercise lessons, and Q occurring at 0.54% and 0.69%, respectively.

The results of the Mann-Whitney U test reveal significant differences in the use of discourse moves between New and Exercise lessons. Expert teachers employed REI significantly more frequently in New lessons (U=305.0, z=2.840, p=0.005, $|r|$=0.449), suggesting a stronger emphasis on encouraging student reasoning. The use of COI (U=317.0, z=3.165, p=0.002, $|r|$=0.500) and CO (U=288.0, z=2.380, p=0.017, $|r|$=0.376) was also significantly higher, reflecting more frequent efforts to guide students in synthesizing ideas. In addition, A occurred more often in New lessons (U=303.5, z=2.800, p=0.005, $|r|$=0.443), indicating a greater tendency to affirm and positively reinforce student contributions.

Table 2. *Frequency, Proportional Distribution, and Statistical Comparison of Discourse Moves Used by Expert Mathematics Teachers in New and Exercise Lessons*

Code	New lessons		Exercise lessons		Mann-Whitney U Test		
	F	%	F	%	U	z	Asymp. Sig. (2-tailed)
ELI	1114	14.38%	1375	18.73%	163.0	-1.001	0.323
EL	298	3.85%	357	4.86%	156.0	-1.190	0.239
REI	447	5.77%	253	3.45%	305.0	2.840	**0.005***
RE	59	0.76%	35	0.48%	254.0	1.461	0.140
COI	160	2.07%	78	1.06%	317.0	3.165	**0.002***
CO	119	1.54%	74	1.01%	288.0	2.380	**0.017***
RB	156	2.01%	127	1.73%	270.5	1.907	0.057
A	626	8.08%	423	5.76%	303.5	2.800	**0.005***
Q	42	0.54%	51	0.69%	212.5	0.338	0.739
OI	1916	24.73%	1870	25.47%	227.0	0.730	0.473
UC	552	7.13%	744	10.13%	172.5	-0.744	0.465

Note. U = Mann-Whitney U statistic; z = standardized test statistic; r = effect size. *Asymp. Sig. (2-tailed)* refers to the asymptotic significance level from a two-tailed test. $p < 0.05$ indicates statistical significance.

4.2 Sequential Patterns of Expert Teachers' Discourse Moves in New and Exercise Lessons

To examine how expert teachers adapt their discourse patterns across instructional contexts, this study modeled classroom discourse move sequences in New and Exercise lessons using

first-order Markov chains and the process mining toolkit *pMineR* (Fan et al., 2025). The resulting process maps (Figures 1 and 2) visualize the sequential flow of interaction, where each node represents a specific teacher or student move, and each connecting line indicates the transition probability (TP) between moves. To enhance interpretability, only transitions with TP>0.2 were retained, emphasizing core discourse pathways. Transition probabilities in New and Exercise lessons are denoted as TP_n and TP_e, respectively. To capture structural contrasts, a transition difference map was also generated (Figure 3), with a threshold of 0.15 applied to highlight only meaningful differences in usage. In this map, green lines indicate transitions more common in New lessons, red lines those more frequent in Exercise lessons, and black lines reflect comparable usage across both. Together, these visualizations reveal how expert teachers flexibly orchestrate discourse in response to lesson-specific goals and task structures.

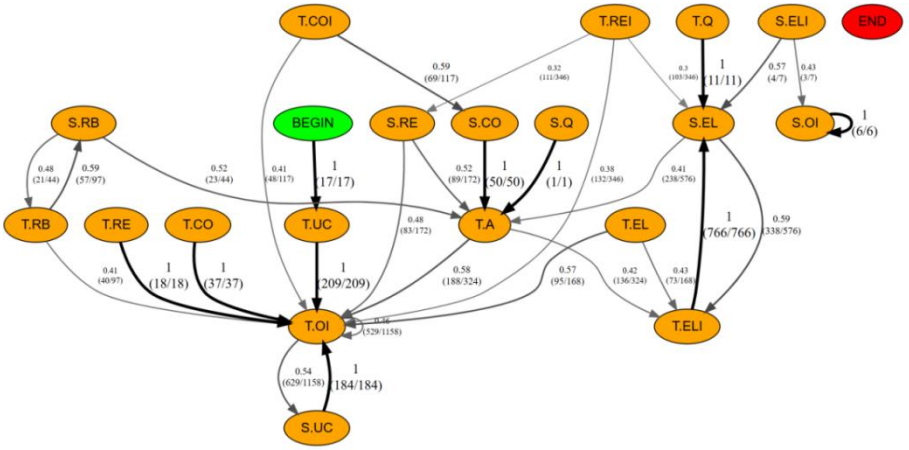


Figure 1. Process Map of Discourse Move Sequences in New Lessons.

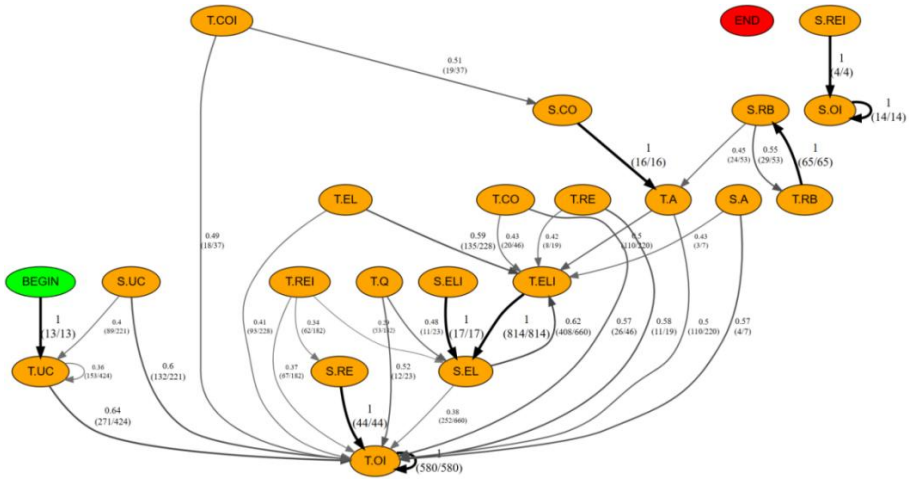


Figure 2. Process Map of Discourse Move Sequences in Exercise Lessons.

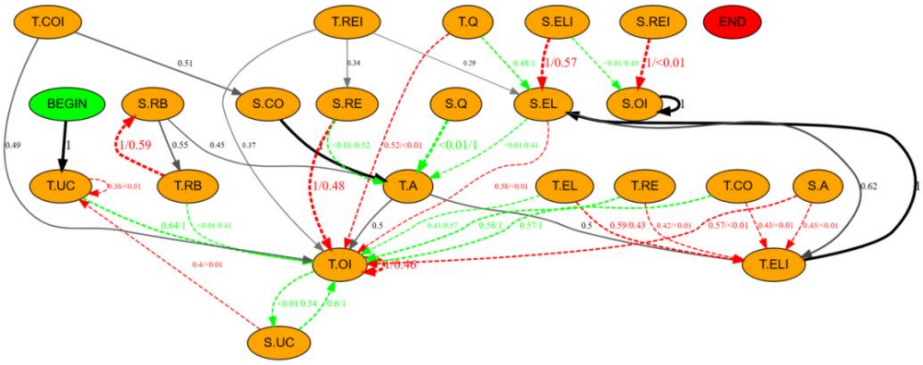


Figure 3. Transition Difference Map of Discourse Moves Between New and Exercise Lessons.

The process maps for both New and Exercise lessons reveal several stable discourse structures reflecting expert teachers' consistent dialogic practices. In both cases, the End node appears as an isolated terminal, indicating flexible, non-formulaic lesson closures. A key feature is the central role of T.OI, acting as the primary interaction hub, receiving transitions from eleven discourse moves in New lessons and thirteen in Exercise lessons. Its self-loop (TP_n=0.46, TP_e=1.00) highlights its function in maintaining interaction flow, particularly in procedurally driven segments. In both lesson types, a recurrent elaboration loop is observed between T.ELI and S.EL, with T.ELI consistently leading to S.EL (TP=1.00), followed by a return to T.ELI (TP_n=0.59, TP_e=0.62), forming a stable pattern of extended elaboration.

While these shared patterns provide a structural foundation for understanding expert discourse practices, clear differences emerge in how teachers adapt their discourse moves to the specific demands of New and Exercise lessons. In New lessons, the instructional opening is highly structured, with teachers consistently transitioning from conventional routines (T.UC) to general invitations (T.OI) with a transition probability of TP_n=1.00. This direct shift often features closed prompts that immediately foreground the lesson's conceptual focus. In contrast, Exercise lessons display more flexible entry patterns. While most begin with T.UC followed by T.OI (TP_e=0.64), a substantial portion (TP_e=0.36) continues with T.UC, during which teachers provide evaluative feedback on homework or learning progress to bridge into task-specific objectives.

The two lesson types also differ in secondary discourse hubs. In New lessons, T.A serves as a secondary node, receiving five transitions and emphasizing frequent affirmative teacher responses. In Exercise lessons, T.ELI takes this role, connecting to six discourse moves and reflecting greater reliance on elaboration invitations to guide interaction.

Foundational interaction patterns further distinguish the two contexts. In New lessons, T.OI frequently leads to brief student responses (S.UC, TP_n=0.54), which are consistently followed by a return to T.OI (TP_n=1.00), forming a recursive loop of confirmation. In contrast, Exercise lessons display a direct self-loop on T.OI (TP_e=1.00), suggesting a procedural use of teacher talk that maintains lesson flow without student response.

Differences are also observed in how teachers follow up on higher-order discourse moves. After T.CO, transitions in New lessons typically lead to T.OI (TP_n=1.00), reflecting efforts to check whole-class understanding. In Exercise lessons, the preferred path is from T.CO to T.ELI (TP_e=0.43), indicating a shift toward individual elaboration. A similar divergence appears after T.RE, which leads to T.OI in New lessons (TP_n=1.00) but to T.ELI in Exercise lessons (TP_e=0.42). Following T.EL, teachers in New lessons more often use T.OI (TP_n=0.57), whereas those in Exercise lessons return to T.ELI (TP_e=0.59). These patterns suggest that New lessons emphasize collective verification, while Exercise lessons tend to foster extended individual responses.

Further contrast is evident in how teachers utilize referential discourse. After T.RB, the transition to student reference (S.RB) occurs more frequently in Exercise lessons (TP_e=1.00) than in New lessons (TP_n=0.59), reflecting a greater emphasis on linking tasks to prior knowledge or strategy transfer in problem-solving contexts.

Teachers' responses to student contributions also differ significantly. In New lessons, teachers often affirm student input through T.A, particularly following S.CO (TP_n=1.00), S.Q (TP_n=1.00), S.RE (TP_n=0.52, TP_e<0.01), and S.EL (TP_n=0.41, TP_e<0.01). In Exercise lessons, however, teachers are more likely to extend the discourse through T.OI, such as in S.RE→T.OI (TP_n=0.48, TP_e=1.00), and S.EL→T.OI appears less frequently (TP_n<0.01, TP_e=0.38). This suggests a stronger emphasis on distributing participation and maintaining lesson rhythm through structured prompts. Lastly, the process maps show more frequent peer-to-peer transitions in Exercise lessons, suggesting a greater degree of student–student interaction in task-oriented contexts.

5. Discussion

Integrating frequency distribution and process-mining analysis, the findings demonstrate that expert teachers maintain consistent discourse practices across lesson types, characterized by shared patterns in structural organization and cognitive scaffolding. Structurally, their discourse reflects a high degree of organization, marked by the use of pre-planned questions and systematic procedural framing to sustain dialogic coherence and align interaction with instructional aims, supporting prior work on the regulatory role of structured discourse in classroom learning (Blatchford & Russell, 2019). Cognitively, expert teachers consistently adopt progressive elaboration strategies, using layered prompts to guide student thinking and build coherent mathematical understanding (Webb et al., 2019).

While expert teachers exhibit consistent discourse practices across lesson types, notable differences emerge in their use of discourse moves, particularly across three dimensions: discourse strategy selection, instructional organization, and interactional structuring.

Expert teachers' discourse strategies diverged notably across lesson types. In New lessons, teachers often responded to student contributions with immediate affirmation, followed by whole-class checks for understanding. This approach helped reduce learners' cognitive uncertainty when encountering new concepts and facilitated the establishment of clear learning expectations through consistent positive feedback and shared validation. From a cognitive load perspective, this pattern reflects a coordinated regulation of intrinsic and extraneous demands, supporting instructional efficiency and conceptual clarity (Zhang et al., 2016). In contrast, Exercise lessons featured a more extended and exploratory discourse pattern. Teachers frequently sustained interaction through elaboration-oriented prompts, encouraging students to extend their own or peers' ideas across multiple turns. These episodes were often interspersed with brief, low-intensity prompts that served as transitional space for cognitive processing. This structure aligns with the pedagogical goals of problem-solving lessons, which emphasize consolidation, application, and transfer. Prior studies have suggested that increasing cognitive demand and extending chains of thought in such contexts can promote deeper understanding and support knowledge transfer (Jou et al., 2016; Moskaliuk et al., 2012).

From the perspective of instructional organization, expert teachers demonstrated context-sensitive strategies that aligned with the goals of each lesson type. In New lessons, discourse was tightly structured and goal-oriented, with teachers frequently using closed questions to focus student attention and establish a coherent conceptual framework. In contrast, Exercise lessons featured more flexible and diagnostic structures. Lesson openings often began with feedback on homework performance, enabling teachers to assess learning needs and link prior knowledge with current tasks. Contextual cues were also used to support knowledge transfer and the internalization of problem-solving strategies. Moreover, teachers placed greater emphasis on expanding dialogic space, encouraging peer interaction and collaborative meaning-making. These practices fostered opportunities for autonomous knowledge reconstruction and dialogic reflection within a framework of consolidation, extension, and transfer (Jen et al., 2016; Zhang, 2018).

In terms of interactional structure, New lessons were characterized by streamlined and rule-governed exchanges, often revolving around confirm–respond–reconfirm sequences. These compact interaction cycles served to regulate classroom pacing and support comprehension monitoring through consistent checks for understanding. In contrast, Exercise lessons relied more heavily on consecutive procedural directives to drive task progression, with fewer contingent responses to individual student contributions. This shift in interactional dynamics reflects a pedagogical focus on efficient content delivery and initial understanding in New lessons, whereas Exercise lessons emphasized sustained task engagement and the practical application of knowledge (Di et al., 2019; Shadiev & Huang, 2020). These differences highlight expert teachers' ability to flexibly structure discourse in response to the functional goals and cognitive demands of each instructional context.

Building on our findings, the study suggests that teacher training should focus on helping novice teachers adapt their discourse strategies to lesson objectives, enhancing teaching flexibility. By examining expert teachers' strategies, this study highlights the contribution of dialogic expertise as an adaptive practice, showing how teachers adjust their

strategies to classroom organization and student needs to promote deep learning and engagement. Future research could expand the sample to include teachers with varying experience and subjects, exploring the diversity and adaptability of discourse strategies.

6. Conclusions

This study examined how expert mathematics teachers orchestrate classroom discourse across New and Exercise lessons by integrating frequency-based statistical analysis with process mining based on first-order Markov modeling. The findings reveal both consistent discourse routines, characterized by stable patterns in structural organization and cognitive scaffolding, and context-sensitive adaptations in strategy selection, instructional organization, and interactional structuring. These adaptations reflect expert teachers' ability to align discourse practices with specific pedagogical goals and cognitive demands. These insights deepen theoretical understanding of dialogic expertise and carry practical implications for enhancing teacher learning. Building on the findings of this study, future work could design teacher workshops as a vehicle for professional development, enabling teachers to adapt context-responsive discourse strategies across different lesson types to further enhance dialogic teaching skills. In addition, it would be valuable to explore the development of AI-driven classroom feedback systems that provide teachers with real-time, context-responsive guidance for adjusting discourse strategies during instruction.

Acknowledgements

This work has been supported by the National Science and Technology Major Project of China (Grant No: 2022ZD0115900) and the National Natural Science Foundation of China (Grant No: 62277006).

References

Adie, L., Van Der Kleij, F., & Cumming, J. (2018). The development and application of coding frameworks to explore dialogic feedback interactions and self‐regulated learning. *British Educational Research Journal*, *44*(4), 704–723.

Alexander, R. (2018). Developing dialogic teaching: Genesis, process, trial. *Research Papers in Education*, *33*(5), 561–598.

Amodia-Bidakowska, A., Hennessy, S., & Warwick, P. (2023). Disciplinary dialogues: Exploring the association between classroom dialogue and learning outcomes within and between subjects in English primary schools. *Learning, Culture and Social Interaction*, *43*, 100742.

Blatchford, P., & Russell, A. (2019). Class size, grouping practices and classroom management. *International Journal of Educational Research*, *96*, 154–163.

Chen, G. (2020). A visual learning analytics (VLA) approach to video-based teacher professional development: Impact on teachers' beliefs, self-efficacy, and classroom talk practice. *Computers & Education*, *144*, 103670.

Derakhshan, A., Karimpour, S., & Nazari, M. (2024). Interactional features in second language classroom discourse: Variations across novice and experienced language teachers. *Applied Linguistics Review*, *15*(6), 2323–2349.

Di, W., Danxia, X., & Chun, L. (2019). The effects of learner factors on higher-order thinking in the smart classroom environment. *Journal of Computers in Education*, *6*(4), 483–498.

Fan, Y., Jovanović, J., Saint, J., Jiang, Y., Wang, Q., & Gašević, D. (2022). Revealing the regulation of learning strategies of MOOC retakers: A learning analytic study. *Computers & Education*, *178*, 104404.

Hennessy, S., Howe, C., Mercer, N., & Vrikki, M. (2020). Coding classroom dialogue: Methodological considerations for researchers. *Learning, Culture and Social Interaction*, *25*, 100404.

Hennessy, S., Rojas-Drummond, S., Higham, R., Márquez, A. M., Maine, F., Ríos, R. M., García-Carrión, R., Torreblanca, O., & Barrera, M. J. (2016). Developing a coding scheme for

analysing classroom dialogue across educational contexts. *Learning, Culture and Social Interaction*, *9*, 16–44.
Jen, T.-H., Yeh, Y.-F., Hsu, Y.-S., Wu, H.-K., & Chen, K.-M. (2016). Science teachers' TPACK-Practical: Standard-setting using an evidence-based approach. *Computers & Education*, *95*, 45–62.
Jou, M., Lin, Y.-T., & Wu, D.-W. (2016). Effect of a blended learning environment on student critical thinking and knowledge transformation. *Interactive Learning Environments*, *24*(6), 1131–1147.
Lachner, A., Jarodzka, H., & Nückles, M. (2016). What makes an expert teacher? Investigating teachers' professional vision and discourse abilities. *Instructional Science*, *44*(3), 197–203.
Mercer, N. (2019). *Language and the joint creation of knowledge: The selected works of Neil Mercer* (First published). Routledge.
Michaels, S., & O'Connor, C. (2015). Conceptualizing Talk Moves as Tools: Professional Development Approaches for Academically Productive Discussions. *Socializing Intelligence Through Academic Talk and Dialogue* ,347–361.
Moskaliuk, J., Kimmerle, J., & Cress, U. (2012). Collaborative knowledge building with wikis: The impact of redundancy and polarity. *Computers & Education*, *58*(4), 1049–1057.
Muhonen, H., Pakarinen, E., Rasku-Puttonen, H., Poikkeus, A.-M., Siekkinen, M., & Lerkkanen, M.-K. (2024). Investigating educational dialogue: Variations of dialogue amount and quality among different subjects between early primary and secondary school classrooms. *Learning, Culture and Social Interaction*, *45*, 100799.
O'Connor, C., & Michaels, S. (2019). Supporting teachers in taking up productive talk moves: The long road to professional learning at scale. *International Journal of Educational Research*, *97*, 166–175.
Omland, M., & Rødnes, K. A. (2020). Building agency through technology-aided dialogic teaching. *Learning, Culture and Social Interaction*, *26*, 100406.
Resnick, L. B., Asterhan, C. S. C., & Clarke, S. N. (2015). *Socializing intelligence through academic talk and dialogue*. American Educational Research Association.
Shadiev, R., & Huang, Y.-M. (2020). Investigating student attention, meditation, cognitive load, and satisfaction during lectures in a foreign language supported by speech-enabled language translation. *Computer Assisted Language Learning*, *33*(3), 301–326.
Smit, R., Hess, K., Taras, A., Bachmann, P., & Dober, H. (2023). The role of interactive dialogue in students' learning of mathematical reasoning: A quantitative multi-method analysis of feedback episodes. *Learning and Instruction*, *86*, 101777.
Tao, Y., & Chen, G. (2024). The relationship between teacher talk and students' academic achievement: A meta-analysis. *Educational Research Review*, *45*, 100638.
Vygotsky, L. S., Cole, M., John-Steiner, V., Scribner, S., & Souberman, E. (1978). *The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
Wang, P., Yang, C., Tong, Y., & Chen, G. (2025). Discursive engagement for intersubjectivity: Investigating the verbal–visual relationship of productive classroom talk in synchronous online lessons. *Learning, Culture and Social Interaction*, *51*, 100892.
Webb, N. M., Franke, M. L., Ing, M., Turrou, A. C., Johnson, N. C., & Zimmerman, J. (2019). Teacher practices that promote productive dialogue and learning in mathematics classrooms. *International Journal of Educational Research*, *97*, 176–186.
Webb, N. M., Franke, M. L., Ing, M., Wong, J., Fernandez, C. H., Shin, N., & Turrou, A. C. (2014). Engaging with others' mathematical ideas: Interrelationships among student participation, teachers' instructional practices, and learning. *International Journal of Educational Research*, *63*, 79–93.
Wray, K. A., & McDonald, S. (2025). Teacher Talk Supporting Student Progressive Discourse in Science. *Science & Education*.
Yu, D., & Huang, Y. (2025). Characteristics of primary school mathematics classroom discourse for cultivating higher-order thinking: Analysis based on classroom videos of Chinese primary school expert teachers. *International Journal of Educational Research*, *131*, 102577.
Zhang, L. (2018). Withholding answers during hands-on scientific *investigations*? Comparing effects on developing students' scientific knowledge, reasoning, and application. *International Journal of Science Education*, *40*(4), 459–469.
Zhang, X., Zhang, X., & Yang, X. (2016). A Study of The Effects of Multimedia Dynamic Teaching on Cognitive Load and Learning Outcome. *EURASIA Journal of Mathematics, Science and Technology Education*, *12*(11).
Zhang, Y., Xin, J. J., Yu, Z., Liu, Y., Zhao, W., Li, N., Li, Y., & Chen, G. (2025). Enhancing Preservice Teachers' Use of Dialogic Teaching and Dynamic Visualizations in Mathematics Classes: Bridging the Knowing–Doing Gap. *International Journal of Science and Mathematics Education*.

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

A Deep-Learning Technique for Converting Bengali Handwritten Answer-Book Images into Text

Moumita MOITRA, Malay Kumar MAJHI & Sujan Kumar SAHA*
Department of CSE, National Institute of Technology Durgapur, India
*sujan.kr.saha@gmail.com

Abstract: Computer-Assisted Assessment (CAA) involves the use of computer technology for educational assessment. Handwritten answer books are the primary medium for various levels of educational assessment in schools in India. Handwritten answer books pose the key challenge in adopting CAA in the Indian language, Bengali. These answer books need to be converted into machine-readable text before applying any tool for evaluation of the answers. The literature on Bengali handwritten word recognition primarily focused on considering the image of a single word as input. However, for automatic answer evaluation, the image of the whole page should be taken as input by the system. In this paper, we aim to develop a system for converting handwritten words from a student's answer book into machine-readable text. The proposed system follows a two-phase architecture. The first phase automatically segments the words using Mask Region-based Convolutional Neural Network (R-CNN) network. Then, the segmented words are used as input for the second phase of image-to-text conversion using a CNN-Transformer network. To train the models, we use an openly available dataset and an in-house dataset comprised of answer books collected from Bengali medium schools. The experimental results show that the proposed model is quite promising.

Keywords: Automatic Answer Grading, Handwritten Character Recognition, Word Segmentation, Deep Learning model.

1. Introduction

The use of computer technology in educational assessment is a widely explored research problem. Several systems have been developed to effectively use computers in educational assessment. Due to various advantages over the traditional manual evaluation, computer-assisted assessment (CAA) has been utilized in various education sectors. Numerous studies have been conducted that demonstrates the effectiveness of CAA. However, due to the lack of necessary infrastructure, use of CAA is hindered and the paper-based examination and assessment is still being continued for educational assessment in India. Most of the existing CAA frameworks needs the examination to be conducted on a digital platform and then automatic evaluation is done. These frameworks cannot be applied directly to paper answer books. In order to adopt CAA here, the first step is to convert the paper answer books into machine readable text.

Handwritten Text Recognition (HTR) is the process of converting handwritten script images into machine-readable text. Automatic Recognition of Handwritten Characters has been a widely explored research area for decades. In the early days, the focus was to convert the images of handwritten characters or digits into corresponding machine-readable characters, and the area was named Handwritten Character Recognition (HCR). With the advancement of information and communication technologies, new application areas emerged, and it became necessary to recognize handwritten words and text. Ample research effort has been devoted to developing HTR systems in various languages (AlKendi et al., 2024; Teslya and Mohammed, 2022; Moitra and Saha, 2024).

Bengali is the sixth most widely spoken languages in the world. It is the second most widely spoken language in India and the official language of Bangladesh. Bengali is the

primary medium of education in a large number of Bengali medium schools in West Bengal and Bangladesh. Handwritten text plays the most crucial role in content delivery and assessment there. So, recognition of handwritten Bengali text is crucial for smart education and assessment. Recognition of Bengali handwriting is comparatively more difficult than English due to its complex shape, matra, and the presence of numerous conjunct characters. Several systems have been developed for recognizing Bengali characters (Bhowmik et al., 2016; Chaudhuri and Bera, 2009; Dutta et al., 2021; Keserwani et al., 2019). Most of these systems were built on handwritten document images written by adults, where handwriting is clearer, matured and more structured. On the other hand, the handwriting of school students is more inconsistent, inter-word and intra-word spacing is not uniform, and contains overlapping of characters. Therefore, recognition of school students handwriting is more challenging. We did not find any openly available system that takes an image of a page of a handwritten Bengali answer book and accurately converts it into the corresponding text.

This paper focuses on developing a system for converting handwritten words from a student's answer book into machine-readable text. A two-phase architecture is proposed for the task. The first phase performs *word segmentation*. It basically marks the word boundaries from the input image of a handwritten whole page containing several lines and words. The output of the first phase, segmented word images, is fed into the second phase, which recognizes the words. The word segmentation model is developed using Mask Region-based Convolutional Neural Network (Mask RCNN). The core component of the RCNN architecture contains a Region Proposal Network (RPN) to generate candidate regions likely containing words, and a Fully Convolutional Network (FCN) is used for pixel-to-pixel segmentation. The word recognition model consists of a Convolutional Neural Network (CNN) followed by a positional encoding layer and an encoder module of the Transformer. The output from the encoder is then finally fed into a linear layer, which produces the predictions.

For the training of the system, we collected answer books from middle school students from multiple Bengali medium schools and extracted and labeled words from them. We evaluated the performance of the individual phases and the whole system using an in-house test dataset comprising images of school students' handwritten pages. We created baseline models in both phases using standard techniques to compare the performance of the proposed systems. The proposed technique outperformed the baseline models in both phases. The system achieved an overall accuracy of 78.12%. To evaluate the potential of the proposed technique, we also tested the system using a dataset containing images of Bengali handwritten pages collected from online sources. Then the system obtained an accuracy of 93.62%. A summary of the relevant literature, the details of the individual components of the proposed system, datasets used to train and test the system and the experimental results are discussed in the subsequent sections of the paper.

2. Literature Review

A number of researchers have proposed various methods for recognizing offline handwritten text in English and other languages. Some models have been developed for various Indian languages, including Bengali. In this section, we review some of these existing works.

2.1 Works on Handwritten Word Segmentation

The segmentation of handwritten text is a well-explored area in document analysis and optical character recognition (OCR). Yet, it presents significant challenges, particularly with complex scripts like Bengali.

In earlier techniques, methods such as Projection Profile (Chaudhuri and Bera, 2009), Smearing method (Roy et al., 2008), Water Reservoir Method (Pal et al., 2001), Connected Component Analysis (Khandelwal et al., 2009), Scale Space method (Manmatha and Rothfeder, 2005) were commonly used to segment characters, lines, and words from handwritten document images. However, all the above methods have some major weaknesses. Many of these methods were extensively explored in Indian handwritten text

segmentation, and some of them are extremely sensitive to parameters like slanted and skewed at the word level, irregular gaps, noise, and language-specific complexities. Some are extremely sensitive to parameter settings and are effective for only one or a few specific types of languages or handwriting.

Researchers also combined multiple techniques to develop hybrid approaches. For instance, Mullik et al. (2015) developed a hybrid approach for text line segmentation in handwritten Bangla documents. They started by smudging the document image to reduce white spaces between words, followed by segmenting the image at its whitest pixels. Thinning was then used to address multi-line components, identifying the most likely separation points. In recent work, Rakshit et al. (2023) used region-based filling operation and an innovative light projection technique that differentiates text lines from the background. However, we feel that the method heavily depends on low-level preprocessing, such as binarization and computation of connected components.

The most recent methods have shifted towards deep learning approaches, overcoming challenges like line skewness, uneven spacing, and irregular paragraph shapes seen in earlier techniques. CNN-based semantic methods (Keserwani et al., 2019; Inunganbi et al., 2021; Dutta et al., 2021; Pramanik and Bag, 2020) were broadly explored for text segmentation. Jindal and Ghosh (2023) utilized an instance segmentation method with an R-CNN, a region proposal network (RPN), and a region of interest pooling layer to segment ancient Indian handwritten scripts.

2.2 Works on Handwritten Word Recognition

The early works focused on recognizing isolated characters by using classical machine-learning techniques. For instance, Das et al. (2010) developed a Multi-Layer Perceptron (MLP) and SVM-based classifier for Bengali HCR. They obtained an average recognition rate of 79.25% using MLP and 80.51% using SVM. Basu et al. (2012) designed a classifier using an MLP for Bengali HCRs using a feature set of 76 elements. Their reported accuracy on the training dataset was 86.46%, and on the testing dataset, it was 75.05%. Roy et al. (2014) implemented a Hidden Markov Model (HMM) based architecture for offline Bengali word recognition. They used zone segmentation techniques to model the various zones and used HMM in the middle zone.

Later, the trend shifted towards the use of deep learning techniques. Purkaystha et al. (2017) developed a convolutional deep learning model that uses kernels and local receptive fields and then employs densely connected layers for the recognition task. They tested their system on the BanglaLekha-Isolated dataset and obtained a testing accuracy of 89.93%. Gupta et al. (2023) used the CNN-Transformer model and built handwritten OCR from eight Indian languages. They worked on word recognition from handwritten text. Haque et al. (2022) proposed an approach for Handwritten Bangla Character Recognition using CNN and the YOLOv5 algorithm, achieving 96% accuracy in character recognition and 91% in word recognition.

However, most of the works mentioned above used an image of a word as input to the system. In the current paper, we consider the image of a whole handwritten page that contains multiple text lines and words as input. So, those recognition methods are not directly applicable. Therefore, we proposed using a segmentation phase before the recognition.

3. Methodology

In this work, we present a two-phase technique for converting an image of a full handwritten page into corresponding machine-readable text. The workflow of the technique is summarized in Figure 1, and the working of the individual phases is discussed below.

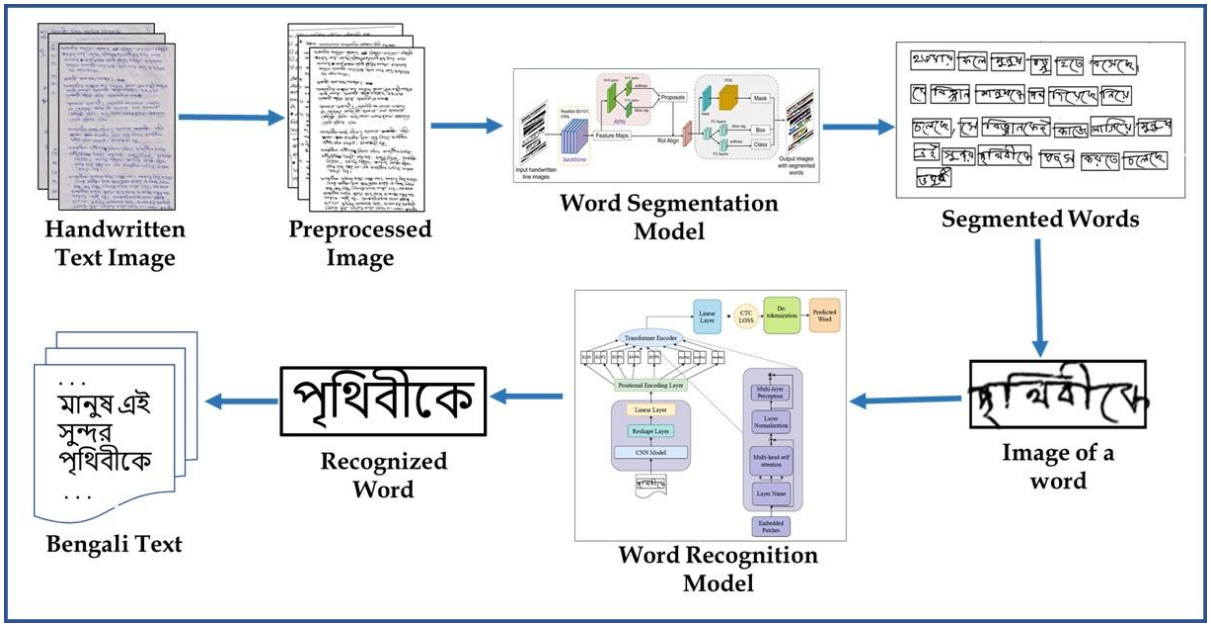


Figure 1. Workflow of the proposed system for converting handwritten page images into text.

3.1 Bengali Handwritten Word Segmentation

This phase aims to extract images of individual words from images containing several words. First, we developed the baseline system using a CNN-based semantic segmentation model. Then, we designed an advanced instance segmentation model using Mask R-CNN, which provided more refined segmentation capabilities.

3.1.1 Baseline System

A CNN-based encoder-decoder architecture for word segmentation has been used as the baseline. This task has been treated as a semantic segmentation problem with two pixel-level classes: words and background. The encoder extracts semantic features through convolutional and pooling layers, while the decoder upsamples and performs pixel-wise classification to generate a dense probability map. The model is trained on a pixel-level annotated dataset, with pre-processing applied to handle challenges like touching characters and marginal noise in handwritten scanned documents. The model uses 5×5 filters, stride 1, and max-pooling for downscaling, doubling the number of filters at each scale to effectively encode complex and fine-grained features. The final output is a dense pixel-wise probability map that predicts whether each pixel belongs to a word or the background. This ensures robust and accurate segmentation of words from scanned handwritten document images. This method aimed to predict the word boundaries within the images directly. However, during experiments, we found that the performance was suboptimal, particularly in handling documents with dense text and small handwriting, where the model struggled to identify word boundaries accurately.

3.1.2 Mask R-CNN based Segmentation

Due to the suboptimal performance of the CNN-based approach, particularly in handling documents with dense text and small handwriting, we employed an instance segmentation method using Mask R-CNN. This method integrates object detection with semantic segmentation by evolving previous methods, including RCNN, Fast RCNN, and Faster RCNN. As shown in Figure 2, Mask R-CNN architecture consists of three key stages: (a) generation

of region proposals, (b) classification of each proposal, and (c) segmentation of each proposal to predict a pixel-level mask for the detected objects.

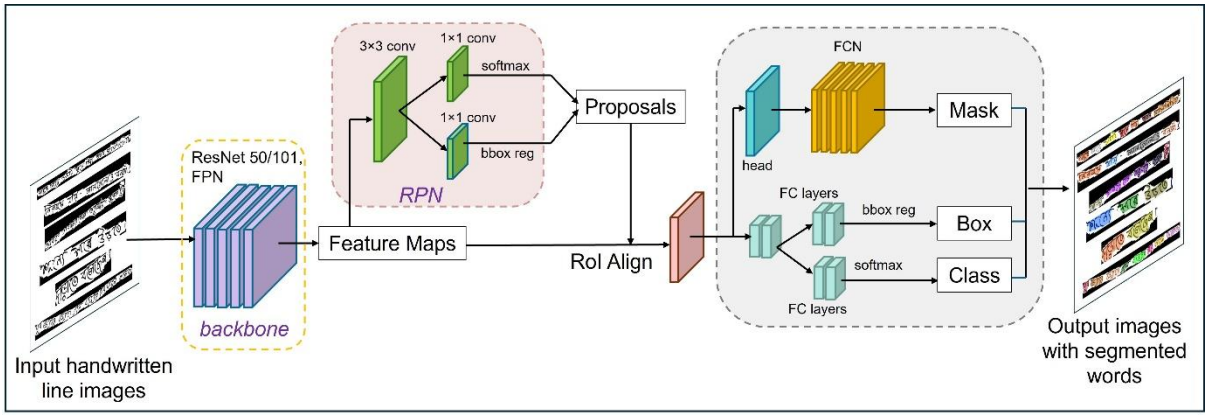


Figure 2. The Mask R-CNN architecture for word segmentation.

Here, we employed full-page handwritten answer sheet images for segmenting or extracting words. A fully convolutional network (FCN) is used for pixel-to-pixel segmentation in the Mask R-CNN, which also comprises the Faster R-CNN for object discovery. We can see a backbone network responsible for convolutions and the generation of feature maps. The backbone network in Mask R-CNN is typically a pre-trained convolutional neural network, such as ResNet, ResNeXt, VGGNet, etc. The FPN is added to this backbone network to create a feature pyramid. ResNet layers act as feature extractors, and the FPN helps to capture features at multiple scales, which is particularly useful for detecting words of varying sizes in handwritten documents. This technique employs a Region Proposal Network (RPN) to generate candidate regions likely containing words, predicting objectness scores and bounding box coordinates. The ROI Align layer ensures precise feature map alignment for each proposal, correcting quantization errors from traditional ROI pooling. This improves word detection and segmentation by retaining spatial feature alignment, which is essential for handwritten documents.

3.2 Bengali Handwritten Word Recognition

This phase aims to convert images of individual words into the corresponding text.

3.2.1 Baseline Model

This study first used a CNN-based architecture with CTC loss as the baseline model for handwritten Bengali word recognition. The CNN acts as a feature extractor, capturing spatial patterns through convolutional and pooling layers. Then, a BiLSTM layer processes the extracted features to model sequential dependencies in the text. The output is passed through a Softmax layer, producing character probabilities, and the final word is obtained using Connectionist Temporal Classification (CTC) loss, which aligns predictions with the target sequence without requiring explicit segmentation. The CNN consists of four convolutional layers, each followed by max-pooling to progressively extract hierarchical representations. The BiLSTM comprises two bidirectional layers with 256 units each, ensuring the model captures both forward and backward contextual dependencies in handwritten words. The model is trained with the Adam optimizer and a learning rate of 1e-4, optimizing character-

level predictions for improved recognition accuracy.

3.2.2 A CNN-Transformer Model for Recognition

In the experiments using the CNN-BiLSTM models, it was found that both models failed to correctly predict words containing constant conjuncts and diacritic forms of vowels (Juktakhors). These limitations motivated us to explore further. It was felt that the mapping of images of handwritten characters to corresponding characters of the Bengali vocabulary could be interpreted as a sequence-to-sequence mapping problem. Transformers were proven to be more efficient in sequence-to-sequence mapping tasks since they allow parallel processing of tokens and employ self-attention mechanisms. So, the final model is based on *CNN-Transformer* architecture and is shown in Figure 3.

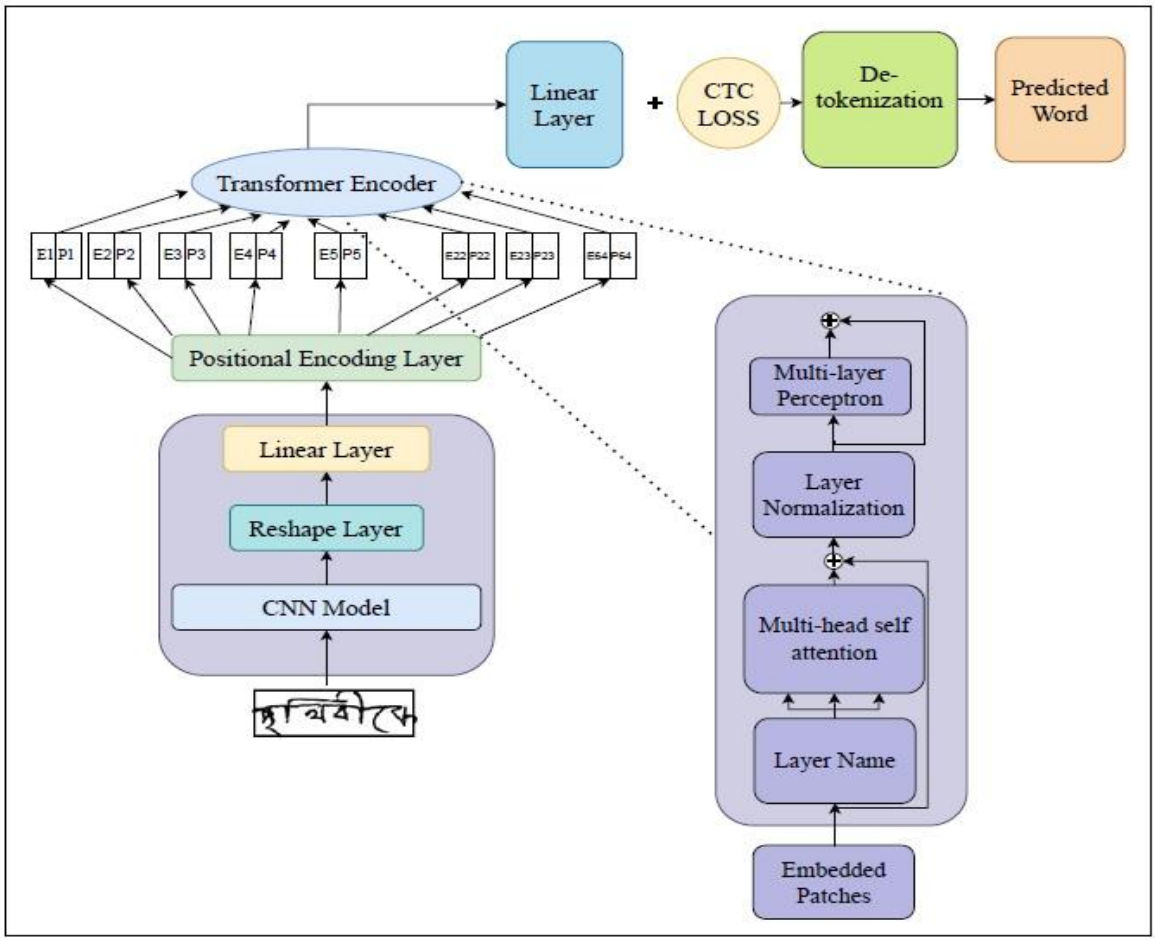


Figure 3. The workflow of the proposed CNN-Transformer model

The proposed CNN-Transformer model has thirteen convolution layers, and the last layer of the CNN is a max pool layer of pool size (2, 2). The depth of CNN was increased by stacking multiple convolutional layers, each with a non-decreasing number of filters (e.g., 64, 128, 256, 512, etc.). This design choice enables the network to learn a diverse range of features at different levels of abstraction, enhancing its capability to represent intricate patterns in image data effectively. Each convolution layer internally uses an activation function, which adds an element of non-linearity and helps the model learn complex patterns and training until convergence. Here, ReLU is used. Additionally, batch normalization normalizes the activation within each mini-batch, leading to faster training and improved model stability by reducing internal covariate shifts and acting as a regularizer. The max pool layer helps reduce computational complexity and control overfitting by extracting dominant features while

retaining crucial spatial information. The hyperparameters of the model were carefully tuned to achieve optimal performance. After processing through the CNN, the output feature map is reshaped and permuted to prepare it for linear transformation. The reshaping operation converts the feature map into a suitable format for subsequent processing by the transformer's encoder layers. For every batch of feature maps that the CNN generates, a positional encoding is applied to inject positional information into the sequence representation before passing it through the encoder layers.

Finally, the transformer's output sequence representation is passed through a linear layer to predict the target sequence. This linear layer maps the transformer's output from the model's internal representation to the target vocabulary size, generating the final predictions. Lastly, the de-tokenization module is used to convert the output tensor from our transformer into the corresponding predicted word.

4. Experimental Results

This section discusses the results obtained from the experiments.

4.1 Dataset for Word Segmentation

There are certain openly available datasets; however, the handwriting of school students contains specific challenges. The openly available Bengali handwritten datasets do not cover these domain-specific issues and are not very suitable for our development. Therefore, we created a domain-specific dataset. To create the dataset, we collected student answer books of grade V-VII students from Bengali-medium schools. These answer books are scanned, and full-page images are used to prepare the dataset.

We first cleaned the images to remove the noises that could interfere with segmentation accuracy. Then, the images were converted into binary format. This binary conversion is crucial as it reduces computational complexity and focuses the model's attention on the text structure, facilitating more effective learning and analysis. Next, the images are manually annotated with word regions using VGG Image Annotator (VIA) to prepare the training data. During annotation, the individual units (essentially words) were marked on a zoomed image by a set of points forming a closed polygon. The training data comprised 205 annotated handwritten answer sheet images. The test set consists of 20 randomly selected answer sheet images, comprising 2098 words in total.

4.2 Dataset for Word Recognition

We used a deep learning architecture to develop our model, so a large training dataset is required. An openly available Bengali handwritten dataset, BN-HTRd[1], is used for the experiments. The BN-HTRd dataset (Rahman et al., 2023) is derived from the BBC Bangla News corpus and encompasses 788 full-page images meticulously compiled from the contributions of 150 distinct writers. Furthermore, it comprises a total of 108,147 instances of handwritten words, organized across 13,867 lines and incorporating 23,115 unique words.

4.3 Evaluation Metrics

To evaluate the performance of the segmentation models, we used % accuracy. The accuracy is computed as the ratio of correctly segmented words to the total number of words in the test data.

The primary metrics used for the evaluation of word recognition models are the Word Recognition Rate (WRR) and Character Recognition Rate (CRR). WRR measures the percentage of correctly recognized words in the dataset, providing an overall assessment of

[1] https://data.mendeley.com/datasets/743k6dm543/1

the system's performance at the word level. On the other hand, CRR evaluates the percentage of correctly recognized characters, presenting a fine observation capability of the model to correctly predict individual characters within word.

4.4 Results of the Word Segmentation Model

In this study, we extracted words from full-page handwritten documents using a baseline semantic segmentation approach employing a CNN model using the aforementioned dataset. The model was trained for ten epochs, each comprising 1,000 iterations. The model achieved an accuracy of 73.02% on the in-house school dataset. We have also tested the system using two other test datasets. On the Bengali online handwritten test images, this model correctly detected 1367 words among 1779 total words with accuracy of 76.84%, as shown in Table 1.

Next, we have adopted an instance segmentation approach for segmenting words from full-page handwritten images using the Mask R-CNN model. When we computed the word count-based segmentation accuracy, we found that the R-CNN model achieves 79.64% accuracy. However, it performed well when the model was tested using the online test dataset. For the Bengali online dataset, the model correctly detected 1701 words out of 1779 total words, achieving an accuracy of 95.61%. As the inter-line and inter-word gap is mostly uniform and the datasets mainly comprise matured handwriting, it identified the boundaries correctly.

Table 1. *Performance of Word Segmentation.*

Model	Dataset	Total Words	Detected Words	Accuracy
Baseline	In-house: answer books	2098	1532	73.02%
	Online Bengali	1779	1367	76.84%
Mask R-CNN	In-house: answer books	2098	1671	73.02%
	Online Bengali	1779	1701	95.61%

4.5 Results of the Word Recognition Model

The baseline model was trained for 120 epochs. The model achieved 81.34% WRR and 89.07% CRR when tested on an in-house dataset. We have also tested the system when it was trained using the BN-HTRd dataset. Then the WRR and CRR values are 73.4% and 82.62%, respectively. These values prove that the handwriting of school students is more difficult to recognize, and available open datasets alone cannot provide high accuracy. Furthermore, experiments were conducted using the proposed CNN-Transformer model. The transformer-based model showed a substantial performance improvement over the baseline. The model was trained until convergence, at 43 epochs. Then, it achieved a WRR of 91.74% and a CRR of 95.18%. The values are summarized in Table 2. This improvement proves the superiority of the proposed transformer-based model in the handwritten word recognition task.

Table 2. *Performance of Word Recognition.*

Model	Training Data	Test Data	WRR	CRR
Baseline	In-house	In-house	81.34%	89.07%
	BN-HTRd	In-house	73.40%	82.60%
CNN-Transformer	In-house	In-house	91.74%	95.18%
	BN-HTRd	In-house	84.81%	89.70%

4.6 Performance of the 2-phase Model

To assess the model's overall performance, we take a few images of handwritten pages from the in-house school dataset as input to the system, run both phases and compute the accuracy of the final word recognition. The performance is measured using percentage accuracy: how many words are correctly recognized by the final system divided by the total number of words. There we found that the system achieved an accuracy of 78.12%. When we tested the performance of the final system using online handwritten images, the system achieved an accuracy of 93.62%. The performance is relatively lower in the handwritten answer-book images because of higher errors in segmentation. When we analyzed the outputs, we found that several words were partially segmented and degraded overall performance.

5. Discussion and Future Work

We proposed a two-phase architecture for converting the image of a handwritten page into machine-readable text. The first phase performs word segmentation, and the second phase recognizes the segmented words. Deep-learning architecture is used in both phases. The system aimed to recognize handwritten answer books of middle-school students.

We found that the model could not segment the words when the words are densely written, and the inter-word gap is not uniform. The model tends to club multiple words into one segment. Again, in some cases, one word is segmented into multiple units due to a larger gap between individual characters of a word. The use of better deep learning architecture and language model-based post-processing might improve the performance of the system. Creation of larger domain-specific training data for segmentation and recognition can be another direction to work in the future.

Funding

The work has been partially funded by ANRF/SERB, India with Project File No: EEQ/2021/000687.

References

AlKendi, W., Gechter, F., Heyberger, L., Guyeux, C.: Advancements and challenges in handwritten text recognition: A comprehensive survey. Journal of Imaging 10(1), 18 (2024).

Basu, S., Das, N., Sarkar, R., Kundu, M., Nasipuri, M., and Basu D. K.: Handwritten Bangla Alphabet Recognition using an MLP Based Classifier. arXiv: arXiv.1203.0882 (2012)

Bhowmik, T.K., Parui, S.K., Roy, U., Schomaker, L.: Bangla handwritten character segmentation using structural features: a supervised and bootstrapping approach. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 15(4), 1–26 (2016).

Chaudhuri, B.B., Bera, S.: Handwritten text line identification in Indian scripts. In: 2009 10th International Conference on Document Analysis and Recognition. pp. 636–640. IEEE (2009)

Das, N., Das, B., Sarkar, R., Basu, S., Kundu, M., Nasipuri, M.: Handwritten Bangla basic and compound character recognition using MLP and SVM classifier, arXiv:1002.4040. (2010).

Dutta, A., Garai, A., Biswas, S., Das, A.K.: Segmentation of text lines using multi-scale CNN from warped printed and handwritten document images. International Journal on Document Analysis and Recognition (IJDAR) 24(4), 299–313 (2021).

Gupta, M.K., Vikram, S., Dhawan, S., Kumar, A.: Handwritten OCR for word in Indic language using deep networks. In: 2023 10th International Conference on Signal Processing and Integrated Networks (SPIN). pp. 389–394 (2023).

Haque, P., Salma, U., Chowdhury, R.: Bangla handwritten character and words recognition-based on the yolov5 algorithm. In: 2022 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE). pp. 24–29 (2022).

Hossain, M.I., Rakib, M., Mollah, S., Rahman, F., Mohammed, N.: Lila-Boti: Leveraging isolated letter accumulations by ordering teacher insights for Bangla handwriting recognition. In: 2022 26th International Conference on Pattern Recognition (ICPR). pp. 1770–1776. IEEE (2022).
Inunganbi, S., Choudhary, P., Manglem, K.: Meitei Mayek handwritten dataset: compilation, segmentation, and character recognition. The Visual Computer 37(2), 291–305 (2021).
Jindal, A., Ghosh, R.: Word and character segmentation in ancient handwritten documents in Devanagari and Maithili scripts using horizontal zoning. Expert Systems with Applications 225, 120127 (2023).
Keserwani, P., Ali, T., Roy, P.P.: Handwritten Bangla character and numeral recognition using convolutional neural network for low-memory GPU. International Journal of Machine Learning and Cybernetics 10, 3485–3497 (2019).
Khandelwal, A., Choudhury, P., Sarkar, R., Basu, S., Nasipuri, M., Das, N.: Text line segmentation for unconstrained handwritten document images using neighborhood connected component analysis. In: Pattern Recognition and Machine Intelligence: Third International Conference, PReMI 2009 New Delhi, India, December 16-20, 2009 Proceedings 3. pp. 369–374. Springer (2009)
Manmatha, R., Rothfeder, J.L.: A scale space approach for automatically segmenting words from historical handwritten documents. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(8), 1212–1225 (2005).
Moitra, M., Saha, S.K.: A review on handwritten text segmentation in Indian languages. *Int. J. Mach. Learn. & Cyber.* (2024). https://doi.org/10.1007/s13042-024-02448-1
Mullick, K., Banerjee, S., Bhattacharya, U.: An efficient line segmentation approach for handwritten Bangla document image. In: 2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR). pp. 1–6. IEEE (2015)
Pal, U., Belaïd, A., Choisy, C.: Water Reservoir based approach for touching numeral segmentation. In: Proceedings of Sixth International Conference on Document Analysis and Recognition. pp. 892–896. IEEE (2001)
Pramanik, R., Bag, S.: Segmentation-based recognition system for handwritten Bangla and Devanagari words using conventional classification and transfer learning. IET Image Processing 14(5), 959–972 (2020).
Purkaystha, B., Datta, T., Islam, M.S.: Bengali handwritten character recognition using deep convolutional neural network. In: 2017 20th International Conference of Computer and Information Technology (ICCIT). pp. 1–5 (2017).
Rahman, M.A., Tabassum, N., Paul, M., Pal, R., Islam, M.K.: BN-HTRd: A benchmark dataset for document level offline Bangla handwritten text recognition (HTR) and line segmentation. In: Computer Vision and Image Analysis for Industry 4.0, pp. 1–16 (2023).
Rakshit, P., Halder, C., Sk, M.O., Roy, K.: A generalized line segmentation method for multi-script handwritten text documents. Expert Systems with Applications 212, 118498 (2023).
Roy, P.P., Dey, P., Roy, S., Pal, U., Kimura, F.: A novel approach of Bangla handwritten text recognition using HMM. In: 14th International Conference on Frontiers in Handwriting Recognition. pp. 661–666 (2014).
Roy, P.P., Pal, U., Lladós, J.: Morphology based handwritten line segmentation using foreground and background information. In: International conference on frontiers in handwriting recognition. pp. 241–246 (2008).
Teslya, N., Mohammed, S.: Deep learning for handwriting text recognition: existing approaches and challenges. In: 2022 31st Conference of Open Innovations Association (FRUCT). pp. 339–346. IEEE (2022)

VisionTutor: An Adaptive Tutoring Platform for Real-Time Progressive Learning

Sunil Gopal C V[1], Sudhan S[1], Sushanth R[1], Shreyas G S[1], Abhilash C B[1*], Rohitaksha K[1], Manjunath K V[2], Ashwin T S[3]
[1]*JSS Academy of Technical Education, Bengaluru, Karnataka, India*
[2]*Indian Institute of Information Technology Dharwad, Karnataka, India*
[3]*Computer Science, Vanderbilt University, United States*
[*]abhilashcb@jssateb.ac.in

Abstract: Advances in artificial intelligence have transformed digital learning, yet current tutoring systems remain limited in their ability to integrate real-time multimodal perception, pedagogical grounding, and contextual responsiveness—particularly in STEM education. We present VisionTutor, a real-time adaptive tutoring platform that leverages Gemini 2.5 Pro to support live screen monitoring, speech-based dialogue, and multimodal input comprehension. VisionTutor provides immediate, context-sensitive feedback through a canvas-based environment and conversational tutoring, enabling step-by-step guidance in mathematics and programming. A central contribution is the Cognitive Learning Scoring Model, trained on 5,000 simulated learner–system interactions using a DistilBERT regression pipeline, which predicts engagement and effectiveness across parameters such as promptness, tool usage, problem-solving strategies, and AI interaction patterns ($R^2 = 0.9856$). Grounded in the ICAP and Self-Regulated Learning frameworks, the scoring model translates behavioral indicators into pedagogically meaningful constructs. Findings suggest that VisionTutor not only enhances personalization but also generates interpretable analytics that support both learners and instructors. This work advances the integration of multimodal AI into education by combining adaptive tutoring, real-time learning analytics, and explainable feedback processes, thereby laying a foundation for next-generation intelligent learning environments.

Keywords: Multimodal AI Tutoring, Context-Aware Learning Analytics, Cognitive Scoring System, Learning Score.

1. Introduction

The field of educational technology has evolved significantly over the past decade, driven by rapid advances in artificial intelligence (AI) and its integration into learning environments. Despite these developments, current tutoring systems remain limited in their ability to deliver truly adaptive, multimodal, and pedagogically grounded support that meets learners' immediate needs and provides comprehensive analytics for instructors (Eden et al., 2024). Many AI-driven platforms operate in silos, restricted to text-based interaction, delayed feedback, or minimal contextual awareness, thereby failing to capture the richness and responsiveness of human tutoring interactions (Alfarra et al., 2024).

This limitation is particularly acute in mathematically intensive and computationally rich domains such as mathematics and computer programming. Effective problem solving in these contexts often requires visual reasoning, step-by-step explanation, debugging, and immediate scaffolding—activities that are poorly supported by conventional tutoring systems. Bloom's classic "2 Sigma Problem" highlighted that individualized tutoring produces learning gains two standard deviations above conventional classroom instruction, setting a benchmark that AI-based tutoring has long sought to approach. Yet, even with these aspirations, most existing

systems have not matched the adaptivity and multimodal responsiveness of expert human tutors.

Recent advances in large language models (LLMs) have substantially improved the ability of AI to process natural language and sustain dialogue. However, their application in education has often been limited to text-based conversational systems, which lack integration with multimodal inputs, real-time analytics, and pedagogical scaffolding frameworks (Cohn et al., 2025; Cohn & Fonteles et al., 2025). As Baker et al., (2016) argue, intelligent tutoring systems must go beyond linguistic competence to provide adaptive, context-sensitive feedback that reflects cognitive and behavioral indicators of learning. Similarly, VanLehn et al., (2011) emphasize the importance of aligning educational technologies with assessment structures that can capture deeper learning processes.

To address these gaps, we propose VisionTutor, a real-time adaptive AI tutoring system designed for STEM education. Built on Gemini 2.5 Pro (https:\cloud.google.com\vertex-ai\generative-ai\docs\models\gemini\2-5-pro), VisionTutor integrates speech, screen activity, code, and visual reasoning into a unified, low-latency platform. By combining multimodal perception with pedagogical frameworks such as ICAP (Interactive, Constructive, Active, Passive) (Chi, M. T. et al., 2014) and Self-Regulated Learning (SRL) (Zimmerman et al., 2002) the system delivers not only instant, personalized support but also interpretable learning analytics for both learners and instructors. A distinctive feature of VisionTutor is its Cognitive Learning Scoring Model, which leverages a DistilBERT-based regression pipeline trained on 5,000 simulated learner–system interactions to predict engagement and effectiveness along parameters such as promptness, tool use, and problem-solving approach (Sanh, V. et al., 2019).

A distinctive strength of VisionTutor lies in its cognitive learning scoring system, which continuously analyzes multimodal interaction signals—including verbal reasoning, problem-solving strategies, screen activity, and canvas engagement—to provide interpretable measures of learning progress (Niyozov et al., 2023). Unlike conventional systems that rely on lagged or periodic feedback, VisionTutor delivers immediate, context-sensitive guidance by mapping behavioral cues into pedagogical constructs such as visual problem solving, adaptive guidance, and contextual interpretation. This real-time adaptivity fosters engagement comparable to one-to-one tutoring and transforms complex interaction data into actionable feedback on learners' strengths, weaknesses, and trajectories. By integrating multimodal tracking with explainable analytics, the system not only supports personalized learning but also empowers students to regulate their progress as an ongoing, self-directed process.

The design of VisionTutor builds on and extends recent AI-in-education innovations. Prior work has explored multimodal analysis of learner engagement (Mousavinasab et al., 2021; Ashwin, T. S et al., 2023, TS, A., et al., 2020), AI-powered personalized learning systems (Rekha et al., 2024), and adaptive data-driven education (Ashwin et al., 2020). However, these systems are typically constrained by narrow modality, lack of explainable feedback, or limited grounding in learning theory. VisionTutor contributes by integrating multimodal adaptivity with interpretable scoring mechanisms that map technical behaviors into established pedagogical constructs.

Overall, our contributions are as follows:

- *Design of VisionTutor*, a real-time multimodal tutoring system that integrates screen, audio, speech, and code to deliver adaptive, personalized support.
- *Introduction of the Cognitive Learning Score*, dynamically computed across five pedagogically motivated dimensions—contextual adaptability, visual problem solving, conversational guidance, adaptive assistance, and screen-context interpretation.
- *Development of a multimodal feedback engine* that continuously monitors learner behavior and produces explainable, actionable feedback to support incremental problem solving, persistence, and motivation.

This study examines whether a multimodal, adaptive tutoring system can improve engagement and cognitive reflectiveness in problem-solving tasks.

The remainder of the paper is organized as follows: Section 2 details the proposed methodology and system components. Section 3 presents results and discusses performance and learning-ability scores. Section 4 outlines future enhancements and concludes the paper.

2. Data and System Methodology

2.1 Dataset Generation

This study required the generation of a synthetic dataset in order to provide VisionTutor with sufficient context for score computation and evaluation. To simulate realistic learner behavior and support robust testing, we created a conversation dataset representing diverse academic scenarios and student personas within the Indian educational context. Each instance in the dataset represents a tutoring dialogue between a student and the AI system across STEM subjects such as Physics, Chemistry, Mathematics, and Programming.

The dataset was enriched with pedagogical and behavioral attributes to enable modeling of both surface-level interactions and deeper cognitive engagement. Key attributes include:

- conversation: A transcript capturing the exchange between the student and tutor, including queries, clarifications, and feedback.
- subject & scenario: The topic under discussion (e.g., wave properties, stoichiometry), providing domain-specific grounding.
- student_persona: Simulated learner profiles (e.g., Analytical, Passive, Impatient) that influence interaction style and support requirements.
- difficulty_level: Represents content complexity (Beginner, Intermediate, Advanced).
- interaction_type: Reflects the overall intensity of interaction (High, Moderate, Low).
- engagement_depth: A numeric score (1–5) indicating how deeply the learner engages with content and tools.
- cognitive_load: An estimate of the mental effort exerted during the session.
- dialogue_flow: Measures how coherent and progressive the conversation is, reflecting mutual understanding.
- learning_behavior: Captures persistence, help-seeking, and constructive tool usage.
- multimodal_integration: Indicates how effectively the student uses voice, canvas, and textual modalities together.
- overall_interactiveness: A composite score reflecting session quality, derived using a DistilBERT-based model trained on real student-AI interactions.

The dataset was generated using heuristic rules informed by prior learner behavior research and validated in consultation with subject-matter experts. While synthetic, the dataset incorporates realistic variation in engagement, errors, and behavioral tendencies, closely approximating classroom dynamics. Academic experts reviewed the dataset to ensure pedagogical plausibility and alignment with real-world conditions.

2.2 VisionTutor and Exiting Systems

The VisionTutor platform integrates a multimodal, context-sensitive tutoring environment on the Gemini 2.5 Pro architecture (Gemini 2.5 Pro 2025). By synchronizing screen activity, speech, and visual problem-solving behavior, the system supports real-time educational interaction.

In contrast, most existing tools operate in single-mode environments. For example, GitHub Copilot focuses on code suggestions, while Google Math Solver addresses symbolic manipulation in mathematics. These systems operate in isolation and lack the integration of modalities, contextual awareness, and real-time adaptivity. VisionTutor bridges this gap by combining speech, image, and code analysis with a canvas-based environment to deliver instantaneous, personalized feedback. The system not only processes learners' direct input but also interprets contextual signals and engagement patterns, thereby providing adaptive support grounded in education theory.

2.3 Multimodal Context Acquisition

To enable dynamic tutoring, VisionTutor utilizes continuous screen monitoring and multimodal input interpretation. Learner behavior is analyzed through light-weight screen monitoring to identify context-dependent learner activity—like open code editors, math solvers, or canvas drawings. Learner speech input is simultaneously transcribed and interpreted with visual inputs, like on-screen diagrams and handwritten mathematical expressions. These inputs are channeled through Gemini 2.5 Pro's multimodal pipeline, blending code understanding, image processing, and conversational inference.

2.4 Adaptive Tutoring and Feedback Generation

Based on the interpreted context, VisionTutor provides individualized feedback in the form of a mix of dialog-based tutoring, visual feedback, and code/debug hints. For instance, while coding, the system detects syntax or logical mistakes and provides step-by-step debugging assistance. In mathematics, it interprets visual steps of problem-solving and provides verbal scaffolding or hints through the in-built canvas environment. This real-time responsiveness is added to mimic human tutoring, allowing students to get instant, relevant assistance specific to their activity.

2.5 Cognitive Learning Scoring Mechanism

The score framework is established through a 5,000-labelled-interactions synthetic dataset in coding and math contexts. A transcript of the conversation, subject, difficulty rating, learner type, interaction type, and engagement level are all included in each record. An overall cognitive score is predicted by a tuned regression model through DistilBERT, which translates these conversation features into a cumulative interactiveness score. This score is a combination of sub-elements such as frequency of tool use, relevance of AI-augmented responses, and interaction pacing.

Sample characteristics are

- Early Prompt Score: How quickly a learner initiates problem-solving.
- Tool Usage: Help tool usage frequency, such as canvas or hints.
- AI Proficiency: Smoothness and independence in using AI guidance.

These are then summed up as:

$$\text{Final Score} = 0.5 \times \text{TaskCompletion} + 0.3 \times \text{Self-confidence} + 0.2 \times \text{AIProficiency}.$$

The model achieved an R^2 value of 0.9856 on the test dataset, which is a measure of high predictive power.

VisionTutor's scoring system maps onto the ICAP model in which greater scores represent more constructive and interactive behaviors such as debugging, step-by-step explanation, or multimodal tool use. The adaptive feedback loop also facilitates Self-Regulated Learning (SRL), with students having control of their performance and being provided with specific hints. Including markers of behavior such as persistence, tool use, and verbal reasoning, the system translates technological interaction into levels of cognitive effort, thus attaining pedagogical validity in the scoring.

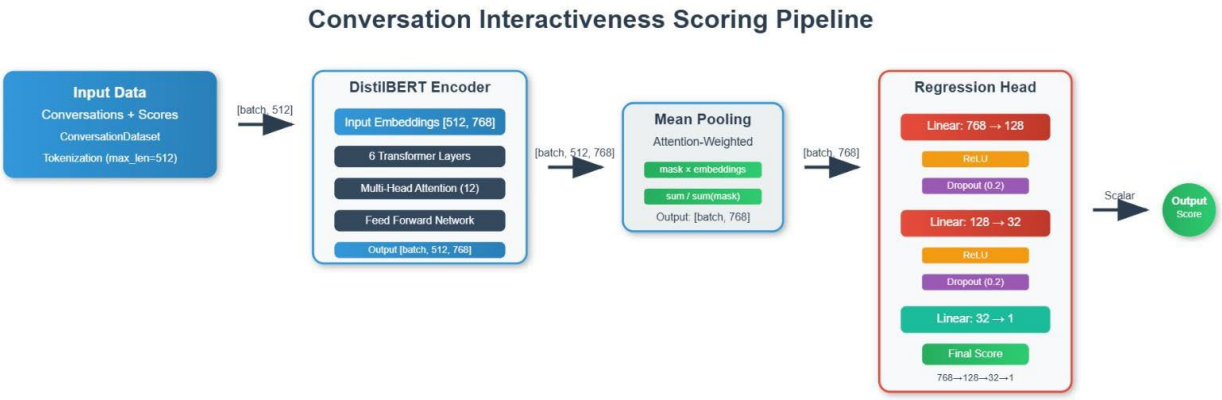


Figure 1: Proposed Methodology of VisionTutor

3. Results and Discussion

3.1 System Responsiveness and Interaction Quality.

VisionTutor demonstrated a strong capacity for real-time responsiveness, with an average system latency of less than 1.5 seconds across multiple sessions. During coding practice, the system readily read on-screen materials and voice-typed questions and typed them into the computer, with timely and appropriate feedback. For instance, during a JavaScript tutorial, the system provided real-time code explanation as well as voice-guided instructions in different languages, as shown in Figure 2.

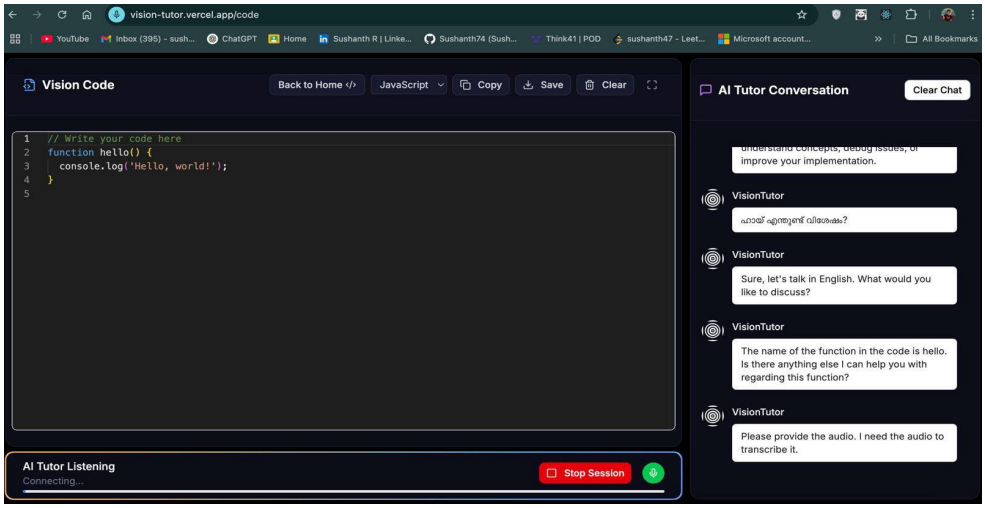


Figure 2. VisionTutor providing live coding support with real-time voice interaction and context-aware code explanations.

3.2 Visual Problem-Solving and Conversational Guidance

In math sessions, learners used a digital canvas to solve problems, writing, drawing, and explaining their solutions. The system accurately recognized handwritten input and provided verbal assistance and visual feedback based on the user's spoken queries. As shown in Figure 3, during a geometry problem involving the Pythagorean theorem, VisionTutor offered step-by-step guidance based on the learner's drawing and voice input, allowing for a more dynamic and interactive learning experience.

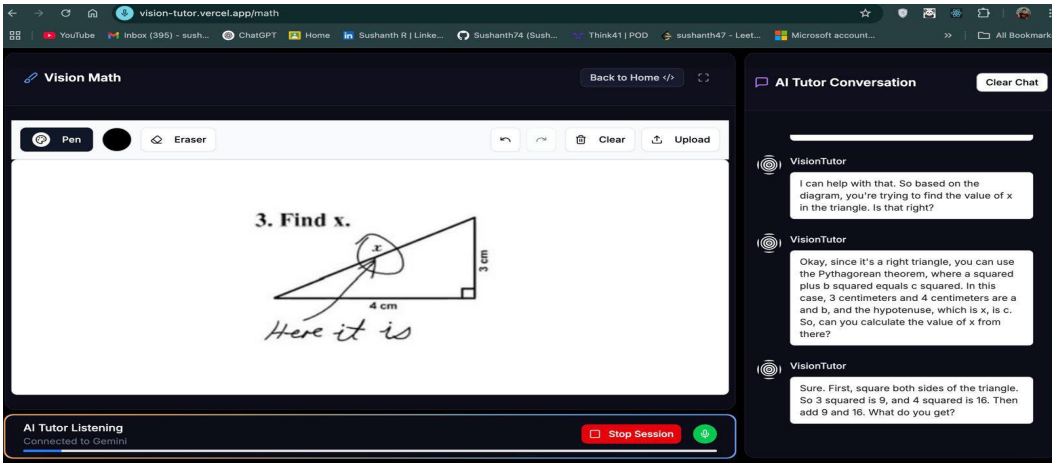


Figure 3: Real-time canvas interaction supporting mathematical reasoning with AI-based speech assistance

3.3 Cognitive Scoring Evaluation

At the conclusion of each session, VisionTutor generated a personalized learning score by analyzing behavioral metrics, such as problem-solving approach, tool usage, and screen interactions. The score was derived from four main factors: conceptual understanding, problem-solving approach, mathematical reasoning, and improvement over time.

For example, Figure 4 presents a learner's session report, which includes a score of 3.5 out of 5 (categorized as Advanced), highlighting strengths in reasoning and providing actionable feedback.



Figure 4. Example session assessment report with cognitive score and personalized feedback

3.4 Cognitive Scoring Breakdown

To further examine VisionTutor's scoring framework, we conducted a study with 20 students (15–inclusive–20 years old) from local secondary and early undergraduate schools via voluntary signup, all having a basic understanding of math and programming. With a within-subject observational design, each participant had one 45–60 minute session of math and coding activities. There was no control group; learning gains were instead measured via initial participation and end-of-session performance scores. Although not a controlled experiment, this study offers preliminary findings about learner behavior and system adaptability. Baseline comparisons and formal experimental verification will be included in future work.

Table 1. Session-wise Scores

User	Duration (min)	Concepts Learned	Initial Engagement	Improvement	Final Performance
U001	42	5	2.8	0.95	3.75
U002	38	4	3.0	0.25	3.25
U003	55	6	2.2	2.05	4.25
U004	30	3	2.5	0.25	2.75
U005	48	5	3.8	0.20	4.00
U006	36	4	2.9	0.60	3.50
U007	60	6	3.3	1.20	4.50
U008	29	3	1.7	1.30	3.00
U009	41	4	3.0	0.25	3.25
U010	33	4	2.5	0.00	2.50
U011	50	5	3.0	0.75	3.75
U012	44	4	2.5	1.00	3.50
U013	37	4	2.2	1.05	3.25
U014	58	6	3.1	1.40	4.50
U015	35	3	2.4	0.35	2.75
U016	46	5	2.7	1.30	4.00
U017	39	4	2.8	0.70	3.50
U018	31	3	2.1	0.90	3.00
U019	53	5	3.2	0.55	3.75
U020	47	5	2.6	0.90	3.50

Table 1 presents the key session metrics we use to inform our cognitive scoring model. Duration measures time and is used to interpret effort and interaction depth. Concepts Learned gives a quantitative indication of content learning and progress. Engagement Score monitors how actively and rapidly a learner begins and early motivation. Improvement monitors intra-session progress and the direction of the learner. Performance monitors task completion, confidence, and autonomy. These parameters collectively inform us about learner behavior and the effectiveness of VisionTutor in facilitating academic progress. We used a multi-component scoring system. Each session was measured in terms of three broad indicators:

Early Engagement*:* Monitors early session activity, including early initiation, tool use, and pace of interaction.

Improvement: Documents behavioral and conceptual gains made during the session.

Final Performance: Symbolizes quality of outcome, confidence, and independence in performing tasks.

Both were based on weighted behavioral markers:

- Early Engagement Score = 0.5 * EarlyPromptScore + 0.3 * ToolUsage + 0.2 * FirstInteractionSpeed
- Final Performance Score = 0.5 × TaskCompletion + 0.3 × Confidence + 0.2 × AIProficiency
- Improvement Score = Final Performance − Initial Engagement

A stacked bar chart in Figure 5, illustrates these three factors across 20 learners. This is supplemented by Table 2, which provides exact figures for session duration, concepts learnt, and the three scores.

3.5 Cognitive Learning Score: Design and Computation

The Cognitive Learning Score (CLS) in VisionTutor is designed as a composite metric that captures learner behavior, tool interaction, and AI-guided engagement in real time. The CLS is computed using five interpretable components, each reflecting a distinct dimension of the

learning process:

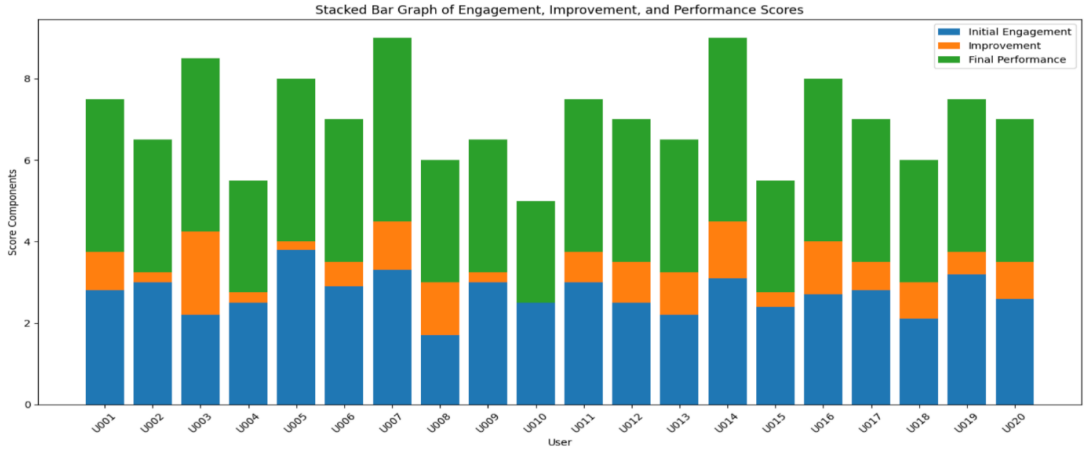


Figure 5. Stacked Bar Graph of Engagement, Improvement and Performance

- Tool Usage: Frequency and diversity of interactive tools employed (canvas, voice, chat, etc.).
- Early Prompt Score: The learner's tendency to initiate problem solving or seek hints early in the session.
- AI Proficiency: Autonomy and fluency in leveraging AI guidance, predicted by a fine-tuned DistilBERT regression model trained on over 5,000 labeled learner–AI conversations.
- Engagement Depth: Ratio of exploratory activity compared to passive viewing or response copying.
- Visual Context Score: Quality and relevance of learner-generated sketches or diagrams on the canvas.

These components are combined into a cumulative score that reflects both immediate engagement and longer-term learning potential. The model achieved an R^2 value of 0.9856 on test data, demonstrating high predictive accuracy.

By design, the CLS maps directly onto the ICAP framework, with higher scores corresponding to more constructive and interactive behaviors such as debugging, step-by-step explanation, or multimodal tool use. In addition, the adaptive feedback loop facilitates Self-Regulated Learning (SRL) by offering learners interpretable analytics and targeted hints. By embedding behavioral markers such as persistence, tool usage, and verbal reasoning, the CLS ensures pedagogical validity while also providing real-time, actionable insights into learning progress. These findings suggest that VisionTutor's adaptive scoring mechanism not only predicts performance but also reflects deeper shifts in learner behavior consistent with the ICAP framework—for instance, movement from passive observation toward more constructive activities such as debugging and multimodal explanation. The real-time analytics further support self-regulated learning (SRL) by encouraging learners to monitor their strategies and adapt their approaches during problem solving.

3.6 Inferential Statistics

To evaluate VisionTutor's impact, we conducted statistical comparisons of learner engagement and learning gains between VisionTutor sessions and a baseline condition. An independent samples t-test revealed that learners using VisionTutor achieved significantly higher scores in both engagement and learning outcomes.

As reported in Table 3, the mean engagement score for VisionTutor learners was 4.21 (SD = 0.49), compared to 3.54 in the baseline condition ($p = 0.003$). Similarly, the mean learning gain was 22.4% (SD = 4.6) with VisionTutor, compared to 15.7% in the baseline ($p = 0.01$). The effect sizes were medium to large (Cohen's $d = 0.72$ for engagement and $d = 0.65$ for learning gain), indicating that the observed improvements are both statistically significant and educationally meaningful. Notably, the observed effect sizes (Cohen's d = 0.72 for

engagement; d = 0.65 for learning gain) are comparable to those reported in meta-analyses of intelligent tutoring systems (e.g., VanLehn, et al., 2011), positioning VisionTutor within the range of interventions that approximate the effectiveness of human tutoring. This alignment indicates that multimodal adaptivity and instant feedback are not merely technical features, but pedagogically significant drivers of improved engagement and learning outcomes.

Figure 6 further illustrates the comparative distribution of engagement scores. VisionTutor learners demonstrated higher median performance and reduced variability, suggesting that the system not only raised average outcomes but also supported a more consistent learning experience across participants. Together, these results provide preliminary evidence that VisionTutor's multimodal adaptivity and real-time feedback mechanisms can enhance both learner engagement and academic progress beyond that of conventional baseline systems.

Table 3: Statistical test for VisionTutor and Baseline Model

Metric	VisionTutor	Baseline	p-value	Effect Size (Cohen's d)
Mean Engagement Score	4.21	3.54	0.003	0.72 (medium-high)
Mean Learning Gain (%)	22.4%	15.7%	0.01	0.65 (medium)

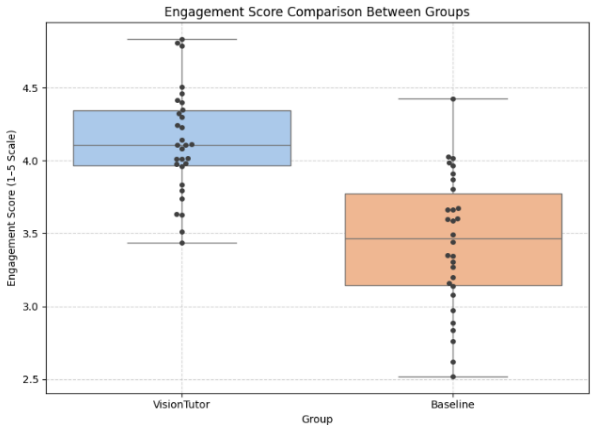


Figure 6: Comparative box plot visualizing the engagement score distributions between the VisionTutor and baseline groups.

While these results are promising, they remain preliminary. The study involved a relatively small sample size and lacked a fully randomized control group, which limits the generalizability of findings. Moreover, reliance on a synthetic dataset for training may not fully capture the variability of real-world learner interactions. Future large-scale studies with authentic learner data are necessary to confirm the robustness and external validity of the observed effects.

4. Conclusion and Future Work

VisionTutor introduces a novel approach to intelligent tutoring by integrating speech, screen interaction, code behavior, and visual problem solving into a unified, adaptive environment. Designed specifically for STEM education, the system delivers immediate, context-aware support that more closely approximates the responsiveness of human tutoring than traditional AI-based systems. At the core of VisionTutor is the Cognitive Learning Score, an interpretable metric that enables learners to monitor their own engagement, progress, and problem-solving strategies in real time. The strong predictive reliability of the scoring model ($R^2 = 0.9856$) underscores the system's capacity to generate trustworthy analytics that inform not only learner self-reflection but also instructional decision-making. In this way, VisionTutor serves a dual purpose: supporting students directly while also assisting instructors by highlighting patterns of engagement, persistence, and conceptual growth.

Looking ahead, future research will focus on validating the system with larger and more diverse learner populations. In particular, we aim to extend evaluation beyond synthetic datasets by analyzing authentic learner logs at scale, thereby strengthening external validity. Further work will also explore refining multimodal recognition (e.g., reducing noise sensitivity in speech and handwriting inputs) and expanding pedagogical alignment to additional learning science frameworks. By combining multimodal adaptivity, interpretable analytics, and theoretically grounded pedagogy, VisionTutor contributes to the development of next-generation intelligent tutoring systems that not only personalize learning but also provide actionable insights for advancing teaching practice.

References

Eden, C. A., Chisom, O. N., & Adeniyi, I. S. (2024). Integrating AI in education: Opportunities, challenges, and ethical considerations. Magna Scientia Advanced Research and Reviews, 10(2), 006-013.

Alfarra, A. H., Amhan, L. F., Mosa, M. J., Alajrami, M. A., El Kahlout, F., Abu-Nasser, B. S., & Abu-Naser, S. S. (2024). AI-driven learning: Advances and challenges in intelligent tutoring systems.

Ashwin, T. S., Prakash, V., & Rajendran, R. (2023). A systematic review of intelligent tutoring systems based on Gross body movement detected using computer vision. Computers and education: Artificial intelligence, 4, 100125.

Cohn, C., Snyder, C., Fonteles, J. H., TS, A., Montenegro, J., & Biswas, G. (2025). A multimodal approach to support teacher, researcher and AI collaboration in STEM+ C learning environments. British Journal of Educational Technology, 56(2), 595-620.

Cohn, C., Fonteles, J. H., Snyder, C., Srivastava, N., Campbell, D., Montenegro, J., & Biswas, G. (2025). Exploring the design of pedagogical agent roles in collaborative stem+ c learning. In Proceedings of the 18th International Conference on Computer-Supported Collaborative Learning-CSCL 2025, pp. 330-334. International Society of the Learning Sciences.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. Educational psychologist, 49(4), 219-243.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. Theory into practice, 41(2), 64-70.

Mousavinasab, E., Zarifsanaiey, N., R. Niakan Kalhori, S., Rakhshan, M., Keikha, L., & Ghazi Saeedi, M. (2021). Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. Interactive Learning Environments, 29(1), 142-163.

Ashwin, T. S., Prakash, V., & Rajendran, R. (2023). A systematic review of intelligent tutoring systems based on Gross body movement detected using computer vision. Computers and education: Artificial intelligence, 4, 100125.

TS, A., & Guddeti, R. M. R. (2020). Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks. Education and information technologies, 25(2), 1387-1415.

Ashwin, T. S., & Guddeti, R. M. R. (2020). Impact of inquiry interventions on students in e-learning and classroom environments using affective computing framework. User Modeling and User-Adapted Interaction, 30(5), 759-801.

Niyozov, N., Saburov, S., Ganiyev, S., & Olimov, S. (2023). AI-powered learning: revolutionizing technical higher education institutions through advanced power supply fundamentals. In E3S Web of Conferences (Vol. 461, p. 01092). EDP Sciences.

Yadannavar, S., Ashwin, T. S., Pathan, R., Vanahalli, M. K., & Rajendran, R. (2024, December). Multimodal Analysis of Learning-Centered Emotions and Cognitive Processes in Open-Ended Learning Environments. In 2024 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE) (pp. 1-8). IEEE.

Baker, R. S. (2016). Stupid tutoring systems, intelligent humans. *International journal of artificial intelligence in education*, *26*(2), 600-614.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational psychologist*, *46*(4), 197-221.

Google. (2025). *Gemini 2.5 Pro architecture* [Generative AI model]. Retrieved August 19, 2025, from https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro

Do Categorical Emotions Have Intensity? Modeling Emotional Intensity Transitions by Task, Peer, and Environment Causes

Ryan LI[a], Ashwin T S[a], Divya MEREDDY[a], & Gautam BISWAS[a]
[a]*Computer Science, Vanderbilt University, USA*
*ashwindixit9@gmail.com

Abstract: Emotions shape students' learning experiences. While research links positive emotions to better outcomes and negative emotions to poorer ones, few studies examine how emotions change in intensity and cause during collaborative learning. This paper introduces and operationalizes a tripartite cause framework—Task, Peer, and Environment—and connects graded learning-centered emotions to their triggers. Using EcoJourneys, a narrative-centered, game-based learning environment, we annotated the emotional states of 21 elementary school students, capturing shifts among five core emotions (Engagement, Boredom, Confusion, Frustration, Delight) across four intensity levels. We investigated four research questions: which causes trigger transitions; how causes affect state duration; whether cause effects vary by intensity; and how theory-grounded causes (CVT, SDT, EST) shape intensity transitions. Our findings show that most transitions increase Engagement intensity. Peer causes frequently support recovery from moderate Confusion to higher Engagement, Task causes often drive within-engagement transitions, and environmental influences show more sustained effects. This work offers a theory-grounded account of emotional intensity dynamics in collaborative learning and informs the design of emotion-aware learning technologies.

Keywords: Learning-Centered Emotions, Collaborative Learning, Emotional Intensity, Emotional Transitions, Emotional Analysis, EcoJourneys, External Causes

1. Introduction

Emotions are integral to understanding students' learning, particularly in Self-Regulated Learning (SRL) (Zimmerman, 1986) and Socially Shared Regulation of Learning (SSRL) (Järvelä *et al.*, 2015). They complement cognitive states and provide insight into engagement, group dynamics, and collaborative learning regulation. In classrooms, learning-centered emotions—Engagement, Boredom, Confusion, Frustration, and Delight—are prevalent compared to universal emotions, forming the basis for affective research in education (D'Mello & Graesser, 2012; TS and Biswas, 2024).

Previous studies treated emotions as discrete states (TS and Biswas, 2024; Ashwin *et al.*, 2024). However, observations indicate emotions like Confusion and Engagement exhibit varying intensities (D'Mello & Graesser, 2014). Mild Confusion may signal curiosity, while intense Confusion implies a higher cognitive load that can hinder learning (D'Mello & Graesser, 2014). These differences necessitate modeling emotions by type and intensity, prompting exploration of emotional dynamics through graded intensities and transitions (Reisenzein & Junge, 2024; Williams *et al.*, 2023). However, modeling intensity alone does not make interventions actionable. Prior approaches emphasize internal appraisals of emotion (e.g., Control–Value Theory, CVT; Pekrun, 2006) and, in empirical practice, often do not link intensity changes to external drivers such as task design, peer interaction, or environmental conditions. To enable targeted, designable interventions, we examine external causes of intensity transitions—Task, Peer, and Environment.

In this paper, we introduce and operationalize a tripartite cause framework—Task, Peer, and Environment—for analyzing emotional intensity transitions in collaborative learning. We ground these causes in established theory: Control–Value Theory for task-related

influences (Pekrun, 2006), Self-Determination Theory (SDT) for peer-related influences (Deci & Ryan, 2012), and Ecological Systems Theory (EST) for environmental influences (Bronfenbrenner, 1979). We select CVT, SDT, and EST as a minimal, complementary set: CVT captures task appraisals (control, value) underlying shifts among engagement, confusion, and frustration; SDT explains how autonomy, competence, and relatedness shape peer-driven affect; and EST models contextual conditions (e.g., noise, interruptions, fatigue) that modulate attention and regulation. Other theories (e.g., ACT, CLT, SCT, EVT) emphasize different phenomena or omit one or more of the task/peer/environment pathways, making them less suitable as a compact explanatory set for our focus on actionable external causes.

Using multimodal data from 21 students in a narrative-centered, collaborative learning environment (EcoJourneys), we analyze how these causes shape transition types, durations, and intensity patterns. Recognizing emotion-intensity transitions is not only theoretically significant but also critical for designing responsive educational technologies. Adaptive learning systems must respond differently to mild disengagement versus deep frustration, and detecting the intensity of an emotion can inform tailored interventions. Modeling transitions across intensity levels enables real-time emotional awareness that supports personalized scaffolding, making these systems more responsive to the evolving affective states of learners. Accordingly, we test how these causes shape transition types, durations, and intensity patterns in collaborative learning.

Specifically, this study addresses the following research questions:

RQ1: What causes among Task, Peer, and Environment trigger emotional transitions?

RQ2: How do these causes affect the duration of emotional states?

RQ3: Do different causes trigger transitions at different intensities of the same emotion?

RQ4: How do Task, Peer, and Environment causes, as conceptualized through CVT, SDT, and EST, shape emotional intensity transitions in collaborative learning?

By addressing these questions, we aim to enhance understanding of collaborative learning dynamics and inform the design of affect-sensitive technologies.

2. Related Work

Emotion Intensity in Learning Theories: Theories have been proposed that prompt the investigation of different intensities that can exist within the same emotion. Control-Value Theory (CVT) explains how internal appraisals influence emotional experiences during learning (Pekrun, 2006). Extensions of CVT include physiological signals to estimate emotion intensity (Barradas et al., 2025), acknowledging graded emotions. However, CVT focuses on internal appraisals and provides limited guidance on external influences, such as task structure or peer interaction. The Circumplex Model (Russell, 1980; Akpanoko et al. 2024; Akpanoko and Biswas, 2024) supports a continuous representation of emotions in valence-arousal space. Yet, empirical applications often fail to distinguish functional variations within emotions or connect shifts in emotion to external causes, such as the impact of emotions on the overall learning process. This leaves gaps in understanding how intensity evolves in collaborative settings. CVT and the Circumplex model drive the investigation of within-emotion intensity and its development. However, empirical research still underexplores the links between changes in intensity and external factors such as tasks, peers, or the environment.

Modeling Emotional Transitions: Markov models are instrumental in analyzing emotional transitions, predicting the likelihood of one emotion following another. Cipresso *et al.* (2023) used a Markov chain model to compute transitional probabilities between emotions like Stressed, Engaged, Bored, and Relaxed, demonstrating the utility of probabilistic models in learning contexts. Prasetio *et al.* (2020) further integrated Time-Delay Neural Networks with Markov chains to predict stress and emotion transitions, highlighting temporal aspects in affect modeling. In the context of learning, emotional state modeling has employed Hidden Markov Models using multimodal data, such as emotion detection from facial expressions (Schmidt *et al.*, 2010; Bosch *et al.*,2016), facial thermal imaging (Liu & Wang, 2011), and sentiment analysis (Ho & Cao, 2012). These studies often use discrete categorical frameworks and overlook transitions within the same emotion at varying intensities, such as escalating from

mild to intense Confusion.

Causal Attribution of Emotional Transitions. Prior work in learning sciences and educational psychology points to three broad classes of external influences on students' affect during collaborative learning: task-related, peer-related, and environment-related factors. Control-Value Theory explains how task appraisals (control and value) shape engagement, confusion, and frustration (Pekrun, 2006). Self-Determination Theory highlights how peer interactions and the fulfillment of competence and relatedness needs regulate motivation and affect (Deci & Ryan, 2012; Järvelä *et al.*, 2015). Ecological perspectives emphasize how environmental conditions (e.g., noise, visual clutter, interruptions) modulate attention and emotional regulation (Bronfenbrenner, 1979; Godwin & Fisher, 2011; Fisher *et al.*, 2014; Gheller *et al.*, 2023). Empirical studies show that well-calibrated tasks sustain engagement, whereas poorly calibrated tasks elicit confusion or frustration (Özhan & Kocadere, 2020; Gijlers & de Jong, 2013). Peer scaffolding can facilitate recovery from negative states (Fu *et al.*, 2009; Järvelä *et al.*, 2015), and environmental distractions are linked to disengagement and boredom (Shernoff, 2013; Godwin & Fisher, 2011; Fisher *et al.*, 2014; Gheller *et al.*, 2023). While CVT and SDT support task- and peer-related influences, and ecological perspectives highlight environmental factors, the literature typically examines these influences in isolation and models emotions as discrete categories. This paper introduces and operationalizes a unified tripartite cause framework—Task, Peer, and Environment—and links these causes to fine-grained emotional transitions across intensity levels.

3. Methodology

Learning Environment and Data: EcoJourneys is a game-based collaborative learning environment where students investigate the illness in tilapia on an island in the Philippines (Acosta *et al.*, 2024). It includes a tutorial and three inquiry quests, with activities like Deduce and TIDE (Talk, Investigate, Deduce, Explain) that promote evidence-based reasoning. In this IRB-approved study, 21 elementary students participated over five days, engaging in about an hour of gameplay daily. All students and guardians provided informed consent before data collection, which included multimodal recordings of video, speech, and logs to capture emotional and behavioral indicators during the collaborative process. To enable analysis aligned with our research questions, we annotate both emotional categories and their intensity levels to model graded transitions rather than only discrete states. This supports examining transition types, durations, and intensity patterns in subsequent analyses (RQ1–RQ3), and provides the basis for theory-grounded interpretation (RQ4).

3.1 Emotion Annotation Scheme and Intensity Definition

We annotated students' emotional states using synchronized video, audio, and screen recordings on the *EcoJourneys* platform (Figure 1). Annotations focused on individual task engagement and group interactions as students worked in teams of 3–4 on inquiry-driven activities. We highlighted five learning-centered affective states—Engagement, Confusion, Frustration, Boredom, and Delight—defined by D'Mello & Graesser (2012).

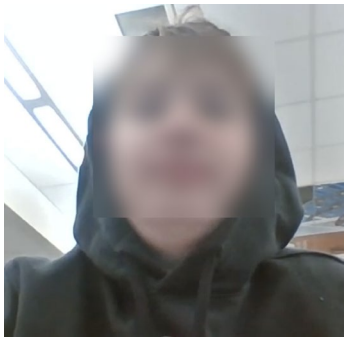
Emotions in learning environments vary in intensity, with observable behaviors and verbal cues indicating different affective levels (Gupta *et al.*, 2016; Xu *et al.*, 2023). Facial Action Coding System (FACS; Ekman & Friesen, 1978), along with other modalities such as hand gestures (Vicario & Newman, 2013; Ashwin and Guddeti, 2020a), body postures (Dael *et al.*, 2012; Ashwin and Guddeti, 2020b), are examples of visual cues that we can directly observe from the learners. Transcribed text, pitch, and tone are examples of verbal cues that we can hear (Fonteles *et al.*, 2024). These multimodal signals help distinguish emotional intensity functionally and perceptually, in which similar observations have been noted in earlier studies on affective expression in learning (D'Mello & Graesser, 2014; Bosch *et al.*, 2016).

Guided by these insights, we observed and classified emotional expressions into four intensity levels: *low, moderate low, moderate high, and high*, based on facial, body, and verbal indicators. These categories emerged through iterative observation and were later formalized. This clustering revealed stable patterns among learners, supported by theory-driven

distinctions in intensity thresholds, especially for Confusion (D'Mello & Graesser, 2014). Only three intensity levels for Boredom were noted in our dataset. Two trained annotators independently coded emotion categories, intensity levels, and causes, achieving high inter-rater reliability (Cohen's κ = 0.91), confirming consistency in our intensity judgments.



(a)



(b)

Figure 1(a) (on the left) A scene where the avatar is controlled by the students. Figure *1(b)* (on the right) The student is using the EcoJourneys software to complete the task. Note that the student's face is not shown for IRB compliance.

Table 1. *Emotion Intensity Scale (Used Across All Emotions)*

Level	Description of Each Emotion Intensity
1	Low – Minimal expression, short duration, subtle or isolated cue
2	Moderate low – Clearly expressed but not dominant, 1–2 cues observed
3	Moderate high – Noticeable intensity, 2+ cues, co-occurring behaviors
4	High – Dominant emotion with rich multimodal expression and sustained cues

Two trained annotators independently coded the emotional states and causes using a standardized rubric. Initial training involved calibration rounds with practice videos, where discrepancies were discussed and resolved until inter-rater reliability exceeded 0.9. During annotation, each coder worked independently and flagged ambiguous cases for joint review with a third expert. Regular consistency checks and review meetings maintained label accuracy throughout the dataset. This structured workflow ensured reliable, theory-aligned labeling across students and sessions. Intensity levels were determined through manual, multimodal observation of student behaviors, based on the idea that higher-intensity emotions are expressed through multiple cues, such as facial expressions, body posture, and vocal signals. Lower-intensity states generally have a single, brief cue. Observed patterns were grouped into four distinct levels for each emotion. Each entry included a timestamp with start time, end time, emotion, intensity level, and cause (e.g., Confusion 2 → Engagement 3, Cause: Peer). This approach allowed systematic analysis of emotion changes in relation to their causal context and intensity fluctuations. Table 1 summarizes the intensity scale used across all emotions.

Cause Annotation Scheme: To capture the context of each emotional transition, annotators assigned a cause label: Task, Peer, Environment, or a combined label (Peer/Task, Task/Environment, Peer/Environment) when influences co-occurred. Task was coded when difficulties in progressing (e.g., ambiguity, navigation/usability issues) precipitated the shift; Peer when interactions with group members altered affect (e.g., help that resolved confusion, social comparison, miscommunication); and Environment when external distractions (e.g., classroom noise, instructor prompts, self-initiated off-task behavior) were primary drivers. We permitted overlapping cause annotations to reflect co-occurring influences in collaborative settings. For interpretation in RQ4, we align Task with CVT, Peer with SDT, and Environment

with EST; the cause framework itself is introduced in this paper. Operational cues for each emotion at each intensity level are listed in Table 2.

Table 2. *Emotion-Specific Intensity Descriptors*

Emotion	Level 1	Level 2	Level 3	Level 4
Engagement	Gaze unfocused, glancing around	Semi-focused, occasional distractions	Focused, limited peer interaction while working	Fully focused, immersed, no external attention shifts
Confusion	Brief hesitation, single cue	Repeated questions, procedural uncertainty	Multiple failed attempts, verbal + nonverbal cues	Facial strain, physical gestures, vocal escalation
Frustration	Mild complaint or sigh	Fast typing, annoyed tone	Physical agitation, task-related frustration	Raised voice, visible anger, peer-directed frustration
Boredom	Stretching, looking away briefly	Longer disengagement, chatting	Extended inattention, no task input	Complete withdrawal, slouched posture, no interaction with task or group *(not observed)*
Delight	Subtle smile or posture shift	Light laughter or excitement	Verbal joy, leaning forward	Expressive body language, vocal exclamations

Table 3. *Cause Category Definitions*

Cause	Definition of the Causes
Task	Triggered by interactions with the learning software or task content
Peer	Triggered by verbal/nonverbal interactions with group members
Environment	Triggered by factors outside the task or group, e.g., classroom noise, instructor intervention
Peer/Task	Interaction with peers about task progress or content
Task/Environment	Task engagement is influenced by the physical or instructional environment
Peer/Environment	Peer interaction involving environmental topics or students outside the group

Annotation Structure and Transition Format: In this experiment, we document types of transitions, including emotion categories, their intensity levels, causes, frequencies, and durations. An annotation example is (0:30, 1:20, Engagement, 4, Peers), indicating the student exhibited a high level of Engagement from 30 seconds to 1 minute and 20 seconds, influenced by peers. To annotate a transition, we consider the next instance. For example, (0:30, 1:20, Engagement, 4, Peers) and (1:20, 1:35, Confusion, 3, Task) can be noted as Engagement 4

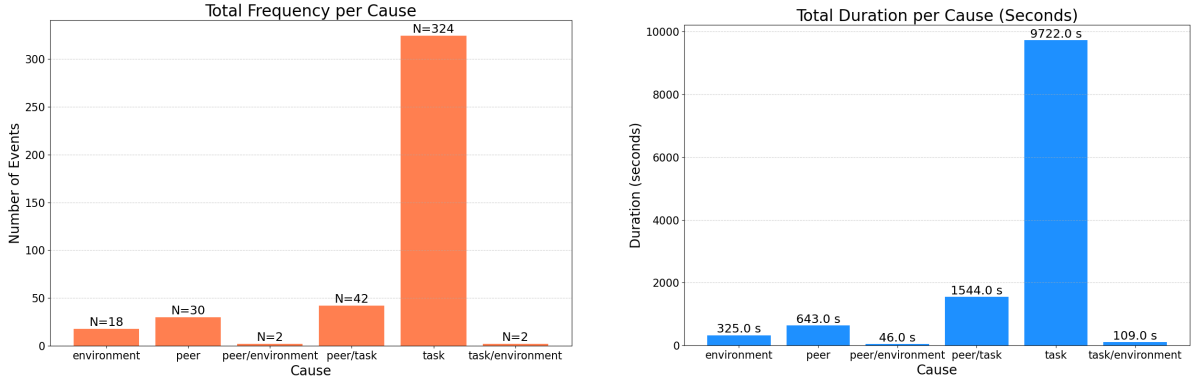
→ Confusion 3 due to Task (as the transition to Confusion 3 stems from the task itself). Definitions and examples of the cause categories appear in Table 3.

4. Results

We analyzed 418 emotional transitions, totaling 12,389 seconds of activity from 21 students. Off-task periods (1,975 seconds) and undetectable facial recordings (391 seconds) were excluded. Below are findings regarding our four research questions.

RQ1: *What causes among Task, Peer, and Environment trigger emotional transitions*?

We first quantify which causes are associated with observed emotional transitions (Figure 2). Transitions were primarily triggered by Task-related causes (324 transitions, 77.51%), followed by Peer (30), Environment (18), Peer/Task (42), and minimal contributions from Peer/Environment (2) and Task/Environment (2). Task causes accounted for the most unique transitions (81 types), compared to Peer (13) and Peer/Task (23). Most transitions led to increased Engagement. For instance, under Task, the top transitions were Engagement 2 → Engagement 3 (10%), Engagement 3 → Engagement 2 (9%), and Engagement 3 → Engagement 4 (7%). Among Peer-triggered transitions, the most common was Confusion 2 → Engagement 3 (17%), indicating peer collaboration is vital for emotional recovery. A classroom study reported a similar cause distribution (Task ≫ Peer > Environment), reinforcing that the prevalence of Task-triggered transitions is a robust pattern in collaborative learning settings.



(a) *(b)*

Figure 2. Distribution of (a) total frequency and (b) total duration counts of causes across all transition types.

RQ2: *How do these causes affect the duration of emotional states*?

We next examine how cause categories relate to the duration of emotional states. In terms of total duration, Task-based transitions dominated (9,722 seconds), followed by Peer/Task (1,544s), Peer (643s), Environment (325s), Task/Environment (109s), and Peer/Environment (46s). Notably, Environment-related causes had a long average duration per instance (mean = 59.87s), suggesting that environmental distractions may have more sustained effects on emotional states.

Transitions from Engagement 2 → Engagement 3 accounted for the most extended cumulative duration (2,798 seconds), indicating that students spent significant time increasing attention during collaborative gameplay. Task-triggered transitions involving non-Engagement states into Engagement also had high durations, with examples such as Boredom 1 → Engagement 4 (131s, single instance) and Boredom 1 → Engagement 3 (91s, single instance). A one-way ANOVA showed marginally significant differences in transition duration across cause types: $F(5, 412) = 2.42$, $p = 0.065$, suggesting that the distribution of durations varies depending on the causal trigger.

RQ3: *Do different causes trigger transitions at different intensities of the same emotion*?

We analyzed transitions across intensity levels using a Markov Chain framework. Task-related causes primarily led to moderate-to-high levels of Engagement, including Engagement 2 → Engagement 3, Engagement 3 → Engagement 4, and Engagement 3 → Engagement 2. Peer causes often helped students recover from Confusion into higher Engagement, most notably Confusion 2 → Engagement 3. High-intensity Frustration 4 and Confusion 4 were only observed under Task, Peer, and Peer/Task conditions, indicating that negative high-intensity states emerge more frequently from direct interaction with content or social dynamics. In contrast, Delight 4 only appeared in transitions related to Task and Environment.

A Chi-square test examining the association between cause and emotion category revealed that Engagement transitions were most diversely caused: $\chi^2(10, N = 418) = 11.65$, $p = 0.0399$, *Cramér's V* = 0.1851, indicating a medium effect size. This suggests that Engagement intensity transitions are particularly sensitive to multiple contextual factors.

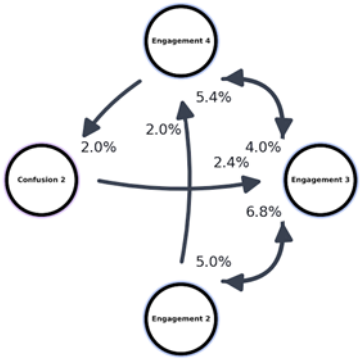


Figure 3. Frequent intensity transitions (≥2% of all transitions).

In Figure 3, nodes denote emotion + intensity level; directed edges denote observed transitions. Labels show % of all transitions (N=418) and edge thickness is proportional to frequency. Transitions are aggregated across causes, complementing Figure 2 (cause distribution) and Figure 3 (durations). Most frequent transitions are within the Engagement spectrum (e.g., Engagement 2→3, 3→4) and recovery from Confusion 2→Engagement 3, complementing cause and duration summaries in Figures 2–3.

RQ4: *How do Task, Peer, and Environment causes (CVT, SDT, EST) shape emotional intensity transitions*?

Observed patterns align with the theoretical lenses used for interpretation. Task-related causes dominated transitions (324/418; 77.51%) and most often produced moderate-to-high Engagement intensity, including Engagement 2→3 (10%), Engagement 3→2 (9%), and Engagement 3→4 (7%). Peer-related causes frequently supported recovery from negative states, notably Confusion 2→Engagement 3 (17% of peer-triggered transitions). Environment-related transitions were fewer (18 total) yet showed a longer average duration per instance (≈59.87s), indicating more sustained contextual effects. Interpreted through CVT/SDT/EST, these results provide a concise, theory-grounded account of how external causes shape emotional intensity transitions in collaborative learning.

Patterns observed here are consistent with prior evidence from a narrative-centered classroom study of middle-school learners. That study reported that Task-related causes were the dominant triggers of emotional change (71.3%), with Peer at 23.2% and Environment at 7.1%; within categories, Task most often co-occurred with Engagement (60%), Peer with Engagement/Delight (50%/30%), and Environment with Boredom (60%). These distributions and within-category associations mirror our findings that Task chiefly drives within-Engagement intensity increases, Peer frequently supports recovery (e.g., Confusion 2 → Engagement 3), and Environment relates to disengagement-prone states. Together, the alignment across studies strengthens the interpretation that cause type systematically shapes intensity transitions in collaborative learning.

5. Discussion

This study examined how emotional transitions caused by Task, Peer, and Environment influence the type, intensity, and duration of a student's emotional state. By annotating multimodal data with emotion and cause, we provide a more nuanced understanding of affective dynamics in collaborative learning.

Our findings show that while Task-related causes are the most common, Peer-related causes effectively promote transitions from negative to positive states, especially from moderate Confusion to high Engagement. This supports claims from SRL and SSRL frameworks emphasizing that peer scaffolding and social regulation can stabilize or elevate emotional states (Järvelä *et al.*, 2015). In contrast, Environment-related causes, though less frequent, result in longer transition durations, suggesting that disruptive stimuli like noise or interruptions have lasting effects on emotional regulation. These findings align with ecological systems theory (Bronfenbrenner, 1979), which posits that environmental conditions significantly impact self-regulation from many perspectives. They highlight the need for thoughtfully designed learning environments. Transitions toward higher engagement were common across cause categories, yet only Task-related triggers were linked to high-intensity negative emotions like Frustration 4 and Confusion 4. This indicates that poorly aligned tasks, including those with characteristics such as redundancy or conflicting features, can disengage learners and induce frustration. Therefore, task calibration must balance challenge and support, especially in adaptive systems.

From a modeling perspective, intensity levels underscore the difference between mild and strong emotional expressions. For example, Confusion 1 may signify curiosity, while Confusion 4 can reflect overload. Recognizing these differences allows for targeted interventions—e.g., possibly addressing Task-induced Confusion at Level 2 with hints, while Level 4 may need peer or instructor support. Finally, Engagement emerged as the most diverse in causes, reinforcing that it is not a simple "on-task" signal but a dynamic state shaped by social, task-based, and environmental factors. This insight suggests that affect-aware learning systems should not only detect discrete emotional states but also understand transitions and context-specific causes.

Taken together, these findings show that Task-related transitions are consistent with Control–Value Theory (CVT), shaping movement toward moderate-to-high Engagement as well as occasional high-intensity Confusion/Frustration; Peer-related influences align with Self-Determination Theory (SDT), supporting recovery transitions such as Confusion 2 → Engagement 3; and Environment-related influences reflect Ecological Systems Theory (EST), exhibiting longer average durations indicative of sustained contextual effects. This mapping provides a concise explanation of how causes shape emotional intensity transitions in collaborative learning (answers RQ4).

5.1 Limitations

This study proposed a framework linking emotional intensity levels with external causes in collaborative learning, but it has limitations. The dataset was small (21 students) and manually coded, leading to limited generalizability despite high inter-rater agreement ($\kappa = 0.91$). Some transitions, like Boredom Level 4, were likely absent due to the game structure. Cause attribution relied on visible cues from video and audio, making these interpretive despite prior validation. Future studies should include system logs or gaze data for more explicit cause identification. All findings are correlational, with patterns emerging, such as peer interactions aiding recovery and tasks triggering diverse transitions, but causal mechanisms need validation through real-time or experimental studies. Despite these constraints, this study offers three key insights. First, categorical learning of emotions shows meaningful intensity levels. Second, transitions are influenced by identifiable causes: Task, Peer, and Environment. Third, modeling transitions reveals dynamics often missed in discrete emotion labeling. This work contributes to developing emotion-aware learning technologies that adapt to emotion types and their contextual evolution, supporting more collaborative and emotionally intelligent learning environments.

6. Conclusion

This study introduced a framework for modeling emotional transitions in collaborative learning by examining five categorical learning-centered emotions and their perceived intensity influenced by external causes—Task, Peer, and Environment. Through detailed annotations of 418 emotional transitions, we found that intensity transitions are patterned and context-sensitive. Task-related causes were most frequent and diverse, resulting in high Engagement and intense negative states. Conversely, Peer-related causes helped recovery from negative emotions, particularly moderate Confusion. Environmental causes, though less frequent, were linked to lasting emotional states.

Our findings underscore the need to integrate emotion intensity and causal context in affect-aware learning technologies. Instead of focusing only on emotional categories, intelligent systems should recognize nuanced emotional trajectories and triggers. Understanding how emotions evolve allows educators and designers to better support emotional regulation and engagement in collaborative learning. Future research should involve larger datasets, explore group-level emotion dynamics, and integrate predictive, real-time adaptive mechanisms addressing emotional intensity and causality.

Acknowledgements

This research was supported by the National Science Foundation AI Institute Grant No. DRL-2112635. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Acosta, H., Lee, S., Mott, B., Bae, H., Glazewski, K., Hmelo-Silver, C., & Lester, J. (2024). Multimodal Learning Analytics for Predicting Student Collaboration Satisfaction in Collaborative Game-Based Learning.

Akpanoko, C. E., Cordell, G., & Biswas, G. (2024). Investigating the relations between students' affective states and the coherence in their activities in open-ended learning environments. In Proceedings of the 17th International Conference on Educational Data Mining (pp. 511-517).

Akpanoko, C. E., & Biswas, G. (2024). The interplay of affective states and cognitive processes in an open-ended learning environment: A case study. In Proceedings of the 18th International Conference of the Learning Sciences-ICLS 2024, pp. 873-880. International Society of the Learning Sciences.

Ashwin, T. S., & Guddeti, R. M. R. (2020). Affective database for e-learning and classroom environments using Indian students' faces, hand gestures and body postures. Future Generation Computer Systems, 108, 334-348.

Ashwin, T. S., & Guddeti, R. M. R. (2020). Impact of inquiry interventions on students in e-learning and classroom environments using affective computing framework. User Modeling and User-Adapted Interaction, 30(5), 759-801.

Ashwin, T. S., SNYDER, C., AKPANOKO, C. E., Srigowri, M. P., & BISWAS, G. (2024, November). Combining Multimodal Analyses of Students' Emotional and Cognitive States to Understand Their Learning Behaviors. In International Conference on Computers in Education.

Barradas, I., Tschiesner, R., & Peer, A. (2025). Dynamic emotion intensity estimation from physiological signals facilitating interpretation via appraisal theory. *PloS one*, *20*(1), e0315929.

Bosch, N., D'mello, S. K., Ocumpaugh, J., Baker, R. S., & Shute, V. (2016). Using video to automatically detect learner affect in computer-enabled classrooms. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *6*(2), 1-26.

Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design.* Harvard University Press.

Cipresso, P., Borghesi, F., & Chirico, A. (2023). Affects affect affects: A Markov chain. *Frontiers in Psychology*, *14*, 1162655.

Dael, N., Mortillaro, M., & Scherer, K. R. (2012). Emotion expression in body action and posture. *Emotion*, 12(5), 1085.

Deci, E. L., & Ryan, R. M. (2012). Self-determination theory. *Handbook of theories of social psychology, 1*(20), 416-436.
D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, *22*(2), 145-157.
D'Mello, S. K., & Graesser, A. C. (2014). Confusion. In *International handbook of emotions in education* (pp. 289-310). *Routledge*.
Ekman, P., & Friesen, W. V. (1978). Facial Action Coding System. *Environmental Psychology & Nonverbal Behavior*.
Fisher, A. V., Godwin, K. E., & Seltman, H. (2014). Visual environment, attention allocation, and learning in young children: When too much of a good thing may be bad. *Psychological science, 25*(7), 1362-1370.
Fonteles, J., Davalos, E., Ashwin, T. S., Zhang, Y., Zhou, M., Ayalon, E., ... & Biswas, G. (2024, July). A first step in using machine learning methods to enhance interaction analysis for embodied learning environments. In International Conference on Artificial Intelligence in Education (pp. 3-16). Cham: Springer Nature Switzerland.
Fu, F. L., Wu, Y. L., & Ho, H. C. (2009). An investigation of cooperative pedagogic design for knowledge creation in web-based learning. *Computers & Education*, *53*(3), 550-562.
Gheller, F., Spicciarelli, G., Scimemi, P., & Arfé, B. (2023). The effects of noise on children's cognitive performance: A systematic review. *Environment and Behavior*, *55*(8-10), 698-734.
Gijlers, H., & de Jong, T. (2013). Using concept maps to facilitate collaborative simulation-based inquiry learning. *Journal of the learning sciences*, *22*(3), 340-374.
Godwin, K., & Fisher, A. (2011). Allocation of attention in classroom environments: Consequences for learning. In *Proceedings of the annual meeting of the Cognitive Science Society* (Vol. 33, No. 33).
Gupta, A., Jaiswal, R., Adhikari, S., & Balasubramanian, V. (2016). DAISEE: dataset for affective states in e-learning environments. *arXiv preprint arXiv:1609.01885*, 1-22.
Ho, D. T., & Cao, T. H. (2012, September). A high-order hidden Markov model for emotion detection from textual data. In *Pacific rim knowledge acquisition workshop* (pp. 94-105). Berlin, Heidelberg: Springer Berlin Heidelberg.
Järvelä, S., Kirschner, P. A., Panadero, E., Malmberg, J., Phielix, C., Jaspers, J., ... & Järvenoja, H. (2015). Enhancing socially shared regulation in collaborative learning groups: Designing for CSCL regulation tools. *Educational Technology Research and Development*, *63*, 125-142.
Liu, Z., & Wang, S. (2011, October). Emotion recognition using hidden Markov models from facial temperature sequence. In *International Conference on Affective Computing and Intelligent Interaction* (pp. 240-247). Berlin, Heidelberg: Springer Berlin Heidelberg.
Özhan, Ş. Ç., & Kocadere, S. A. (2020). The effects of flow, emotional engagement, and motivation on success in a gamified online learning environment. *Journal of Educational Computing Research*, *57*(8), 2006-2031.
Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational psychology review,* 18, 315-341.
Prasetio, B. H., Tamura, H., & Tanno, K. (2020). Deep time-delay Markov network for prediction and modeling the stress and emotions state transition. *Scientific Reports*, *10*(1), 18071.
Reisenzein, R., & Junge, M. (2024). Measuring the intensity of emotions. *Frontiers in Psychology*, *15*, 1437843.
Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161.
Shernoff, D. J. (2013). Optimal learning environments to promote student engagement.
TS, A., & Biswas, G. (2024, November). Relating students cognitive processes and learner-centered emotions: An advanced deep learning approach. In Proceedings of the 26th International Conference on Multimodal Interaction (pp. 575-584).
Vicario, C. M., & Newman, A. (2013). Emotions affect the recognition of hand gestures. *Frontiers in Human Neuroscience*, 7, 906.
Williams, T. F., Vehabovic, N., & Simms, L. J. (2023). Developing and validating a facial emotion recognition task with graded intensity. Assessment, 30(3), 761-781.
Xu, Y., Li, Y., Chen, Y., Bao, H., & Zheng, Y. (2023). Spontaneous visual database for detecting learning-centered emotions during online learning. *Image and Vision Computing*, 136, 104739.
Zimmerman, B. J. (1986). Becoming a self-regulated learner: Which are the key subprocesses?. *Contemporary educational psychology*, 11(4), 307-313.

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Evaluating Students' Scientific Inquiry Strategies in Large-Scale Digital Assessments

Namrata SRIVASTAVA[a*], Emma LINSENMAYER[b], Mario PIACENTINI[b] & Gautam BISWAS[a]
[a]*Vanderbilt University, USA*
[b]*OECD, France*
*namrata.srivastava@vanderbilt.edu

Abstract: As education shifts toward digital learning environments, new assessments are needed to evaluate not only what students know but also how they learn using available tools and resources. The OECD Programme for International Student Assessment (PISA) 2025 Learning in the Digital World (LDW) assessment addresses this challenge by engaging students in open-ended, interactive tasks that require them to apply learning content to solve inquiry-based problems. This paper focuses on two key scientific inquiry strategies assessed through LDW tasks: Control of Variable (CoV), which reflects how systematically students conduct inquiry experiments, and Deriving Relationship from Data (DRD), which involves interpreting relationships or patterns from the experimental results. We present a scoring rubric designed for fine-grained analysis of student performance, allowing for partial credit to be linked to degrees of applying and reasoning with the strategies to derive solutions to assigned problems. Using a learning-by-modeling task in the LDW framework, we apply this rubric to pilot data collected from 6,800 students across 63 countries. Our findings show that students' success with the CoV and DRD strategies is influenced not only by prior knowledge but also by how they engage with instructional phases and utilize digital and scaffolding tools during the task.

Keywords: computational problem-solving, scoring rules, digital learning environment

1. Introduction

As digital technologies become more prevalent in education, there is a growing recognition that assessments must evolve beyond traditional testing measures of static content knowledge (NRC, 2012; OECD, 2017; Vo & Simmie, 2025). Assessments need to evaluate how students learn, apply, and adapt their learning and understanding in interactive, real-world problem-solving contexts – key competencies emphasized globally as part of 21st century education reform (Chu et al., 2017; Griffin et al., 2012). The Organisation for Economic Co-operation and Development's (OECD) Programme for International Student Assessment (PISA) 2025 Learning in the Digital World (LDW) assessment marks a significant advance in this area. This innovative assessment immerses students in open-ended digital tasks, requiring them to engage with instructional content and demonstrate strategic problem-solving skills using digital tools to solve problems (OECD, 2023).

Unlike traditional assessments that rely on selected responses, LDW tasks are open-ended, interactive, and simulate authentic learning. Students build and debug computational artefacts, conduct experiments, analyze data, and reflect on progress. This shift aligns with a trend towards active learning (Prince, 2004) and simulation-based assessments that better capture complex inquiry and reasoning (Baker et al., 2016; Gobert et al., 2013). It embodies the concept of "learning by doing", allowing students to build and apply knowledge through experimentation, feedback, and exploration. Consequently, LDW marks a significant shift from

traditional approaches, emphasizing active processes over static recall. Students must demonstrate their development, integration of new information, testing ideas, and application of strategies while using digital tools.

This change in assessment design demands new interpretations of student responses. Traditional scoring methods that categorize answers as right or wrong in open-ended and interactive tasks, like those in LDW assessments, fail to capture partial understanding, emerging reasoning, and strategic problem-solving. Effective assessments should evaluate both students' final products and their processes. This highlights the need for scoring frameworks that provide final scores (summative assessment) and detailed feedback to foster critical thinking and problem-solving skills. Such feedback helps students and educators understand key learning processes and identify improvement areas (formative insight).

This paper introduces a scoring rubric for evaluating students' problem-solving strategies within an LDW computational problem-solving task. The rubric uses partial-credit scoring to capture the nuances of student reasoning and awards points that reflect their use of productive problem-solving methods to achieve goals. The goal is to acknowledge students for engaging in meaningful inquiry strategies and to gain insights into their problem-solving processes. We focus on two core strategies that are crucial to scientific inquiry and central to the LDW competency model: **Control of Variables (CoV)** and **Deriving Relationships from Data (DRD)**. CoV assesses how effectively students can design inquiry experiments that systematically change one independent variable while keeping other potential influencing factors constant to isolate the effect of the variable being tested. DRD evaluates how well students can interpret experimental results by analyzing generated data to identify patterns, correlations, and causal relationships between variables. We chose to focus on these strategies because both are essential for interpreting scientific processes and are consistently emphasized in science education frameworks as fundamental to inquiry and causal reasoning.

We tested our scoring rubric using data collected during a PISA LDW pilot study in 63 countries with 6800 students. Our findings show that students' success with CoV and DRD strategies is shaped not only by prior knowledge but also by how they engage with instructional phases and use available tools during the task.

2. Background

Inquiry-based learning (IBL) facilitates active exploration and investigation of topics that interest students (Pedaste et al., 2015). It encourages students to ask questions, gather evidence, draw conclusions, and construct knowledge, fostering deeper understanding and critical thinking skills similar to the practices followed by professional scientists (Keselman, 2003). This method is often regarded as supporting the development of problem-solving skills (Pedaste & Sarapuu, 2006). In a comprehensive review of the literature, Pedaste et al. (2015) proposed a synthesized framework for IBL consisting of five distinct inquiry phases: *Orientation, Conceptualization, Investigation, Conclusion*, and Discussion. These phases form a cyclical and iterative cycle of learning and problem solving and can be further divided into sub-phases. For instance, the Investigation phase is divided into three sub-phases: *Exploration, Experimentation*, and *Data Interpretation*, involving tasks such as generating hypotheses, planning and designing experiments, and analyzing data to draw conclusions.

A central aspect of the Experimentation and Data Interpretation sub-phases is the ability to design and conduct valid experiments. In science education, this centers on two crucial inquiry skills: **control of variables (CoV)** and **deriving relationships from data (DRD)** strategies, which are the focus of the current study. The CoV strategy refers to a student's ability to manipulate one independent variable at a time while keeping others constant to establish relationships between the independent and outcome variables. This skill is essential for ensuring that observed effects can be attributed to the variables being tested, rather than being confounded by uncontrolled factors (Chen & Klahr, 1999). A substantial body of research has explored how this strategy develops (Kuhn, 2010; Zimmerman, 2007) and how educators can best support its development (see Schwichow et al. (2016) for meta-analysis). However, findings remain mixed, and only a few empirical studies have examined how students enact this

strategy using process-level data. For instance, Schwichow et al. (2016) conducted a comprehensive meta-analysis of 72 intervention studies focused on CoV instruction, reporting an overall positive effect of teaching CoV (mean effect size = 0.61). However, they also found substantial variability across studies, influenced by factors such as the type of instructional support, the assessment method, and whether feedback or demonstrations were included in the intervention. These contrasting outcomes suggest that students' performance on CoV tasks may depend on additional contextual or behavioral factors, highlighting the need to go beyond static outcome measures and explore how students engage with tasks at the process level.

Complementing CoV is the DRD strategy, emphasizing students' ability to interpret data patterns and infer relationships between variables. Although DRD is a critical aspect of the inquiry process, it has attracted comparatively less attention in the literature. For instance, prior studies have shown that students struggle to apply DRD strategies such as recognizing relevant trends, distinguishing between causal and correlational patterns, and expressing findings through models or graphs (Donnelly-Hermosillo et al., 2020; Masnick & Klahr, 2003). These challenges highlight the need for targeted assessments that can capture students' developing inquiry skills.

Digital assessments like the PISA LDW offer significant potential for exploring students' understanding of variable reasoning and data interpretation skills. The interaction data, along with process data captured through digital logs, can provide deeper insights not only into whether students solve tasks correctly or incorrectly but also into how they approach, engage with, and navigate the investigative phases of inquiry. This study aims to provide a process-oriented perspective on how students develop inquiry skills in digital environments by analyzing these strategies through log-based analysis and rubric-aligned scoring.

3. Methodology

3.1 LDW Unit Description

To assess students' application of CoV and DRD strategies in a digital inquiry setting, we analyzed student interaction data for the "Increasing Tomato Yield" Unit within the PISA LDW framework. The Example Unit is a 30-minute interactive assessment task that follows the four-phase structure of LDW: *Show, Learn, Apply*, and *Reflect*.

1. **Intro and Show phase (Pre-test):** The unit begins with a static introductory page that outlines the unit's overall goals and illustrates a real-world scenario related to variable relationships. This is followed by the Show phase, which includes four pre-test items designed to assess students' prior understanding of core concepts. These items measure students' ability to design controlled experiments, interpret variable relationships in graphical form, and draw inferences from visual data. They establish a baseline before students engage in scaffolded learning activities.
2. **Learn phase:** During this phase, students receive guided instruction from a virtual tutor as they complete a series of scaffolded tasks. These tasks help familiarize students with the experimental interface and provide hands-on experience using the CoV strategy to design controlled experiments. Additionally, students learn to apply DRD strategies, such as generating graphs to interpret their experimental results. Scaffolding is provided through example solutions and correct answers after each task. This approach allows students to practice selecting variables, conducting controlled experiments, and matching outcome patterns to graphs. The goal of this phase is to ensure that all students have the foundational knowledge needed to engage productively with the remaining tasks in this unit.
3. **Apply phase (Comprehensive Challenge task):** In this phase, students encounter an open-ended task in a new but related context, where they must independently apply the strategies they learned earlier. This phase focuses on three key objectives:
   i) Designing valid experiments using the CoV strategy.

ii) Applying the DRD strategy to choose the graph that accurately represents the relationship between variables.
iii) Identifying the subset of experiments that support the relationships between independent and outcome variables.

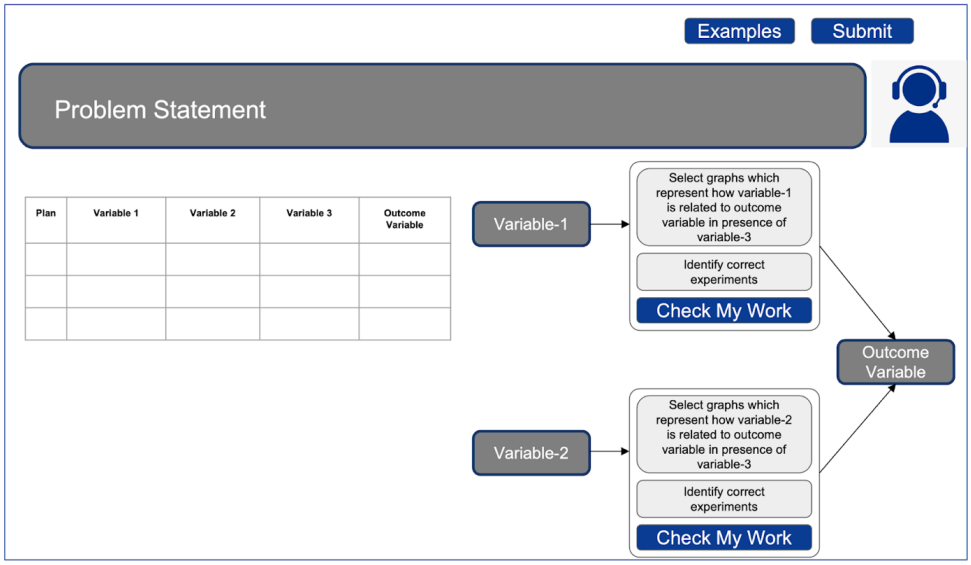


Figure 1. A prototype of the Challenge task.

In contrast to the Learn phase, students do not receive real-time guidance or access to final solutions during the Challenge task. However, they can refer back to the annotated solutions from the Learn phase for help. Additionally, a "Check My Work" feature is available that provides feedback on whether the graphs and experiments they have chosen accurately represent the underlying relationships (see Figure 1). The overall assessment adopts “a low threshold, high ceiling” approach that enables all students to make some progress while allowing more advanced learners to demonstrate deeper reasoning and mastery of problem-solving using the CoV and DRD strategies.

4. **Reflect phase:** This phase consists of three brief self-assessment tasks. First, students evaluate whether they could not accomplish, partially accomplished, or fully accomplished each of the three sub-goals in the Challenge task. Second, they reflect on their emotional state during the problem-solving (for example, feelings of confusion or boredom). Third, they indicate how difficult they found the task to be, and the amount of effort required to solve it. These reflections provide valuable insights into students' self-regulation and perceived performance.

While each phase of the task contributes to the overall learning experience, this paper specifically evaluates student performance during the Challenge task. In this task, students independently demonstrate their problem-solving strategies. Additionally, students’ responses to the Show phase questions were scored to assess their prior knowledge, which is included in the analysis to provide context for their performance in the Challenge task.

3.2 Rubric Development

The framework for the challenge task was developed using a partial-credit system that evaluated students' use of the CoV and DRD strategies. Instead of assigning binary scores, we awarded points for both complete and partially correct responses.

The rubric is structured around three sub-goals of the Challenge task: (1) selecting correct relationships and graphs (DRD), (2) identifying relevant experiments (CoV), and (3) assessing the quality and coverage of the designed experiments (CoV). For each sub-goal, the rubric criteria were defined to reflect levels of correctness, completeness, and strategic execution. A summary of these components, illustrated with an example task – “Increasing

Tomato Yield," which includes three variables of interest (sunlight intensity, soil type, and amount of water used) is provided in Table 1.

To assess DRD strategies, the rubric evaluated whether students correctly identified if each independent variable (sunlight intensity, soil type, and amount of water used) influenced the outcome (tomato yield) and selected the appropriate graph to represent that relationship. Partial credit was awarded to students who recognized the general direction of the relationship but overlooked more nuanced patterns. Additionally, students received points for selecting a subset of experiments that clearly demonstrated the effect of a variable, reflecting their skill in interpreting and choosing informative data.

For CoV, the rubric assessed both the quality and coverage of students' experiments. Quality was evaluated based on whether students systematically held one variable constant while varying another, which indicates strategic planning. Coverage was determined by the number of unique experiments conducted by the students. Full credit was awarded for complete, well-controlled experiments across both soil conditions, while partial credit was given for incomplete or partially controlled experimental designs.

Table 1. *Rubric for scoring students' inquiry tasks*

Goal	Sub-goals	Example ***Design experiments to investigate how light and water affect the number of tomatoes produced, and check whether the relationships depend on soil type.***	Points Awarded	Max points
DRD	Choosing the correct relationship and graph (variable-1)	Tomato yield is not impacted by soil type (partial credit for choosing "no")	1	8
		Tomato yield increases from low to normal sunlight intensity conditions, but drops for high light intensity conditions (e.g., non-linear pattern)	2	
	Choosing the correct relationship and graph (variable-2)	Tomato yield is influenced by the amount of water used (partial credit for choosing "yes")	1	
		Tomato yield increases linearly with the amount of water used for a soil type	2	
	Choosing the correct experiments (variable-1)	Selecting experiments that show the correct water use-tomato yield relationship	1	
	Choosing the correct experiments (variable-2)	Selecting experiments that show the correct sunlight intensity-tomato yield relationship	1	
CoV	Coverage of experiments	3–5 unique experiments conducted; may cover only part of a variable set	1	3
		6–11 experiments conducted; possibly covers 2 conditions for 1 variable	2	
		12 experiments conducted; covers all combinations of 2 variables while holding one constant	3	
	Quality of experiments	**Unpaired Incomplete:** Held water use or sunlight intensity constant for only one soil condition and only one other variable	1	3
		Unpaired Complete: Held water or light constant for only one soil condition, but did this for both light and water	2	
		CoV Complete: Exhaustively held soil type constant while varying both water and light	3	

3.3 Data Collection

This data was collected as part of an OECD PISA 2025 LDW pilot study. The pilot was administered in 63 countries and involved 6,800 students. Students completed a series of LDW

prototype units under standardized assessment conditions using a digital platform that automatically recorded detailed interaction log data. Ethics approval and data handling protocols adhered to OECD standards and local requirements in the participating countries.

For this paper, we specifically focus on data from one prototype unit – the Increasing Tomato Yield Unit – designed to assess students' ability to conduct controlled experiments and derive relationships from data. The dataset contains a varying number of log events across countries, with an average of 34,275 events (Minimum = 16,536 events, Maximum = 91,462 events, and standard deviation = 12,465 events). Regarding student participation, the number of students ranged from 52 to 239, with a mean of 108 students per country.

We extracted students' interaction logs from the Show phase and the Apply phase to answer questions. We preprocessed them to capture prior knowledge (via pre-test questions) and to analyze the development and application of CoV and DRD strategies in the Challenge task (Apply phase).

3.4 Data Preprocessing

We first transformed the raw interaction log data from the Challenge task into a structured format to facilitate rubric-based scoring. This process involved extracting key events from each student's log file, including (1) experimental trials (i.e., selected values for sunlight intensity, soil type, and the amount of water used); (2) corresponding outcome values (e.g., number of tomatoes); (3) chosen graphs; (4) subsets of experiments selected to justify variable relationships; and (5) the use of the "Check My Work" feature. We then parsed and aggregated these events to identify unique experiments and evaluate whether students systematically varied one variable while keeping others constant, which is a key criterion for scoring the CoV strategy. For scoring DRD, we analyzed students' graph selections and assessed whether the chosen subsets of experiments provided sufficient evidence for determining the variable relationships, and if the chosen relation was correct. This processing was conducted using custom scripts developed in Python. The final dataset included individual student scores for each rubric item, along with metadata such as country and pre-test scores.

3.5 Data Analysis

Following preprocessing and rubric-based scoring, we conducted descriptive and inferential analyses to examine student performance patterns during the Challenge task and identify factors that may influence performance, such as prior knowledge and task engagement.

First, we calculated summary statistics, including means, standard deviations, and score distributions for total performance scores in the Challenge task, individual DRD and CoV components of the scores, and pre-test scores across all students to identify overall trends. Next, to analyze learning transitions, we divided the students into low and high-performance bands using a median split for both pre-test scores and Challenge task scores. This categorization resulted in four transition groups:

- Group-1: LowPre-test → LowTotal-score
- Group-2 HighPre-test → LowTotal-score
- Group-3: LowPre-test → HighTotal-score
- Group-4: HighPre-test → HighTotal-score.

These groups allowed us to explore how students' performance changed from the pre-test to the final task. For instance, we identified how many low pre-test scorers' performances improved in the final task (Group 3) and how many high pre-test scorers' performances declined (Group 2) in the final task. Finally, we investigated the role of potential mediating factors such as students' engagement during the Learn phase and Challenge task. Specifically, we compared metrics such as the amount of time spent in each phase and the use of digital tools (e.g., graph or feedback buttons) across the four groups (Groups 1–4). To ensure robust analysis of timing-related metrics, we applied a 98% winsorization to mitigate the influence of spurious outliers and removed any missing values to maintain data quality.

4. Results and Findings

4.1 Descriptive statistics

Challenge task: Figure 2 (left panel) shows the distributions of students' total scores and their individual DRD and CoV scores (bottom panel) during the Challenge task. In terms of students' overall performance in the task, we found that most students scored between 0 and 5 (Mean = 3.62, SD = 3.5, out of a maximum of 14). Additionally, the right-skewed distribution of the total score suggests that only a small number of students achieved high scores. At the same time, a large proportion demonstrated either partial understanding or incomplete execution of strategies.

When examining DRD and CoV scores separately, we found that most students struggled with the DRD strategy. The average DRD score was relatively low (Mean = 1.57, SD = 2.23, out of a maximum of 8), with many students (50%) scoring 0 points. These findings suggest that students had difficulty recognizing patterns in the experimental table, selecting correct graphs, or determining whether a variable had an effect. In contrast, students performed relatively better on the CoV strategy (Mean = 2.04, SD = 1.59, out of a maximum of 6). While a few students achieved full points for implementing the CoV strategy, a larger proportion scored between 1 and 3 points, demonstrating partial understanding of designing controlled experiments.

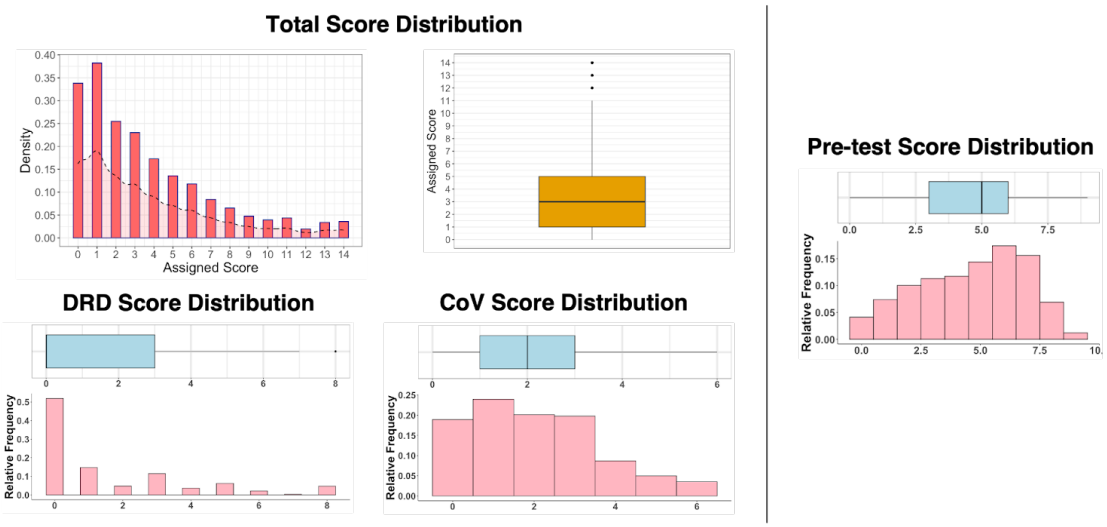


Figure 2. Distribution of total scores for the Challenge task (top-left panel), individual DRD score and CoV score distribution (bottom-left panel), and pre-test scores (right panel)

Pre-test scores: Figure 2 (right panel) displays the distribution of students' pre-test scores. On average, students scored 4.59 points (SD = 2.26) out of a maximum of 9 points, indicating that most approached the Challenge task with some preliminary understanding of CoV and DRD concepts.

4.2 Relationship between prior knowledge and Challenge task performance

Figure 3 presents two complementary views of the relationships among pre-test scores, total score in the Challenge task, and individual CoV and DRD scores. Treating the score as a ranked variable, we conducted a Spearman rank-order correlation to determine the relationship between students' prior knowledge and their total score in the Challenge task.

Correlation Analysis: The correlation matrix (left panel) reveals a strong positive correlation between CoV and DRD scores (ρ = 0.62), indicating that students who effectively designed control experiments were also more likely to accurately identify and represent relationships from the data. We also found moderate correlations between students' pre-test scores and their performance on the Challenge task: CoV (ρ = 0.51), DRD (ρ = 0.47), and total scores (ρ = 0.54). These results suggest that students with greater prior knowledge were generally more successful in applying inquiry strategies during the task. However, the moderate strength of these correlations implies that prior knowledge alone does not fully explain students' success in applying CoV and DRD strategies, reinforcing the importance of process-based assessment.

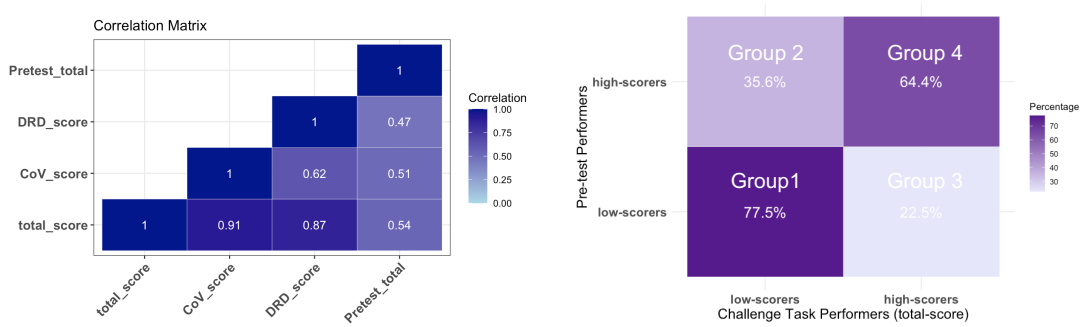


Figure 3. Correlation Matrix (left panel) and Score Distribution Plot across pre-test score and total score (right panel)

Transition groups: The heatmap in Figure 3 shows the distribution of students in four performance transition groups (Groups 1-4 defined in Section 3.5). Results indicate that students with higher prior knowledge were more likely to achieve high scores in the Challenge task, with 64.4% of high pre-test scorers remaining in the high-performance band (Group 4). In contrast, 77.5% of students with low prior knowledge stayed in the low-performance band (Group 1), suggesting limited improvement. Additionally, 22.5% of students with low prior knowledge performed better in the Challenge task (Group 3), while 35.6% of high prior knowledge students showed declining performance (Group 2). These transitions highlight variability in students' performance and raise questions about factors affecting their success or struggle.

4.3 Exploratory Analysis

Table 2 summarizes the CoV and DRD scores, time spent in task phases, and digital tool usage across four transition groups. Kruskal-Wallis tests showed significant differences among groups in score and time metrics ($p < .001$), confirmed by post-hoc analyses. Chi-squared tests also indicated significant differences in tool usage (Graph-button use: $\chi^2(3) = 2507.60$, $p < .001$; Check-My-Work button: $\chi^2(3) = 688.09$, $p < .001$). Students who improved (Group-3: Low→High) and those who remained high performers (Group-4: High→High) had higher CoV and DRD scores, spent more time in the Learn and Challenge phases, and utilized digital tools such as the graph and check-my-work buttons more frequently. Meanwhile, students who remained low performers (Group 1) showed the least engagement across all dimensions.

Table 2. *Summary of score-metrics, time-duration metrics, and tool-use metrics for the four transition groups*

	Score-related metrics		Time-related metrics (in seconds)			Tool-related Metrics	
Transition Group	Average CoV	Average DRD Score (SD)	Average Time spent in Show	Average Time spent in	Average Time spent in	% who used	% who used Check-

	Score (SD)		Phase (SD)	Learn Phase (SD)	Chal-lenge task (SD)	Graph Button	My-Work Button
Group-1: Low→Low	0.99 (0.89)	0.18 (0.45)	219 (177.8)	194 (193.0)	67.5 (75.4)	30.1%	28.6%
Group-2: High→Low	1.50 (1.05)	0.29 (0.55)	298 (175.8)	266 (194.8)	91.8 (75.0)	52.3%	37.3%
Group-3: Low→High	2.91 (1.15)	2.79 (1.74)	505 (96.6)	570 (201.0)	223 (109.3)	91.3%	62.3%
Group-4: High→High	3.74 (1.29)	4.10 (2.34)	539 (73.1)	655 (176.3)	271 (109.6)	96.2%	62.5%

5. Discussion, Limitations and Future Work

In this study, we examined how students applied two key scientific inquiry strategies – CoV and DRD – during a digitally administered assessment. Combining rubric-based scoring with log data provided insights not only into student outcomes but also into how engagement and tool use shaped performance.

Students generally performed better on CoV than DRD, aligning with prior evidence that CoV is more teachable and frequently emphasized (Schwichow et al., 2016). DRD remained challenging, with nearly half of students scoring zero, underscoring persistent difficulties in recognizing patterns and casual relationships from data (Donnelly-Hermosillo et al., 2020; Masnick & Klahr, 2003). Importantly, our transition analysis showed that performance was not determined by prior knowledge alone: while many students remained within their initial performance bands, a subset improved or declined significantly, highlighting diverse learning trajectories.

Engagement emerged as a key mediating factor. Students who invested more time in the Learn and Challenge phases generally outperformed their peers, particularly those who remained in the low-performance group. This suggests that scaffolded practice, where students first observe worked examples and then apply strategies independently, supports deeper understanding and transfer of CoV and DRD skills. The use of digital tools also differentiated higher-performing groups: frequent interaction with the graphing feature promoted data interpretation, while the “Check-My-Work” button provided timely formative feedback, enabling students to refine experimental designs. These results align with prior research indicating that structured opportunities for practice, combined with feedback, enhance inquiry processes (Hattie & Timperley, 2007; Schwichow et al., 2016).

Despite these insights, several limitations should be acknowledged. First, our analyses were exploratory and limited to two strategies within a single LDW unit; task-specific effects, therefore, constrain generalizability. Future work should expand to additional inquiry components (e.g., hypothesis generation, error analysis), apply sequential or temporal methods (e.g., Markov models) to capture strategy development, and explore cross-national differences in performance. Second, engagement was proxied primarily by time, which may also include off-task behaviors. Richer interaction measures are needed to distinguish productive form unproductive engagement. Third, while our rubric was grounded in theory, formal validation of its psychometric reliability and transferability across contexts remains an important next step. Lastly, this study did not explore motivational or self-regulated learning (SRL) factors. Future research could investigate how students’ motivation, goals, and self-regulated learning (SRL) behaviors contribute to performance differences.

Overall, this study demonstrates how open-ended digital assessments, combined with process-level analysis, can provide rich insights into not only what students learn but also how they engage with and learn core scientific inquiry strategies. These findings offer valuable implications for designing learning environments that more effectively scaffold students’ use of CoV and DRD strategies in authentic, inquiry-driven contexts.

Acknowledgements

We thank the OECD team, participating researchers, students, and teachers for their contributions and support in this research. This work was partially supported by funding from the OECD.

References

Chen, Z., & Klahr, D. (1999). All Other Things Being Equal: Acquisition and Transfer of the Control of Variables Strategy. *Child Development*, *70*(5), 1098–1120. https://doi.org/10.1111/1467-8624.00081

Chu, S. K. W., Lee, C. W. Y., Notari, M., Reynolds, R. B., & Tavares, N. J. (2017). *21st Century Skills Development Through Inquiry-Based Learning: From Theory to Practice* (1st ed. 2017). Springer Singapore : Imprint: Springer. https://doi.org/10.1007/978-981-10-2481-8

Donnelly-Hermosillo, D. F., Gerard, L. F., & Linn, M. C. (2020). Impact of graph technologies in K-12 science and mathematics education. *Computers & Education*, *146*, 103748. https://doi.org/10.1016/j.compedu.2019.103748

Griffin, P. E., McGaw, B., & Care, E. (Eds.). (2012). *Assessment and teaching of 21st century skills*. Springer.

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Keselman, A. (2003). Supporting inquiry learning by promoting normative understanding of multivariable causality. *Journal of Research in Science Teaching*, *40*(9), 898–921. https://doi.org/10.1002/tea.10115

Kuhn, D. (2010). *The Wiley-Blackwell Handbook of Childhood Cognitive Development* (1st ed.). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781444325485

Masnick, A. M., & Klahr, D. (2003). Error Matters: An Initial Exploration of Elementary School Children's Understanding of Experimental Error. *Journal of Cognition and Development*, *4*(1), 67–98. https://doi.org/10.1080/15248372.2003.9669683

NRC (Ed.). (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. The National Academies Press.

OECD. (2017). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving*. OECD. https://doi.org/10.1787/9789264281820-en

OECD. (2023). *Innovating Assessments to Measure and Support Complex Skills* (N. Foster & M. Piacentini, Eds.). OECD. https://doi.org/10.1787/e5f3e341-en

Pedaste, M., Mäeots, M., Siiman, L. A., De Jong, T., Van Riesen, S. A. N., Kamp, E. T., Manoli, C. C., Zacharia, Z. C., & Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review*, *14*, 47–61. https://doi.org/10.1016/j.edurev.2015.02.003

Pedaste, M., & Sarapuu, T. (2006). Developing an effective support system for inquiry learning in a Web-based environment. *Journal of Computer Assisted Learning*, *22*(1), 47–62. https://doi.org/10.1111/j.1365-2729.2006.00159.x

Schwichow, M., Croker, S., Zimmerman, C., Höffler, T., & Härtig, H. (2016). Teaching the control-of-variables strategy: A meta-analysis. *Developmental Review*, *39*, 37–63. https://doi.org/10.1016/j.dr.2015.12.001

Vo, D. V., & Simmie, G. M. (2025). Assessing Scientific Inquiry: A Systematic Literature Review of Tasks, Tools and Techniques. *International Journal of Science and Mathematics Education*, *23*(4), 871–906. https://doi.org/10.1007/s10763-024-10498-8

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Developmental Review*, *27*(2), 172–223. https://doi.org/10.1016/j.dr.2006.12.001

A Secured Knowledge Management System to enhance the User Experience in Engineering Education

K. RAJESWARIa*, N. VIVEKANANDANb, SAKSHI KULKARNIc, JANHAVI MALId & VINAYAK SHETEe

[a]*Faculty Mentor, Pimpri Chinchwad College of Engineering, Pune, India*
[b]*Faculty Mentor, Pimpri Chinchwad College of Engineering, Pune, India*
[c]*Student, Pimpri Chinchwad College of Engineering, Pune, India*
[d]*Student, Pimpri Chinchwad College of Engineering, Pune, India*
[e]*Student, Pimpri Chinchwad College of Engineering, Pune, India*
*kanna.rajeswari@pccoepune.org

Abstract: With the increasing complexity of institute management systems, there is a growing need for an integrated digital solution that streamlines academic, administrative, and extracurricular activities. The system is a unified institute management application designed to simplify processes for students, faculty, and administrators. The platform offers academic tools such as course registration, attendance tracking, and examination applications, along with administrative features like fee payment and digital certificate requests. Additionally, it enhances student engagement through event management, resource management, and knowledge sharing forums. It also provides data driven dashboards and analytics, enabling users to track progress and institutional efficiency. This paper explores the functional architecture, key features, and impact of the system, highlighting its role in improving efficiency and user experience in higher education. Future enhancements and the potential of such systems in digitalizing institutes are also discussed.

Keywords: Institute management system, digital integration, academic tools, administrative efficiency, student engagement, event management, knowledge sharing, data analytics, higher education, Insightify, digital transformation.

1. Introduction

In the rapidly evolving landscape of higher education, institutes are facing increasing challenges in managing academic, administrative, and extracurricular activities efficiently. Traditional methods of institute management often involve fragmented systems that require students, faculty, and administrators to navigate multiple platforms for different tasks. This lack of integration leads to inefficiencies, miscommunication, and delays in critical processes such as course registration, attendance tracking, fee payments, and event management.

To address these challenges, Insightify is designed as an all-in-one institute management application that serves as a centralized digital platform for students, faculty, heads of departments (HODs), and administrators. By integrating multiple features into a single application, Insightify streamlines essential institute functions such as academic management, attendance tracking, course registration, faculty-student interactions and knowledge sharing.

2. Literature Review

The development of institute management systems has been a topic of extensive research in recent years. Various studies have proposed and evaluated different models, frameworks, and implementations that aim to enhance the efficiency of academic institutions.

2.1 Knowledge Management in Higher Education Institutions

Research highlights the role of knowledge management (KM) in universities, emphasizing the importance of structured KM systems for academic and administrative efficiency. It discusses the impact of KM frameworks on improving decision-making, student engagement, and faculty collaboration.

2.2 Impact of Digital Platforms on University Administration

This study examines the implementation of digital university management systems that integrate academic, financial, and administrative services into a unified platform. It compares traditional systems with modern cloud-based solutions and evaluates their scalability and security.

2.3 A Framework for University Knowledge Management System

A proposed framework highlights the importance of integrating KM strategies with university management platforms. The study suggests using AI-driven recommendation systems to enhance personalized learning experiences and streamline administrative processes.

2.4 Technology-Driven University Management Systems

This research focuses on the technological advancements in university management, emphasizing cloud computing, AI, and big data analytics. The study discusses the implementation of web-based platforms that facilitate course registration, examination management, and student performance tracking.

2.5 Student-Centric Digital Transformation in Higher Education

The paper explores how digital transformation impacts student experiences in higher education. It discusses the role of mobile applications and dashboards in providing realtime academic insights, attendance tracking, and performance analytics.

3. Methodology

3.1 Introduction to Methodology

The efficient management of academic and administrative processes is a significant challenge in modern institutes. Institutions require an integrated, scalable, and secure system to streamline student services, faculty interactions, and administrative operations. This research paper presents a comprehensive university management system designed to centralize various academic, financial, and extracurricular activities into a single digital platform.

This section outlines the methodology employed in the design, development, and evaluation of the system. The methodology ensures a structured approach that aligns with user requirements while optimizing system performance, security, and scalability. The research integrates both theoretical frameworks and practical implementations, offering a holistic understanding of the architectural, functional, and usability aspects of this system. To achieve these objectives, the methodology is structured into several key components:

- Requirement Analysis to identify the essential features based on user needs.
- System Architecture Design to establish a scalable and modular structure.
- Data Collection and Processing methods to handle user information securely and efficiently.
- Development and Optimization Techniques for ensuring high performance and usability.
- Testing and Evaluation Strategies to validate functionality and user experience.

- Deployment Planning and Maintenance to ensure long-term sustainability and updates.

4. System Design and Model

The system design follows a modular and scalable approach to accommodate the diverse functionalities required by students, faculty, and administrators.

4.1 Overview

The system follows a three-tier architecture:

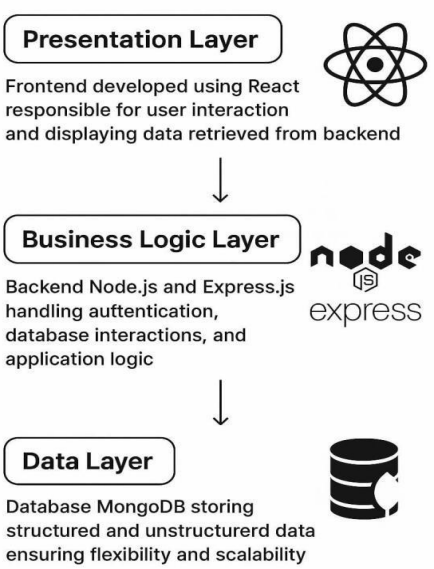


Figure 1. System Design

4.2 Proposed Solution: System Overview

The system is designed to be an all-in-one application offering distinct features for students, faculty, and administrators. The system integrates various modules to ensure smooth management of academic and administrative processes. The primary components include:

- Student Dashboard: Provides access to course registration, attendance tracking, examination applications, result viewing, timetable management, fee payments, and hostel facilities.
- Faculty Dashboard: Facilitates course allocation, attendance marking, grading, result generation, and scheduling of classes.
- Admin Dashboard: Oversees digital applications, meeting bookings, certificate issuance, and centralized academic records

4.3 System Architecture

The architecture is built on the MERN stack, comprising React.js for the frontend, Node.js with Express.js for the backend, and MongoDB as the database. Users interact with the system through a responsive React interface, which communicates with backend APIs to perform actions like login, data submission, and content retrieval. The backend handles logic, user authentication via JWT/OAuth, and role-based access control. MongoDB stores all academic and administrative data, including user profiles, timetables, and attendance. Cloud deployment ensures scalability, while monitoring and testing tools like Postman and Prometheus maintain system reliability. This architecture enables secure, modular, and efficient data flow across all components.

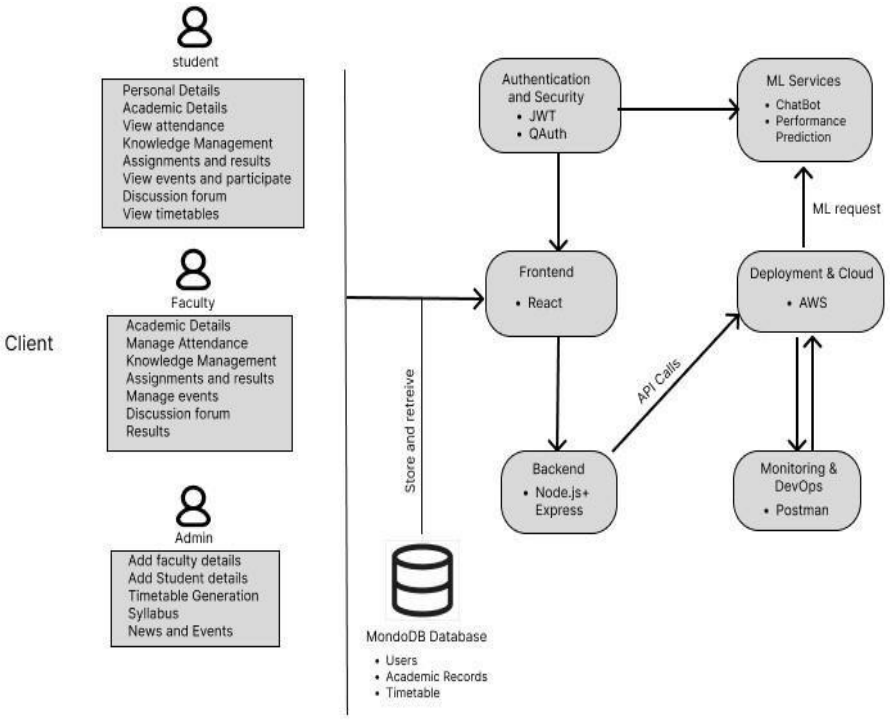


Figure 2. System Architecture

4.4 Features and Functionalities

4.4.1 Student Functionalities

The Student module of this system includes the following feactures:

- Academic functions: Course registration, timetable viewing, examination applications.
- Administrative functions: Fee payments, certificate applications, leave requests.
- Canteen and hostel services: Online food ordering, seat reservations, hostel applications.
- Community engagement: Event registrations, discussion forums, knowledge sharing.

4.4.2 Faculty Features

The faculty module of this system includes the following features:

- Course management: Viewing and managing assigned courses.
- Assessment management: Attendance, internal evaluations, result generation.
- Administrative tasks: Leave applications, meeting bookings, student guidance.
- Community engagement: Event management, knowledge sharing, discussion forums.

5. Results

Table 1. *Results*

Previous systems	Hurdles overcame	System enhancement
Manual course registration	Time delays, errors	Online academic process
Paper-based attendance	Lack of tracking	Digital attendance and results

No unified student portal	Scattered systems	Centralized dashboard
Manual search for free resource availability	Scheduling issues	Convenient and proper resource usage
Limited communication	Low engagement	Forums and event management
Offline fee/certificate process	Manual delays	Online payments and applications
No KMS or academic resources	Limited knowledge sharing	Integrated Knowledge Management System (KMS)
No doubt resolution platform	Student queries went unanswered	Doubt-raising and discussion forum



Figure 3. Results

6. Benefits and Impact

- Efficiency: Automates manual processes, reducing administrative workload.
- Transparency: Improves information flow between students, faculty, and administration.
- Accessibility: Enables remote management of academic functions.
- Engagement: Encourages knowledge sharing and collaboration through discussion forums.

7. Conclusion

The system is composed to revolutionize institute management by integrating all essential academic, administrative, and extracurricular functionalities into a single platform. By

leveraging modern technology, the system enhances efficiency, reduces administrative overhead, and fosters a collaborative learning environment.

With AI-driven analytics and real-time data processing, institutes can optimize course scheduling, monitor student performance, and facilitate personalized learning. The automation of administrative tasks, such as fee management and event coordination, will help institutions allocate resources more effectively and improve operational efficiency.

References

D.-G. J. Dei and T. B. van der Walt, ”Knowledge Management Practices in Universities: The Role of Communities of Practice,” *International Journal of Knowledge Management, vol. 12, no. 4, pp. 1–14, 2016*. DOI: 10.4018/IJKM.2016100101.

Mouna Ben Chouikha Zouari, Salem Ben Dhaou Dakhli, "A Multi-Faceted Analysis of Knowledge Management Systems*", Procedia Computer Science*, *2018*,

K. Iskandar, M. I. Jambak, R. Kosala, and H. Prabowo, ”Current Issue on Knowledge Management System for Future Research: A Systematic Literature Review,” *Procedia Computer Science, vol. 116, pp. 68–80, 2017*. DOI:10.1016/j.procs.2017.10.011.

Ahmed, Professor & Sutton, Michael. (2017). Gamification, serious games, simulations, and immersive learning environments in knowledge management initiatives. *World Journal of Science, Technology and Sustainable Development.* 14. 78-83.10.1108/WJSTSD-02-2017-0005.

Kudozia, R. , Kudozia, C. and Owusu-Donkor, D. (2023) Knowledge Management Infrastructure in Organisational Knowledge Management Processes—A Case of Aquantuo LLC. *Open Journal of Business and Management, 11, 1248-1270*. doi: 10.4236/ojbm.2023.113070

T. Gao, Y. Chai, and Y. Liu, ”A Review of Knowledge Management about Theoretical Conception and Designing Approaches,” *International Journal of Crowd Science, vol. 2, no. 1, pp.42–52,2018*. DOI: 10.1108/IJCS-08-2017-0023.

Challenge-Based Learning and ICT Integration in Mechanical Engineering Education: A Case Study on Mechanical Engineering Design Courses

Vivekanandan N [a*], Rajeswari K[b] & Yuvraj Kanna N V[c]
[a]*Associate Professor, Department of Mechanical Engineering, Pimpri Chinchwad College of Engineering, Pune, India*
[b] *Professor, Department of Mechanical Engineering, Pimpri Chinchwad College of Engineering, Pune, India*
[c]*Data Science for Business, University of Stirling, Stirling, United Kingdom*
*n.vivekanandan@pccoepune.org

Abstract: This paper presents an innovative, practice-driven approach to mechanical engineering education by integrating Challenge-Based Learning (CBL) and digital tools in two core courses: Mechanical Vibrations and Acoustics (MVA) and Mechanical System Design (MSD). In MSD, a focused design hackathon was introduced, engaging students in real-world, interdisciplinary challenges in pressure vessels, gearboxes, and material handling systems. Students employed ICT tools such as MATLAB, AutoCAD, and ChatGPT to simulate, model, and present integrated design solutions. In MVA, simulation-based and video-driven assessments enabled students to visualize vibratory phenomena and communicate engineering concepts effectively.
Analysis of course-end survey data from 82 students revealed strong outcomes: over 85% reported high achievement of course outcomes, especially in practical application and integration of design standards. Program Outcomes such as teamwork, ethics, and societal awareness were rated as significantly enhanced, alongside Program Specific Outcome 01 (PSO01)—the ability to apply mechanical design principles and codes to real-world problems. Students expressed high satisfaction with the clarity of learning objectives, relevance of assessments, and the creative, collaborative learning environment.
These findings underscore the pedagogical effectiveness of CBL and ICT-enhanced strategies in fostering deep learning, professional skills, and engineering judgment. The approach offers a scalable and replicable model aligned with NEP 2020 and ABET standards, preparing students for complex, multidisciplinary challenges in modern engineering practice.

Keywords: Challenge-Based Learning, ICT Tools, Mechanical Engineering, Hackathon, Vibrations, System Design, AI in Education

1. Introduction

Engineering education is seeing a revolutionary paradigm shift due to the imperative of providing students with more interesting, application-driven, and trans disciplinary methods. Traditional practices based on theory teaching and tests are unable to prepare students with hands-on design thinking, problem-solving capabilities, and integration across domains. Challenge-Based Learning (CBL), supplemented with ICT technologies and online platforms, has been found to be an effective pedagogical model in filling these lacunae (Chien & Tsai, 2021; Savery, 2015). This paper examines the use of CBL frameworks in two mechanical engineering courses—Mechanical Vibrations and Acoustics (MVA) and Mechanical System Design (MSD)—in Pimpri Chinchwad College of Engineering (PCCoE), Pune. These courses used novel formative assessments, such as simulation-based modeling, design hackathons,

and video-based product storytelling, with the aim of developing technical knowledge as well as professional skills among undergraduate students.

2. Literature Review

Challenge-Based Learning (CBL) and Problem-Based Learning (PBL) have been widely recognized as effective pedagogical models in engineering education. Kolmos et al. (2016) emphasized that curricula should address authentic problems to foster systems thinking and collaboration, while Brereton et al. (2015) showed design challenges stimulate creativity and mirror professional practice. Savery (2015) distinguished between surface and deep learning, noting that open-ended challenges promote interdisciplinary problem-solving. Extending this view, Chien and Tsai (2021) found that CBL enhances motivation and computational thinking, leading to improved outcomes. The role of information and communication technologies (ICT) in engineering education has been equally significant. Ferguson (2012) highlighted learning analytics as a way to track engagement and personalize instruction. Nouri et al. (2023) showed simulation tools increase technical confidence and communication, while Shanbhag and Al-Ammar (2023) demonstrated that virtual design studios enhance creativity and accuracy. Dillenbourg et al. (2009) also confirmed that computer-supported collaborative learning improves teamwork and communication. Together, these works emphasize the importance of tools such as MATLAB, Python, and AutoCAD in fostering technical and collaborative skills.

Artificial intelligence (AI) tools are now complementing traditional pedagogy. Zawacki-Richter et al. (2019) reviewed AI in education, noting benefits in feedback and learner autonomy, while VanLehn (2011) showed intelligent tutoring systems can approach human tutoring effectiveness. More recent studies highlight ChatGPT's value: Yeo and Quek (2022) found AI-generated feedback boosts participation and performance, Wang and Han (2021) linked chatbot-supported environments to improved metacognition, and Bravo and Cruz-Bohorquez (2024) discussed chatbots' impact on engineering education. Kumar (2021) further noted the potential of educational chatbots for project-based learning. Research on active and cooperative learning confirms its advantages. Prince (2004) showed interactive tasks enhance understanding compared to lectures, and Freeman et al. (2014) demonstrated that active learning reduces failure rates while improving STEM performance. Felder and Brent (2016) advocated cooperative strategies for engagement and retention, while Boud and Falchikov (2007) emphasized sustainable assessment for lifelong learning.

Design thinking and reflective practice also underpin modern engineering pedagogy. Adams et al. (2003) showed iterative projects strengthen reflective thinking, while Biggs and Tang (2011) stressed constructive alignment between goals, activities, and assessment. Schön (1983) argued that reflection-in-action is essential for professional growth, reinforcing the role of project-based learning. Finally, the rise of simulation and digital twins demonstrates the importance of ICT-driven training. Hussain and Kamarudin (2024) highlighted how interactive simulations deepen understanding in mechanical contexts, and Kritzinger et al. (2018) classified digital twins as critical for bridging theory and practice. Similarly, Lee et al. (2024) and Calvani et al. (2021) reinforced the importance of computational thinking and digital competence in higher education.

Collectively, these studies underscore that CBL, ICT, AI, and active learning approaches converge toward preparing students with technical expertise, collaboration, and reflective practice—competencies central to modern engineering education.

3. Methodology and Course Context

This study was conducted in two undergraduate mechanical engineering courses—Mechanical Vibrations and Acoustics (MVA) and Mechanical System Design (MSD)—at Pimpri Chinchwad College of Engineering (PCCoE), Pune, with ~60 students each. Both courses followed NEP 2020's experiential learning approach, integrating simulations,

hackathons, and digital storytelling. In MVA, students modeled real-world vibration scenarios (e.g., bridge oscillations, vehicle suspensions) using MATLAB/Python and analyzed responses of single-degree-of-freedom systems. A second task required short promotional videos on vibration/noise-control products, emphasizing professional communication. In MSD, a design hackathon engaged teams in challenges such as gearbox layouts, pressure vessel sizing, and conveyor systems. Evaluation was based on technical accuracy, innovation, and teamwork under time constraints.

4. ICT Tools and Digital Pedagogy

Students used a range of ICT and AI tools to support simulation, modeling, design, and communication. MATLAB and Python enabled vibration analysis and automation, while AutoCAD/SolidWorks were applied for schematic layouts in MSD. ChatGPT supported idea generation, logic validation, and report drafting, complementing traditional instruction. For presentations, Canva and PowerPoint were employed in both courses. Instead of a static table, this integration is summarized as a workflow where simulation tools enhanced conceptual clarity, CAD platforms aided design accuracy, AI tools improved reasoning, and visual tools strengthened communication. Collectively, these resources ensured alignment with Challenge-Based Learning, enabling both technical fluency and professional communication skills..

5. Results and Student Feedback

5.1 Student Engagement and Skill Development

Survey responses showed strong engagement: 89% agreed that the MVA simulation improved conceptual clarity, 78% rated the MSD hackathon as highly engaging, and 85% felt the video assignment enhanced communication skills. However, more than 70% reported time management as a challenge. As illustrated in Figure 1, these perceptions were reflected in self-assessed skill growth. Communication improved from 2.0 to 4.6 on a five-point scale, ICT proficiency from 2.3 to 4.4, and teamwork from 2.8 to 4.7, while time management also showed modest gains.

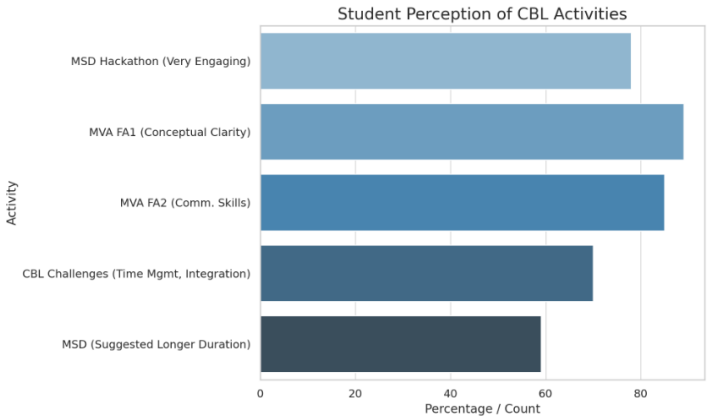


Figure 1. Student Perception of Various CBL Activities in Terms of Engagement and Skill Development.

5.2 Hackathon Performance and Conceptual Gains

Hackathon performance, evaluated through a rubric (innovation, technical accuracy, teamwork), yielded average scores of 8.1, 7.6, and 8.4 respectively, showing that students delivered both innovative and technically sound solutions under time constraints. Complementing this, the MSD pre- and post-test on stress analysis and gearbox kinematics showed scores rising from 42% to 76%, a 34-point improvement. Together, these results

(Figure 2) demonstrate measurable conceptual gains alongside professional skill development.

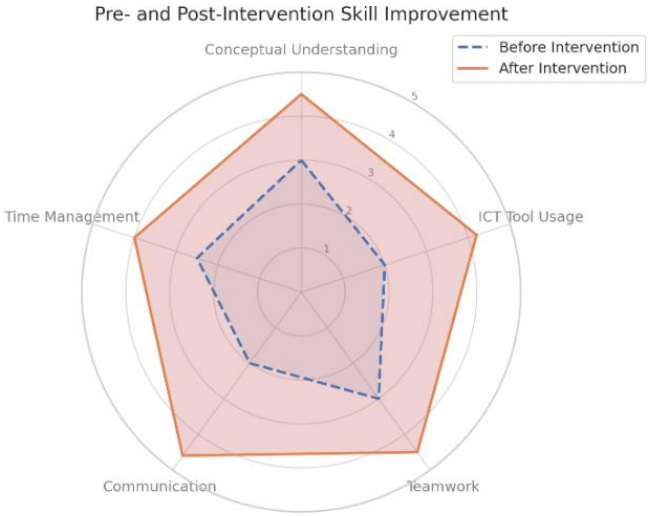


Figure 2. Radar chart of student skill levels before and after CBL interventions in MVA and MSD.

5.3 Course and Program Outcomes

As shown in Figure 3, more than 85% of students reported achieving MSD Course Outcomes at a “High” level, with CO1 (material handling systems) and CO2 (pressure vessel analysis) exceeding 90%. Broader Program Outcomes were also reinforced: over 90% of students noted gains in solution design (PO3), societal awareness (PO6), and ethics (PO8). Teamwork (PO9) was particularly strong, highlighted by 95% of respondents. Program Specific Outcome 01 (PSO01)—application of mechanical design principles and codes to real-world challenges—was validated by 92%, underscoring the industry-readiness of the courses.

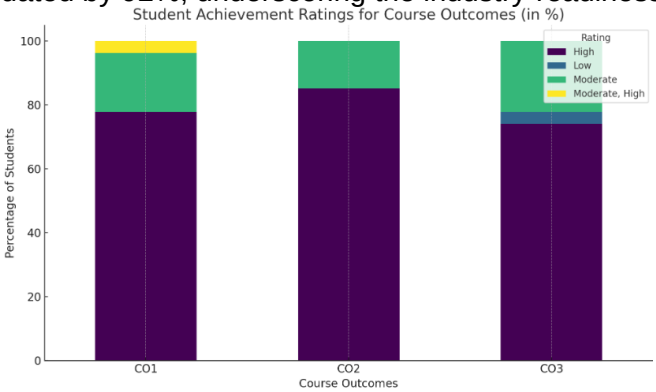


Figure 3. Student Ratings of Course Outcome Achievement for MSD (N=82).

5.4 Activity Glimpses and Ethical Considerations

Student participation was highly interactive and collaborative, as shown in Figures 4 and 5. In MVA, students created promotional videos to communicate vibration and noise-control concepts (Figure 4). In MSD, design hackathons featured innovative projects such as compressed-air-powered gearboxes, pneumatically powered conveyors, and CNG-powered conveyor vehicles (Figures 5). All activities were conducted with informed consent, and identifiable student faces were blurred to comply with ethical standards



Figure 4. FA2 Activity of the course MVA video making to advertise the Noise and Vibration Measuring Instruments.



Figure 6. Students involved in Design Hackathon for the course MSD

In compliance with ethical publication standards, all identifiable student faces in these images have been blurred, and informed consent for participation and documentation was obtained.

6. Challenges and Future Scope

Key challenges included varying student proficiency with tools, limited hackathon time, and subjectivity in rubric scoring, along with suggestions to add automation in conveyor design. Future iterations should include pre-hackathon training, longer design sessions, and integration of automation and sensors. Greater use of AI, standardized rubrics, and pre/post-tests across courses can enhance consistency, while multi-semester projects and virtual labs would further build real-world readiness.

7. Conclusion

This study shows that integrating Challenge-Based Learning (CBL) with ICT and AI tools significantly enhanced student learning in Mechanical Vibrations and System Design courses. Activities such as simulations, video storytelling, and hackathons improved conceptual understanding, communication, teamwork, and ICT fluency. Over 85% of students reported high achievement of Course Outcomes, and more than 90% affirmed gains in Program Outcomes related to solution design, ethics, and teamwork, with strong validation of PSO01. The use of tools like MATLAB, Python, AutoCAD, and ChatGPT supported both technical accuracy and professional communication, aligning with NEP 2020 and ABET standards. While time constraints and varied digital proficiency posed challenges, these can be addressed through extended hackathons, pre-activity training, and AI-assisted feedback. Overall, the approach offers a scalable model to prepare students for multidisciplinary, industry-relevant engineering challenges.

Acknowledgements

We would like to thank Pimpri Chinchwad College of Engineering (PCCoE), Pune — for providing the platform, resources, and encouragement to implement the courses.

References

Adams, R. S., Turns, J., & Atman, C. J. (2003). Educating effective engineering designers: The role of reflective practice. *Design Studies, 24*(3), 275–294. https://doi.org/10.1016/S0142-694X(02)00056-X

Biggs, J., & Tang, C. (2011). *Teaching for quality learning at university* (4th ed.). McGraw-Hill Education.

Boud, D., & Falchikov, D. (2007). *Rethinking assessment in higher education: Learning for the longer term*. Routledge.

Brereton, M., Donovan, J., Viller, S., & Biddle, R. (2015). Design challenge-based learning in engineering education. *Design Studies, 40,* 49–72. https://doi.org/10.1016/j.destud.2015.06.001

Calvani, A., Fini, A., & Ranieri, M. (2021). Digital competence in higher education: Students' perception and personal factors. *Sustainability, 13*(21), Article 12184. https://doi.org/10.3390/su132112184

Dillenbourg, P., Järvelä, S., & Fischer, F. (2009). The evolution of research on collaborative learning. In P. Dillenbourg, M. Specht, & J. D. Hoppe (Eds.), *Technology-enhanced learning* (pp. 3–19). Springer. https://doi.org/10.1007/978-1-4020-9827-7_1

Felder, R. M., & Brent, R. (2016). *Teaching and learning STEM: A practical guide*. Jossey-Bass.

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning, 4*(5/6), 304–317. https://doi.org/10.1504/IJTEL.2012.051816

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences, 111*(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Kolmos, A., Hadgraft, R. G., & Holgaard, J. E. (2016). Response strategies for curriculum change in engineering. *International Journal of Technology and Design Education, 26*(3), 391–411. https://doi.org/10.1007/s10798-015-9319-y

Kritzinger, W., Karner, M., Traar, G., Henjes, J., & Sihn, W. (2018). Digital twin in manufacturing: A categorical literature review and classification. *IFAC-PapersOnLine, 51*(11), 1016–1022. https://doi.org/10.1016/j.ifacol.2018.08.474

Kumar, J. A. (2021). Educational chatbots for project-based learning: Investigating learning outcomes for a team-based design course. *International Journal of Educational Technology in Higher Education, 18*(1), Article 65. https://doi.org/10.1186/s41239-021-00302-w

Lee, H.-Y., Wu, T.-T., Lin, C.-J., Wang, W.-S., & Huang, Y.-M. (2024). Integrating computational thinking into scaffolding learning: An innovative approach to enhance STEM hands-on learning. *Journal of Educational Computing Research, 62*(2), 233–256. https://doi.org/10.1177/07356331231211916

Prince, M. J. (2004). Does active learning work? A review of the research. *Journal of Engineering Education, 93*(3), 223–231. https://doi.org/10.1002/j.2168-9830.2004.tb00809.x

Savery, J. R. (2015). Overview of problem-based learning: Definitions and distinctions. *Interdisciplinary Journal of Problem-Based Learning, 1*(1), 9–20. https://doi.org/10.7771/1541-5015.1002

Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. Basic Books.

Shanbhag, S., & Al-Ammar, E. A. (2023). Enhancing design pedagogy using virtual studios. *Computer Applications in Engineering Education, 31*(2), 456–471. https://doi.org/10.1002/cae.22595

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*(4), 197–221. https://doi.org/10.1080/00461520.2011.611369

Bravo, F. A., & Cruz-Bohorquez, J. M. (2024). Engineering education in the age of AI: Analysis of the impact of chatbots on learning in engineering. *Education Sciences, 14*(5), Article 484. https://doi.org/10.3390/educsci14050484

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence in education. *International Journal of Educational Technology in Higher Education, 16,* Article 39. https://doi.org/10.1186/s41239-019-0171-0.

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

User Evaluation of a Virtual Patient for Philippine Medical Education

Maria Mercedes T. RODRIGO [a*], Samantha CASTAÑEDA[b], James Alvir Maclin V. ALAAN [a], Paolo Santino P. CAOILE [a], Alec Isaiah DAYUPAY [a], Roswold Jemuel SANCHEZ [a], Eric Cesar VIDAL, Jr. [a]

[a]*Ateneo de Manila University, Philippines*
[b]*Ateneo School of Medicine and Public Health, Philippines*
*mrodrigo@ateneo.edu

Abstract: Caladrius is a virtual patient system designed for use in Philippine medical schools. It responds to the need for medical interview training, for greater variety in teaching-learning strategies, and for culturally appropriate technologies. It has two versions: a text-only version whose interface is similar to a chat interface and an audio version that accepts speech input and responds with speech output. In a prior test of Caladrius, users requested improved audio response time and the inclusion of different patient personalities. A subsequent version of Caladrius was created to comply with these requests. As much of the lag was attributable to the need for Internet connectivity, we attempted to use offline speech-to-text and text-to-speech models for local processing. We also provided three possible personalities: neutral, talkative, and aggressive. We conducted a user test to determine whether the offline version of Caladrius performed better and if the personalities were distinct. The test with the offline version was unsuccessful, as the model was unable to recognize the test participant's speech accurately. The personalities were distinct, although the aggressive personality was more subtle and difficult to trigger. The paper ends with plans for continuing development.

Keywords: Generative AI, Patient History, Philippines, User Test, Virtual Patient

1. Introduction

Medical interviews are conversations between patients and health providers intended to collect patient information related to health (Goldberg, 2024). These include current symptoms, past medical conditions, work and hobbies, consumption of nicotine and alcohol, and so on (Berman et al., 2016). A medical interview is an essential process that contributes to overall health outcomes and doctor-patient rapport (Keifenheim et al., 2015).

Because of the importance of this process, medical interview training is part of the medical school curriculum (Novak, et al., 1993). In recent years, medical schools have integrated the use of virtual patients (VPs) among their teaching-learning strategies. Medical students interview VPs as they would human patients to hone their interview skills. These VPs provide answers to interview questions based on programmed patient information. Existing VPs discussed in the literature are typically developed in Western countries such as the US (see Scherr, et al., 2023) or the UK (see Enzer, et al., 2003).

Caladrius is a VP designed and developed for use specifically in Philippine medical schools. It responds to the need for medical interview training, for greater variety in teaching-learning strategies, and for culturally appropriate technologies.

As discussed in Rodrigo, Castañeda, Hernandez, and colleagues (2025), Caladrius is composed of a patient database, the VP, a mentor AI, and text and speech/audio (henceforth "audio") interfaces (Figure 1). Each row of the patient database is the full history of a single patient, provided by the second author of this paper who is a medical doctor. At the start of the interaction, the medical student selects from among the patients and begins the interview. Caladrius is a few-shot, lightweight, field-based retrieval system where structured patient data

is dynamically injected into LLM prompts; therefore, the responses of the VP are constrained to the record of the selected patient. The text and audio interfaces are built on top of ChatGPT, Google Cloud, and ElevenLabs technologies.

In addition to the patient persona, Caladrius also has a mentor AI that monitors the quality and flow of the interview. After each exchange, the mentor persona provides the student with a grade based on a rubric provided by the medical school faculty. The rating indicates which details of the medical history the student was able to collect and whether the student demonstrated professionalism and empathy (Rodrigo, Castañeda, Hernandez, et al., 2025).

An initial version of Caladrius was tested with students and faculty of a Philippine medical school in order to determine whether the content, flow, and performance of the VP were sufficient to make the system useful and usable. A more complete report of this test is detailed in (Rodrigo, Castañeda, Alaan, et al., 2025). In summary, the users agreed that the text version of the system was sufficiently useful and usable for student practice. The mentor AI was also useful and usable because it provided students with guidance about what else they needed to ask, though the grading tended to be somewhat inconsistent, e.g., repeated questions along the same theme caused grades to fluctuate from high to low. The feedback for the audio version was less positive. The users agreed that it was too slow to be practical, with response lag times sometimes exceeding 30 seconds. Furthermore, Caladrius's ability to understand spoken Filipino or Taglish was inconsistent, and the quality of non-English responses tended to be stilted.

Given these comments, the development team improved upon Caladrius in two ways: We improved upon the speech-to-text (STT) and text-to-speech (TTS) capabilities and created more patient personalities. Details of these improvements will be discussed in sections 2 and 3 respectively. We also conducted a user test with non-medical students to determine whether personality differences and the improvements on the audio interface were noticeable.

2. Improved Audio Version

The audio version of Caladrius had a different interface from the text version. The patient was a humanoid animated figure, not realistic in appearance. A clipboard displayed the transcription of the user's speech but did not show the response of the VP, the history of the conversation, or the mentor feedback. The mentor feedback was given as a single report at the end of the interview.

Many of the performance problems of the audio version stemmed from bandwidth issues. The audio version required an Internet connection, and therefore the speed of both the STT and TTS options was heavily dependent on the Internet speed of the device running the program. A slow internet connection or low bandwidth results in a significant delay between the end of the user's speech and the start of the virtual patient's audio, making conversation feel unnatural and frustrating.

Another limitation is that both the STT and TSS were dependent on the continued and uninterrupted operation of the online STT and TTS services. If either of these services were to become unavailable, the STT and TTS services would no longer function.

To eliminate the STT and TTS subsystems' dependence on external services and an Internet connection while also addressing the issue with speed, we attempted to run each subsystem locally. When the local version was selected, the STT and/or TTS subsystems ran on the device with the program.

For STT, a locally running Whisper model is used to transcribe the user's spoken words into text. This was made possible using a slightly modified version of the GodotWhisper (V-Sekai, 2024) plugin for Godot. For TTS, the built-in TTS system of Godot was used to synthesize speech.

Although making the TTS and STT subsystems run locally removed the dependence on connectivity, this approach was still subject to limitations. The performance of the audio version became dependent on the hardware running the program, i.e. the better the hardware, the faster these subsystems ran. This was most evident with the STT, as performance was affected by the processing speed of the hardware running the program and the hardware used

to record the audio of the user. Compared to its online counterpart, the offline STT could have difficulty with background interference present in the user's recorded dialogue.

Another limitation was that it limited both TTS and STT to only the English language. Although Filipino would sometimes work with these local versions, performance tended to be inconsistent. For TTS specifically, Filipino words were sometimes mispronounced, which led to a lack of clarity of the VP's response. The current implementation of the local TTS also has only one voice, whereas the online options could change their voice based on the chosen virtual patient's sex.

3. Inclusion of Patient Personalities

As the audio version was being revised, we also addressed the request to add more patient personality types. The original version of Caladrius had one base model which was "neutral." That is, the patient responded concisely to the medical student's queries. Responses were typically short, objective, and straightforward. They were accurate, only to the extent of the details provided in the medical history database. The neutral model could converse in English, Filipino, and Taglish, a combination of English and Filipino.

In addition to the neutral model, there are two personalities that can be applied to a patient and their details: talkative and uncooperative or aggressive (henceforth "aggressive"). The talkative model will also answer accurately and provide more information than what was asked, sometimes to the point that the patient forgets the question. The answers are usually long and can often appear rambling, overly worried, or unfocused. All additional information will also be to the extent of the details provided in the database. The uncooperative/aggressive model is straightforward and objective, but when asked about sensitive questions like family or mental health, the answers will be defensive, and information might not be shared. The answers are typically short, and for sensitive questions, may seem rude or offended. The talkative and aggressive personalities could only converse in English.

Implementing these three personalities required fine-tuning or training of three separate VPs, one per personality. To fine-tune these models, we provided ChatGPT with training examples. Table 1 illustrates the training example for the user question, "Do you have a history of psychiatric consults?"

A training example consists of messages with two components: roles and content. There are three role types. Both the system and the assistant refer to different aspects of the VP. The messages to the system (Table 1, blue text) set up the VP, informing it of its background or constraints. The assistant messages (Table 1, red text) provide the VP with a template of a response. Note that the VP can paraphrase its responses. The user messages (Table 1, green text) represent the inquiry that the user might ask. In all three cases, the question is "Do you have a history of psychiatric consults?". Table 2 contains examples of questions and responses from the three personality types.

4. User Test

Once the updates were completed and with clearance from the University Research Integrity Office, we conducted a user test to determine if the personality types were evident and if the offline version of the audio interface improved performance. We invited undergraduate students to test the text and audio versions of Caladrius. Five volunteers participated in the test. There were 2 females and 3 males, aged 19 to 22. All were computer science students. Since the focus of the test was the user interface and not the content, it was not necessary to recruit medical students for this test.

Each test was performed on an individual basis on the same computer. The computer had an AMD Ryzen 9 processor, a Radeon RX 6800S graphics card, and a 16 GB DDR5 RAM.

The first author of this paper served as the test monitor. At the start of the test, the test monitor explained the informed consent form and secured each participant's signature. The test monitor asked the participants to focus on the user interface and to think aloud as they interacted with Caladrius.

Table 1. *Training Examples for "Do you have a history of psychiatric consults?"*

Personality	Training Example
Neutral	{"messages":[{"role":"system","content":"You are a patient with history of psychiatric consult. You are a patient that only speaks the English language, You should only use English when communicating. When asked in another language, you must act confused."},{"role":"user","content":"Do you have history of psychiatric consult?"},{"role":"assistant","content":"Yes, I have history of psychiatric consult."}]}
Talkative	{"messages":[{"role":"system","content":"You are a patient with history of psychiatric consult. You are a talkative patient that must answer questions with more detail than asked but must not say information that you were not given. You must not make up information about yourself. You are a patient that only speaks the English language, You should only use English when communicating. When asked in another language, you must act confused."},{"role":"user","content":"Do you have history of psychiatric consult?"},{"role":"assistant","content":"Yes, I do. I saw a psychiatrist before because I was having anxiety attacks. It felt like my heart was racing all the time, and I would randomly start sweating, even if I wasn't doing anything. Sometimes I had trouble sleeping, and I kept overthinking things. The doctor helped me a lot, but I still get anxious sometimes."}]}
Aggressive	{"messages":[{"role":"system","content":"You are a patient with history of psychiatric consult. This is a sensitive question and you must answer aggressively. You are an uncooperative and aggressive patient that must answer questions very shortly and to the point but must be defensive when asked about sensitive topics. You are a patient that only speaks the English language, You should only use English when communicating. When asked in another language, you must act confused."},{"role":"user","content":"Do you have history of psychiatric consult?"},{"role":"assistant","content":"Why are you even asking? You think I'm crazy?"}]}

Table 2. *Sample Responses of Different Personalities.*

Question	Neutral Personality	Talkative Personality	Aggressive Personality
Do you have a history of psychiatric consultation?	Yes, I have a history of psychiatric consultation	Yes, I do. I saw a psychiatrist before because I was having anxiety attacks. It felt like my heart was racing all the time, and I would randomly start sweating, even if I wasn't doing anything. Sometimes I had trouble sleeping, and I kept overthinking things. The doctor helped me a lot, but I still get anxious sometimes.	Why are you even asking? You think I'm crazy?

The test monitor wrote down the comments and suggestions of the participants together with any observations about participants' difficulties. The interactions with Caladrius took between 30 to 40 minutes, after which the test monitor asked the participants to complete an online evaluation form.

The evaluation form asked the participants to rate different tasks within Caladrius (entering the user ID, selecting a patient, and conducting a conversation) based on ease of navigation, intuitiveness, and efficiency. The participants also answered open-ended

questions about the software's aesthetics, ease of use, and the distinctiveness of the personalities.

5. Results

The test participants also provided extensive comments on the user interfaces, recommending improvements to make the look and feel sleeker and more professional. For the text version, these suggestions included making the text box cursor more visible, adding cell padding, and adding an "exit" button. For the audio version, suggestions included making the field borders thinner, adding a "save" button for the configuration, and adding an "exit" button.

They also recommended consistency between the two versions. They noted, for example, that the text version provided mentor feedback after every dialogue exchange but did not provide the mentor's overall ratings at the end of the conversation. The audio version was the opposite—it provided the mentor's overall ratings at the end of the conversation but did not provide feedback after every dialogue exchange. Test participants noted that the text version allowed them to start a conversation with another VP without exiting the program, whereas the audio version required them to exit and restart the program in order to converse with another VP.

Despite these comments, the test participants did not have difficulty using Caladrius. We converted the ratings to numerical values: low was 1, medium was 2, and high was 3. On average, the participants' ratings ranged from medium to high, meaning that the participants found the interface easy to navigate, intuitive, and efficient.

The qualitative feedback, both written and oral, provided helpful insight about the audio interface and the personalities. The participants noted that, even though the audio interface provided a more realistic doctor-patient experience, they could not speak naturally. They had to hyperarticulate when using the audio interface. They had to enunciate their words and speak more slowly than normal in order to be understood. When they spoke normally, the system sometimes failed to understand their sentences.

The STT system had particular difficulty with Filipino and Taglish. It was not able to detect Filipino accurately and would sometimes transcribe Filipino sentences into other languages like Spanish.

When Caladrius would attempt to respond in Filipino, whether in the text or audio version, the Filipino grammar was incorrect. For example, when the test participant asked if the VP was taking any medicine, it replied, "*Aking gamot po ay hindi ako sigurado*" which literally translated to "My medicine I am not sure."

The STT system had a harder time discerning deeper voices. One of our male testers who had a deep voice struggled to be understood.

None of our participants successfully tested the offline versions of the STT or TTS. It is possible that the hardware configuration was insufficient to support the offline model.

The participants were able to notice differences in the personalities, however, some differences were more subtle than others. The aggressive personality, in particular, was not readily apparent. It was necessary to ask personal questions for the aggressiveness to surface. The personalities were less obvious in the audio version than in the text version.

The test participants noted that, while the VPs were different people, their characters seemed very similar. They suggested making the VP personalities more distinct by varying each VP's mannerisms, tone, energy levels, and speech patterns.

A further observation was that the VP tended to change persona from patient to doctor when the test participant asked an irrelevant question or else said something neutral like "hello" and nothing else. Another research team also building a VP on top of ChatGPT had the same experience.

6. Discussion and Next Steps

At present, the text version of Caladrius is able to converse with an acceptable level of performance in English, Filipino, and Taglish. The audio version works acceptably well if the

language is constrained to English and the Internet connection is reasonably fast. The performance of the offline STT and TTS models is unacceptable at this point.

There are several ways in which Caladrius may be improved as we move forward. The user interface issues are currently being addressed. More substantially, we are continuing to look for ways to improve upon the STT and TTS performance. We are also going to try to make the personalities more distinct and, in addition, provide the VPs with their own characters. The suggestions of the test participants to vary tone, mannerisms, and so on are possible ways of proceeding. We note at this point that the prior version of Caladrius also had a virtual reality version. Following the first user test, we decided to suspend development of this version, as the virtual reality experience, while more immersive, did not provide the medical students with any additional, useful information about the patient's health. In the future, we may continue development of this version, but only if the immersion adds pedagogical value. The most essential next step is to determine the extent to which Caladrius can be used in actual classes and what impact it has on learning. To this end, we have scheduled to conduct a field trial with one medical school class. The results from that field trial will be the subject of a future study.

Acknowledgements

This study is funded by the Office of the Vice President for Higher Education, Ateneo de Manila University. We thank the Ateneo School of Medicine and Public Health and the Ateneo Laboratory for the Learning Sciences. We also thank Dr. Reynan B. Hernandez, Riva Kaitleen Paras, and the students who participated in this study.

References

Berman, N. B., Durning, S. J., Fischer, M. R., Huwendiek, S., & Triola, M. M. (2016). The role for virtual patients in the future of medical education. *Academic medicine, 91*(9), 1217-1222

Enzer, I., Robinson, J., Pearson, M., Barton, S., & Walley, T. (2003). A reliability study of an instrument for measuring general practitioner consultation skills: the LIV-MAAS scale. *International Journal for Quality in Health Care, 15*(5), 407-412.

Goldberg, C. (2024). History taking. In Goldbert C., (eds)*, Practical Guide to History Taking, Physical Exam, and Functioning in the Hospital and Clinic*. McGraw Hill Professional. Accessed December 30, 2024. https://accessmedicine-mhmedical-com.asmphlibrary.remotexs.co/content.aspx?bookid=3505&sectionid=288414503

Keifenheim, K. E., Teufel, M., Ip, J., Speiser, N., Leehr, E. J., Zipfel, S., & Herrmann-Werner, A. (2015). Teaching history taking to medical students: a systematic review. *BMC medical education, 15*, 1-12.

Novack, D. H., Volk, G., Drossman, D. A., & Lipkin, M. (1993). Medical interviewing and interpersonal skills teaching in US medical schools: progress, problems, and promise. Jama, 269(16), 2101-2105.

Rodrigo, M.M.T., Castaneda, S., Hernandez, R. B., Alaan, J. A. M. V., Caoile, P. S. P., Chidrome, K. D., Dayupay, A. I., Lim, J., Redoble, A. P., & Vidal, E. C. Jr. (2025). Project Caladrius: The design of a virtual patient for Philippine medical education. Project Caladrius: The Design of a Virtual Patient for Philippine Medical Education. In International Conference on Human-Computer Interaction (pp. 232-244). Cham: Springer Nature Switzerland.

Rodrigo, M.M.T., Castaneda, S., Alaan, J. A. M. V., Caoile, P. S. P., Dayupay, A. I., Sanchez, R. J. C. & Vidal, E. C. Jr. (2025). Initial evaluation of a GenAI-based Virtual Patient for Philippine Medical Education. In: Cristea, A.I., Walker, E., Lu, Y., Santos, O.C., Isotani, S. (eds) Artificial Intelligence in Education. AIED 2025. Lecture Notes in Computer Science(), vol 15882. Springer, Cham. https://doi.org/10.1007/978-3-031-98465-5_45

RxList. (n.d.). Medical history definition. RxList. Retrieved February 6, 2025, from https://www.rxlist.com/medical_history/definition.htm

Scherr, R., Halaseh, F. F., Spina, A., Andalib, S., & Rivera, R. (2023). ChatGPT interactive medical simulations for early clinical education: case study. JMIR Medical Education, 9, e49877.

V-Sekai. (2024, April 18). *godot-whisp*er (Version 1.0.5) [Computer software]. GitHub. https://github.com/V-Sekai/godot-whisper

Critical Thinking as a Dynamic System: Tracking Nonlinear Development through Reading Study Worksheets in Elementary Science

Manyu HUANG[a], Xueqing ZHOU[b] & Guang CHEN[a*]
[a]*Faculty of Education, Beijing Normal University, China*
[b]*Organization of Synthesis, Longquanyi Shiyan Elementary School, China*
*guang@bnu.edu.cn

Abstract: This study applies Dynamic Systems Theory and the moving min-max graph to analyze the nonlinear development of critical thinking skills in elementary science education. Based on an eight-session reading worksheet intervention with 49 fifth-grade students, the study reveals three-stage trajectories, four types of learner behavior, and asynchronous development across critical thinking dimensions. This process-oriented approach offers a significant departure from traditional pre-post measurement paradigms.

Keywords: critical thinking, dynamic systems theory, science education, reading study worksheets, learning analytics

1. Introduction

Traditional developmental assessment measures, such as gain scores, often focus on overall changes in outcomes while neglecting the complexity and dynamics inherent in the developmental process. These metrics fail to capture the developmental range and variability of individuals at each time point—a critical oversight, as such variability may contain key transition points. Within science education, where critical thinking, recognized as a core 21st-century skill, develops through iterative engagement, this limitation is problematic. Reducing critical thinking development to a linear pre-post comparison fundamentally misrepresents its nonlinear trajectory.

Dynamic Systems Theory (DST) addresses this gap by positioning variability as a fundamental indicator of development, offering a transformative perspective by conceptualizing cognitive development as a complex, self-organizing process characterized by instability and phase transitions (Thelen & Smith, 1994). Through visualization techniques like the moving min-max graph (van Geert & van Dijk, 2002), DST enables researchers to detail the fluctuation range and variability of individuals at different measurement points, revealing the development trend and identifying critical transition points.

Critical reading, which demands an understanding of both content and context (Comber & Nixon, 2011), provides a vital foundation for fostering critical thinking in elementary science. This study uses structured science reading worksheets with stepped questions to explore the dynamics of CT development from a dynamic perspective. It shifts the focus from "how much change occurs" to "how change unfolds," empowering educators to decode the complexity of students' thinking processes.

2. Literature Review

2.1 Dynamic Systems Theory

Dynamic Systems Theory (DST) offers a theoretical framework for understanding learning as a complex, nonlinear, and self-organizing system. Many elements of learners' differences constitute a complex system. Development is not proportional to input; small changes can trigger significant shifts, and progress can regress or plateau (Granic et al., 2003). Self-organization refers to the process by which a system spontaneously forms an ordered structure through the interactions of its internal elements without external directives.

The states of a complex system are not unique. An attractor refers to a stable state toward which a system naturally evolves and tends to remain. A repellor refers to an unstable state that the system tends to move away from (Hollenstein, 2013). Phase transitions mark significant shifts between these states, often triggered by learning experiences and observable as periods of heightened variability in performance.

2.2 Critical Thinking in Science Education

Definitions converge on Critical Thinking (CT) as higher-order, purposeful thinking (Ennis, 1987; Facione, 1990), encompassing both cognitive skills and affective dispositions (Facione, 1990). This study focuses on the cognitive skill dimensions outlined in the Delphi Report, specifically interpretation, analysis, evaluation, inference, and explanation.

Cultivating CT in science education is crucial, given its unique demands for evidence-based reasoning and skeptical inquiry. Strategies often involve engaging with scientific texts. In College, interventions focus on critical analysis of primary literature (Fischer, 2021) or the C-QRAC collaboration scripts (Lee, 2015). In secondary, strategies include analyzing science-based news reports (McClune & Jarman, 2010), newspaper articles (Oliveras et al., 2013), and higher-order thinking tasks (Miri et al., 2007). Research at the elementary level is sparse, though science-based reading expands instructional time for deeper inquiry (Romance & Vitale, 1992), and concept mapping shows promise for skill development (Akinoglu, 2013). In conclusion, a gap exists in understanding the dynamic and nonlinear development of CT skills in young learners and how specific tools, such as worksheets, facilitate this process over time.

2.3 Reading Study Worksheets

Reading study worksheets are instructional tools that guide students through text by providing structured tasks. While prevalent in language arts (Anwar et al., 2020; Haswani et al., 2023), the use of these approaches in science, especially elementary science, is less explored. Emerging evidence suggests their efficacy in improving scientific explanation and confidence in junior high (Ikeda et al., 2023) and critical analysis skills in college (Fischer, 2021). However, studies typically rely on pre-post designs, failing to capture the dynamics of how worksheet-guided activities influence CT skill emergence and stabilization on a moment-to-moment and session-to-session basis.

3. Research Questions

Guided by DST, this study focuses on the development of CT skills in science for elementary students by using reading study worksheets. The research questions are as follows:

1. What is the nonlinear developmental trajectory of critical thinking skills in elementary school students guided by science reading study worksheets?
2. Is there consistency in the effects of students' critical thinking skills being influenced by reading study worksheets?
3. Is there consistency in the development of the dimensions of critical thinking skills among elementary school students?

4. Methods

4.1 Participants

The participants consisted of 52 fifth-grade students from an elementary school in Sichuan Province, China, with ages ranging from 10 to 11 years and a gender ratio approximating 1:1. Informed consent was obtained from all participating students and their teachers. Due to 1-2 absences by 3 students, 49 valid data points concerning the development of CT skills were included, which aligns with the requirement of DST for observing the entire process.

4.2 Research Procedure

The procedure was structured into two sequential phases to investigate the development of CT skills through science reading interventions. Each session lasted 30 minutes and took place once a week. Phase 1 (Sessions 1-2) focused on establishing baseline performance. Students engaged in unstructured reading of science articles without worksheet guidance, allowing researchers to capture initial skill levels in interpretation, analysis, evaluation, inference, and explanation. During Phase 2 (Sessions 3-8), participants transitioned to structured reading activities guided by stepped-question reading study worksheets.

4.3 Research Tools

The reading study worksheets were developed using literature-based adaptation, pilot testing, and expert validation. Every question was mapped to one of the five cognitive dimensions defined by the Delphi Report and then adapted from validated instruments such as A Test of Science Critical Thinking for fifth-grade students (Mapeala & Siew, 2015). Then, a pilot study was conducted with 15 fifth-graders from a parallel class; students thought aloud while completing the worksheets. Feedback indicated some items needed to be simplified in scientific terms or replaced due to overly advanced concepts.

The reasonableness of the worksheets' content and evaluation criteria was validated through an expert review method. Three reviewers rated each item's relevance to its appropriateness of the 5-point rubric (0 = no understanding, 5 = sophisticated, context-rich response). Mean ratings were 4.47 for interpretation, 4.22 for analysis, 4.00 for evaluation, 4.14 for inference, and 4.22 for explanation. All ratings were 4.0 or higher on a 5-point scale, confirming that the worksheets and criteria are theoretically grounded and empirically refined.

4.4 Data Analysis Methods

DST provides unique analytical methods, with the moving min-max graph as a time-series visualization tool (van Geert & van Dijk, 2002). The moving min-max graph is built in three steps. First, set a moving window (e.g., 8 data points) based on the data's time structure. This window slides through the dataset, each time excluding the oldest point from the previous window and adding the following new point. Second, for each window, calculate the minimum and maximum values of the target variable. Third, plot these minimum and maximum values over time—their vertical span forms a bandwidth, showing variability within each window. This technique shows the data using the bandwidth of observed scores.

In this study, three scientific readings were used as a moving window, and the data from 8 measurements were divided into 6 windows, e.g., (t1, t2, t3), (t2, t3, t4), ..., (t6, t7, t8). To visualize the entire 8-session trajectory, the min and max values for the first window (sessions 1-3) were assigned to sessions 1 and 2, and the min and max values for the last window (sessions 6-8) were assigned to sessions 7 and 8. This allows the moving extrema lines to span the full x-axis alongside the total score line.

The visual patterns of the min-max graph provide crucial insights into the learner's cognitive state dynamics. A narrowing bandwidth (decreasing vertical distance between min and max lines) suggests the system is stabilizing and may be approaching or residing within

the basin of an attractor state. A widening bandwidth suggests instability and can be a sign that the system is in a transient state, potentially moving away from a repellor or undergoing a phase shift.

5. Results

5.1 Nonlinear Developmental Pathways

Analysis of total score-moving min-max graphs for 49 participants revealed two key patterns in critical thinking development. First, an overall upward trajectory was observed across sessions. Second, the process was characterized by significant nonlinear fluctuations, as 40 participants exhibited multiple performance peaks, specifically 37 with two peaks and 3 with three peaks, while only 9 showed minimal fluctuations with zero or one peak.

Cluster analysis using between-groups linkage on total scores from eight sessions demonstrated a three-stage developmental trajectory. Readings 1-2 formed a cluster representing initial sessions without worksheet intervention. Readings 3-5 formed a cluster characterized by upward fluctuations and the emergence of the first peak following the implementation of the worksheet. Readings 6-8 formed a cluster where skills stabilized despite continued fluctuations, resulting in subsequent peaks. This segmentation confirms a nonlinear developmental pattern initiated by pedagogical intervention.

5.2 Individual Differences and Classification

Participants were classified into four distinct developmental profiles based on a combination of cluster analysis and the number and timing of performance peaks.

Group 1 is characterized by three performance peaks. This group demonstrated lower performance in Reading Sessions 1-2, significant fluctuations in Sessions 3-6, and higher performance in Sessions 6-8. The moving extrema gap exhibited an increase-decrease trend, reflecting that fluctuations in students' CT levels first intensified and then diminished. According to DST, this fluctuation pattern suggests that the participants experienced a repellor state. These students exhibit highly active cognitive development. Their critical thinking is readily stimulated by interventions and responds rapidly, classifying them as efficient learners.

Group 2 exhibited two peaks, in Sessions 3 and 5. The moving extrema gap showed a decrease-increase trend, reflecting reduced then amplified fluctuations. This suggests that the participant experienced an attractor state. These students' thinking stabilizes quickly after an initial rapid improvement, and they adapt to the guidance provided by the reading worksheets.

Group 3 also exhibited two peaks, around Sessions 3 and 6/7. The moving extrema gap displayed a decrease-increase-decrease trend, pointing to an early attractor and a later repellor. These students require a longer period of accumulation for cognitive development, with intervention effects showing delayed results. Comprising the majority of participants, they represent the common trajectory that requires sustained practice to overcome thinking bottlenecks.

Group 4 showed minimal fluctuation, manifesting as either a single peak or a steady rise. Despite lacking clear peaks, the moving extrema gap still demonstrated a trend of decrease-increase-decrease, implying weak but detectable attractor-repellor dynamics. Correspondingly, the growth of these students' CT was slow and incremental, consistent with a more passive learning mode.

5.3 Asynchronous Development Across Dimensions

This study creates dimensional moving min-max graphs to examine how each dimension develops. Participant 51 from Group 3 (see Figure 1) was selected for detailed presentation as it exemplifies the complex interplay of attractor and repellor states within a single system, a characteristic feature observed in this largest group. Interpretation revealed decreasing-

increasing trends with a marginally narrowing extremum bandwidth between moving extremes, suggesting insufficient evidence for attractor or repellor states. Analysis revealed up-down-up-down fluctuations amid progressive bandwidth contraction, indicating probable attractor stabilization with extended tracking. In evaluation, performance oscillated up-down-up-down as bandwidth first decreased, then increased, then decreased again, indicating the presence of one attractor and one repellor. Inference followed an up-down-up pattern, with bandwidth changing by first increasing, then decreasing, then increasing, similarly suggesting state competition. Explanation featured up–down–up–down fluctuations, yet bandwidth steadily shrank toward zero, indicating the presence of a repellor and an attractor. In the 7th science reading, the two moving lines coincide, indicating precisely the attractor state.

In addition, cross-dimensional comparison revealed marked asynchrony. During Sessions 5-6, Participant 51 improved in evaluation while regressing in explanation, suggesting resource competition between dimensions. These asynchronous patterns underscore DST's principle of multidimensional self-organization, where subsystems interact asynchronously to shape holistic skill development.

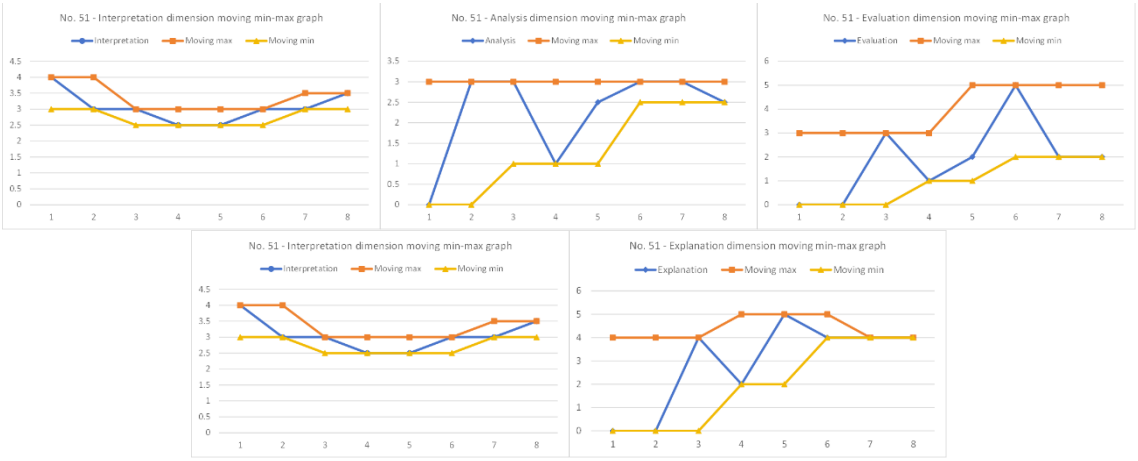


Figure 1. Moving Min-Max Graphs for Five Dimensions of Critical Thinking Skills (No.51).

Similar analyses for participants from the other three groups consistently revealed that students' overall attractor or repellor states are not equal to the simple sum of their subdimension states. It demonstrates that developing critical thinking skills is not a consistent process across all dimensions, but rather a complex and dynamic one.

6. Discussion and Conclusion

This study applied Dynamic Systems Theory to explore the nonlinear development of CT skills in elementary science education through reading study worksheets. The findings revealed that CT development follows a three-stage trajectory (baseline, transitional fluctuation, consolidation) characterized by variability and asynchronous progression across dimensions. This confirms DST's premise that cognitive growth is a dynamic, self-organizing process marked by instability (Thelen & Smith, 1994).

Reading study worksheets effectively scaffolded CT skills, as evidenced by the overall upward trend. However, their impact varied across learners, yielding four developmental profiles: efficient learners, early stabilizers, late stabilizers, and passive learners. This heterogeneity may stem from individual differences in receptiveness, metacognitive strategies, and learning pace. Educators can leverage these profiles by using the moving min max graph to identify attractor and repellor states and deliver personalized interventions, including metacognitive prompts for unstable dimensions and challenging tasks for efficient learners.

Critically, CT dimensions developed asynchronously, reflecting system-level instability. Interpretation and Analysis stabilized early (Sessions 5-6), suggesting that foundational skills are easier to consolidate. Evaluation and Inference fluctuated until Session 7, indicating higher

cognitive load in integrating concepts. Notably, many students exhibited cross-dimensional competition, mirroring DST's concept of resource redistribution within a complex system.

However, the 8-session intervention captured initial development but limited observation of long-term consolidation or skill transfer. Future studies should extend duration and investigate interactions between CT skills and dispositions.

References

Akinoglu, O. (2013). Effects of Concept Maps on Students Critical Thinking Skills in Science Education. JOURNAL OF ENVIRONMENTAL PROTECTION AND ECOLOGY, 14(3A), 1424–1431.

Anwar, L., Furwana, D., & Iksan, M. (2020). Developing Reading Worksheet for Tenth Grade Students of SMA Negeri 4 Palopo. FOSTER: Journal of English Language Teaching, 1(1), 1–8. https://doi.org/10.24256/foster-jelt.v1i1.2

Comber, B., & Nixon, H. (2011). Critical reading comprehension in an era of accountability. The Australian Educational Researcher, 38(2), 167–179. https://doi.org/10.1007/s13384-011-0022-z

Ennis, R. H. (1987). A taxonomy of critical thinking dispositions and abilities. In Teaching thinking skills: Theory and practice (pp. 9–26). W H Freeman/Times Books/ Henry Holt & Co.

Facione, P. (1990). Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction. Research Findings and Recommendations. https://www.semanticscholar.org/paper/Critical-Thinking%3A-A-Statement-of-Expert-Consensus-Facione/dfd988ab957da7816826034712451adc73dc5065

Fischer, K. M. (2021). Using Critical Analysis of Scientific Literature To Maintain an Interactive Learning Environment for In-Person and Online Course Modalities. JOURNAL OF MICROBIOLOGY & BIOLOGY EDUCATION, 22(1). https://doi.org/10.1128/jmbe.v22i1.2523

Granic, I., Hollenstein, T., Dishion, T. J., & Patterson, G. R. (2003). Longitudinal analysis of flexibility and reorganization in early adolescence: A dynamic systems study of family interactions. Developmental Psychology, 39(3), 606–617. https://doi.org/10.1037/0012-1649.39.3.606

Haswani, F., Erlita, Y., & Rika. (2023). Integrating extensive and intensive reading worksheets on digital platforms. JOALL (Journal of Applied Linguistics and Literature), 8(2), 415–432. https://doi.org/10.33369/joall.v8i2.28043

Hollenstein, T. (2013). State space grids: Depicting dynamics across development (pp. xiii, 127). Springer Science + Business Media. https://doi.org/10.1007/978-1-4614-5007-8

Ikeda, K., Yamasita, S., Sunaga, K., & Miyake, Y. (2023). Development and evaluation of a reading and worksheet on the subject of superconductivity and linear bullet trains for 2nd year junior high school students. JSSE Research Report, 38(1), 7–10. https://doi.org/10.14935/jsser.38.1_7

Lee, Y.-H. (2015). Facilitating critical thinking using the C-QRAC collaboration script: Enhancing science reading literacy in a computer-supported collaborative learning environment. Computers & Education, 88, 182–191. https://doi.org/10.1016/j.compedu.2015.05.004

Mapeala, R., & Siew, N. M. (2015). The development and validation of a test of science critical thinking for fifth graders. 4, 741.

McClune, B., & Jarman, R. (2010). Critical Reading of Science-Based News Reports: Establishing a knowledge, skills and attitudes framework. International Journal of Science Education, 32(6), 727–752. https://doi.org/10.1080/09500690902777402

Miri, B., David, B.-C., & Uri, Z. (2007). Purposely Teaching for the Promotion of Higher-order Thinking Skills: A Case of Critical Thinking. Research in Science Education, 37(4), 353–369. https://doi.org/10.1007/s11165-006-9029-2

Oliveras, B., Márquez, C., & Sanmartí, N. (2013). The Use of Newspaper Articles as a Tool to Develop Critical Thinking in Science Classes. International Journal of Science Education, 35(6), 885–905. https://doi.org/10.1080/09500693.2011.586736

Romance, N. R., & Vitale, M. R. (1992). A curriculum strategy that expands time for in-depth elementary science instruction by using science-based reading strategies: Effects of a year-long study in grade four. Journal of Research in Science Teaching, 29(6), 545–554. https://doi.org/10.1002/tea.3660290604

Thelen, E., & Smith, L. B. (1994). A dynamic systems approach to the development of cognition and action (pp. xxiii, 376). The MIT Press.

van Geert, P., & van Dijk, M. (2002). Focus on variability: New tools to study intra-individual variability in developmental data. Infant Behavior and Development, 25(4), 340–374. https://doi.org/10.1016/S0163-6383(02)00140-6

Assessment of Comment Quality in Active Video Watching using Deep Learning

Chantelle BALDWIN* **& Antonija MITROVIĆ**
University of Canterbury, New Zealand
chantelle.baldwin@pg.canterbury.ac.nz

Abstract: This study investigates the effectiveness of deep learning (DL) methods for classifying student comments based on quality within the AVW-Space video-watching platform. Given the limitations of existing machine learning (ML) techniques, this research explores whether DL models can improve classification performance. The study compared six DL models (BERT, RoBERTa, ALBERT, ELECTRA, GPT, and GPT-2) by training them on a dataset of 13,440 student comments. The results show that RoBERTa outperforms all models, demonstrating precision, recall, and F1-score improvements. Fine-tuning experiments led to an optimised RoBERTa model. We examine methods to address class imbalance, with weighted loss functions and random undersampling proving ineffective. This study contributes to the automation of comment assessment and supports personalised educational experiences.

Keywords: LLMs, Deep learning, Classifiers, Learning analytics, Video-based Learning, Automated Evaluation

1. Introduction and Related Work

Using video media for learning has become prevalent. Video-based learning (VBL) has unique features supporting various learning styles and has the potential to lead to better learning outcomes (Zhang et al., 2006). Recent VBL systems also facilitate self-regulated learning (Dimitrova & Mitrovic, 2022), but the lack of direct interaction with teachers can lead to passive learning (Yousef et al., 2014). Active Video Watching (AVW) mitigates inefficient learning and encourages engagement in video-based teaching.

AVW-Space is a controlled video-based learning platform that enables teachers to create customised spaces by selecting YouTube videos and specifying aspects to direct students' attention (Mitrovic et al., 2016). These aspects prompt students to reflect on their knowledge and experiences. Students can pause the video to leave comments at any point, and teachers may select comments to be anonymously shared for peer review, enabling classmates to rate and respond. Continuous enhancements have optimised AVW-Space for learning and engagement (Dimitrova & Mitrovic, 2022; Mitrovic, Dimitrova, et al., 2017; Mitrovic et al., 2016, 2019; Mohammadhassan et al., 2020). This study focuses on comments from a presentation skills module, which contained four tutorial videos on delivering presentations and four actual recordings of presentations.

Students who reviewed and wrote comments improved their knowledge, while those who passively watched did not (Mitrovic, Dimitrova, et al., 2017). Initial studies demonstrated the benefits of behaviour-based hints designed to encourage commenting and offer exemplar responses, known as personalised nudges, leading to increased student comments (Mitrovic et al., 2019). AVW-Space was enhanced to deliver personalised nudges; however, students did not yet receive real-time feedback on comment quality.

In response, Mohammadhassan et al. (2020) explored whether machine learning (ML) could assess comment quality, introducing a five-class scheme specific to tutorial videos. Class 1 (Affirmative, negative, off-topic) contains irrelevant or purely affirmative/negative remarks without explanation. While Class 2 (Repeating) consists of comments restating video content, both are considered low quality. The remaining classes represent higher quality. Class 3 (Critical and analytical) reflects critical thinking, while Class 4 (Self-reflective) connects

to the learner's prior experiences or behaviours. In Class 5 comments, students write about how to improve in future. Weighted classes and cost-sensitive error handling addressed the dataset's class imbalance. Misclassifying Class 3 as Class 2 is less severe than misclassifying Class 1 as Class 5, as the latter represents the most significant gap in engagement and thus incurs the highest penalty. A weighted random forest classifier achieved an F1-score of 0.68 and an average cost of 3.53. Performance was improved by merging classes, though this reduced the specificity of the feedback.

Reflective writing classification is challenging due to its unstructured nature and varied vocabulary. Literature shows deep learning (DL) methods often outperform traditional ML. Wang et al. (2024) achieved an F1-score of 0.699 using BERT, improving to 0.7474 with GPT-4-generated data and contrastive learning to stabilise training. Wulff et al. (2023) classified teachers' reflections with BERT, but faced tokenisation issues due to the niche domain. Li et al. (2023) found that BERT outperformed ML for reflective writing (F1-score: 0.7217).

This study investigates whether DL can outperform existing ML methods for AVW-Space comment classification. We evaluate transformer-based models; BERT (Devlin et al., 2019), ALBERT (Lan et al., 2020), RoBERTa (Liu et al., 2019), ELECTRA (Clark et al., 2020), GPT (Radford et al., 2018), and GPT-2 (Radford et al., 2019). Experiments assess hyperparameter tuning (dropout, hidden layers, batch size, epochs, learning rate), and testing strategies for handling class imbalance. Findings will inform NLP-driven educational tools and deepen understanding of DL's effectiveness for educational text classification.

2. Proposed Method

Mohammadhassan et al. (2020) used 2,245 manually classified comments, collected from 2016 to 2019, to develop an ML-based method. They split the data into 80% training and 20% testing. Since that study, more data is available for training and testing the model. We experimented with 13,440 comments written by first-year engineering students and collected from 2018 to 2023 (Dimitrova & Mitrovic, 2022; Mitrovic et al., 2019; Mohammadhassan et al., 2020, 2022). The studies were approved by the Human Research Ethics Committee (HREC) of the University of Canterbury.

We split the data randomly into a training, validation and testing set of 70%, 15% and 15%, respectively. The class distribution of the comments is uneven. There are 7,599 comments in Class 2 (Repeating), accounting for 57% of the dataset. Class 3 (Critical and analytical) has 2,813 comments, making up 21%. Class 4 (Self-reflective) has 2,253 comments, 17% of the dataset. Class 1 and Class 5 contain the smallest number of comments, both holding 3% of the total collected comments, 341 and 334 comments, respectively.

2.1 Model Comparison

The DL models selected for comparison were BERT, ALBERT, RoBERTa, ELECTRA, GPT, and GPT-2. This selection provides a broad overview of the performance of two leading architectural frameworks, BERT-based and GPT-based models. We selected these models over larger, more recent architectures due to practical and ethical considerations. Fine-tuning and deploying larger-scale models exceeded available computational resources. Also, AVW-Space requires near-instantaneous comment evaluation, favouring smaller models with faster inference. Additionally, performing evaluations locally safeguards student privacy by avoiding transmission of sensitive data to external entities. Each model is publicly available via Hugging Face (bert-base-uncased, albert-base-v2, roberta-base, google/electra-base-discriminator, openai-community/openai-gpt, and openai-community/gpt2) (Hugging Face, 2025).

We used the F1-score, precision, recall and average cost to compare with the results obtained from the ML benchmark testing. The goal of the evaluation was to help identify the model that best captures the complex nature of student comments, ultimately improving the accuracy of the comment classification system in AVW-Space. Model performance determined the selection of a single model for the remaining tests.

2.2 Fine-tuning and Weighted Training

Fine-tuning a model to find the optimal hyperparameters involved training models using different parameter combinations to maximise performance while ensuring generalisability to unseen data. We tested dropout rates of 0.1, 0.2, 0.3, 0.4, and 0.5; batch sizes of 16, 32, and 64; training for 2 to 6 epochs; learning rates of 1e-5, 2e-5, and 5e-5; and hidden layer counts ranging from 2 to 12. The evaluation covered 2,046 configurations, each trained and validated on the same dataset split for comparability. To address variability in DL training, the six top-performing configurations from the initial tests underwent four additional training runs. Subsequent experiments utilised the best-performing configuration.

To address class imbalance, we modified the default cross-entropy loss using focal loss for hard-to-classify examples. We applied inverse frequency weighting through both weighted cross-entropy and alpha-balanced focal loss. In addition, random undersampling balanced all classes to 334 samples to create an even distribution of comment classification.

3. Results

We tested six models to assess whether a DL approach would outperform the existing ML method to classify the quality of student comments. The selected models are BERT, ALBERT, RoBERTa, ELECTRA, GPT, and GPT-2. The BERT-based models were trained with a batch size of 64, while GPT models used 32 due to higher computational costs. Preliminary tests showed validation loss typically increased after four epochs, suggesting overfitting; thus, all the DL model training consisted of four epochs. Table 1 presents the aggregated precision, recall, F1-score and average cost obtained for each of the DL models.

Table 1. *Recall, precision, F1-score and average cost of different models*

Model	Recall	Precision	F1-Score	Average Cost
RoBERTa	**0.740**	**0.735**	**0.734**	**3.418**
ALBERT	0.738	0.734	0.732	3.513
GPT	0.731	0.728	0.728	3.658
BERT	0.724	0.722	0.723	3.738
GPT-2	0.729	0.721	0.712	3.823
ELECTRA	0.718	0.693	0.704	3.943
Retrained ML	0.719	0.706	0.692	4.030

The retrained ML model achieved a recall of 0.719, a precision of 0.706, and an F1-score of 0.692. All the DL models outperformed the retrained ML model across F1-score and average cost. RoBERTa achieved the highest performance across all metrics: a recall of 0.740, precision of 0.735, F1-score of 0.734, and the lowest average cost of 3.418. Overall, while the model performs well, further fine-tuning could improve its performance further.

ALBERT showed promising results with a recall of 0.738, precision of 0.734, and an F1-score of 0.732. ALBERT's compact architecture provides advantages in terms of computational efficiency, making it a good option for resource-constrained environments. However, RoBERTa remains the best-performing model due to its superior balance of accuracy and efficiency, which justifies its selection for further experimentation. Our findings were consistent with previous findings, which demonstrate that RoBERTa outperforms other DL alternatives (Sy et al., 2024, p. 20).

Additionally, GPT models obtained lower performance metrics than both RoBERTa and ALBERT. GPT models are generally less suited to tasks requiring fine-tuning for specific downstream applications, as their pre-training objective is broad in scope. This generality could have led to the suboptimal performance compared to models like RoBERTa, which are trained on a diverse vocabulary and further optimised for targeted tasks. However, it is important to note that the GPT models used a smaller batch size of 32 for training. The smaller batch size could have led to overfitting and reduced its ability to generalise

3.1 Fine-Tuning and Weighted Training

We tested 2,046 hyperparameter configurations for fine-tuning RoBERTa on student comment classification. Of these, 72 outperformed the default RoBERTa model (F1-score: 0.734), with the highest achieving 0.746, a marginal improvement of 0.012.

To ensure consistency, we trained the top six configurations three additional times. Results showed slight improvements over the default model but revealed that fine-tuning does not always guarantee better performance. The best performing configuration (dropout of 0.2, 10 hidden layers, batch size of 16, 6 epochs, and learning rate of 5e-5) achieved the highest F1-score of 0.747, precision of 0.749, and recall of 0.753. The configuration reduced the average cost to 3.29 from the RoBERTa default of 3.42, and therefore served as the configuration for further testing. This configuration also improved performance across all classes (Figure 1). Class 1 achieved an F1-score of 0.789, precision of 0.811 and recall of 0.768. Class 2 maintained strong performance with an F1-score of 0.829, precision of 0.794, and recall of 0.868, comparable to the retrained ML model.

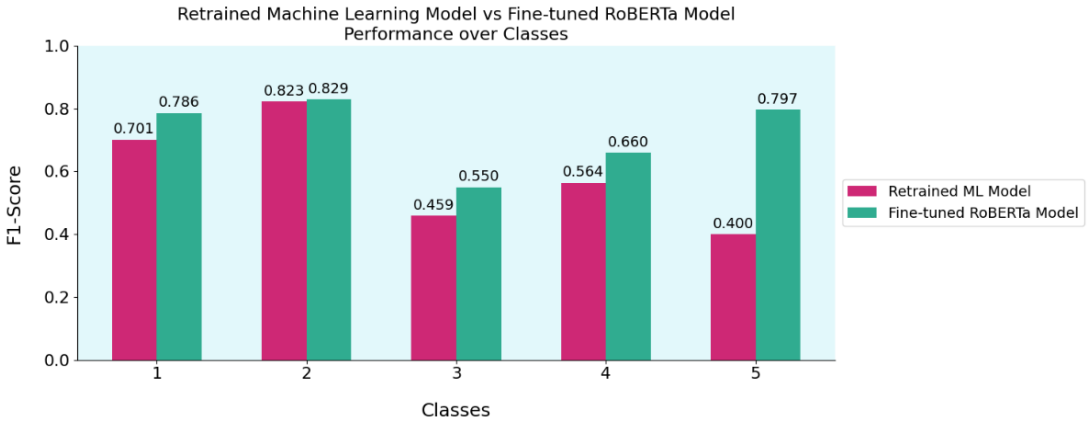


Figure 1. Retrained ML method versus the fine-tuned RoBERTa model performance

Class 3 continued to underperform with an F1-score of 0.550, precision of 0.577, and recall of 0.525, possibly due to the factual nature of comments, which made it difficult to differentiate from Class 2. For example, one comment classified correctly as Class 2, with the aspect "I did/saw this in the past," stated, "Only using meaningful numbers and statistics that enhance your point,". However, another comment with the same aspect, "Making the audience imagine works as a good opener," was misclassified as Class 2 instead of Class 3. The Class 3 classification was dependent on the content of the video. Incorporating video transcripts might improve classification performance but risks reducing generalisability.

Class 4 also improved with an F1-score of 0.660, a precision of 0.777, and a recall of 0.573. The lower recall compared to precision indicates that the model has a conservative approach to classifying into Class 4, often misclassifying as Class 2 comments. An example is a Class 4 comment with the aspect "I am rather good at this," stated, "Show you are actually interested in your topic," but was misclassified as Class 2. The example comment highlights the challenge of distinguishing self-reflective content when the reflective components are contained in the aspect rather than the comment.

Class 5 saw the most significant gain, achieving an F1-score of 0.797, a precision of 0.731, and a recall of 0.875, marking a 0.397 increase in F1-score over the retrained ML model. The higher cost of misclassifying Class 5 comments suggests that the average cost reduction was primarily due to the improved classification of these comments.

Although the aggregated F1-score presents marginal improvement from the RoBERTa model, there is a decrease in average cost and improvement in performance across all classes. The difference in each class's performance highlights the limitations of evaluating multiclassification models such as these using aggregated metrics.

The top fine-tuned model performed well for minority Class 1 and 5 but struggled with Class 3 and 4, which were harder to distinguish from the majority Class 2. To address the

class imbalance, weighted cross-entropy, focal loss, alpha-balanced focal loss, and random undersampling were tested. The default cross-entropy achieved the strongest performance with an F1-score of 0.747, while focal loss produced a comparable result with an F1-score of 0.735. Random undersampling performed considerably worse, with an F1-score of 0.624, reflecting the loss of training data. Weighted cross-entropy and alpha-balanced focal loss produced F1-scores of 0.669 and 0.674, respectively. Although these approaches improved recall for Classes 3 and 4, they also increased false positives. The models more readily classified comments into Classes 3 and 4, but did so inaccurately, resulting in decreased F1-score. Overall, the alternative methods reduced overall performance, suggesting that more targeted strategies are required to address the imbalance.

4. Discussion and Conclusions

AVW-Space enables students to leave comments that are automatically classified into quality classes using a weighted random forest classifier developed in 2020 (Mohammadhassan et al., 2020). Our goal was to determine whether DL techniques outperformed the existing ML method and whether fine-tuning and weighted techniques could optimise performance.

Training of the 2020 ML model utilised a smaller dataset. To ensure a fair comparison, we retrained it using the expanded dataset, yielding marginal improvements in the F1-score. We then evaluated six transformer-based DL models: BERT, RoBERTa, ALBERT, ELECTRA, GPT, and GPT-2. All DL F1-scores outperformed the retrained ML model, with RoBERTa emerging as the best performer. Our findings confirm RoBERTa's superior performance over other DL and ML approaches, aligning with prior research (Mohammadhassan et al., 2020; Sy et al., 2024). Fine-tuning RoBERTa by adjusting hyperparameters (dropout rate, hidden layers, batch size, number of epochs, and learning rate) further refined its performance. Weighted testing used the configuration that obtained the highest F1 score.

A limitation of our study is the imbalanced dataset, with most comments falling into Class 2. Differentiating Class 3 and 4 from Class 2 proved challenging. The factual nature of Class 3 comments and the importance of the aspect in some Class 4 comments likely contributed to this issue. Attempts to address class imbalance using alternative loss functions showed mixed results. Focal loss performed similarly to standard cross-entropy, while weighted cross-entropy and alpha-balanced focal loss reduced overall accuracy: although they increased classification into Class 3 and 4 over Class 2, these predictions were often incorrect. Undersampling also decreased performance, as the benefit of a balanced class distribution did not compensate for the information lost.

As AVW-Space evolves with new features, the nature and quality of comments will change. Future research should consider synthetic data generation or selective subsampling of recent comments to adapt to these evolving standards. Additionally, the fine-tuned model is specialised for comments on presentation skills, meaning its performance may not generalise to other topics. Further testing is needed to assess its effectiveness outside the trained domain. Also, aggregated performance metrics obscured class performance variations, suggesting that alternative evaluation methods, such as class-wise thresholds or tailored metrics, could provide a better assessment of model performance. Also, annotator bias remains a concern, emphasising the need for standardised labelling to enhance classification consistency. Incorporating video transcripts might improve classification performance but could also reduce generalisability.

Overall, this research identified a fine-tuned RoBERTa model as the best-performing DL approach, achieving an increased performance across precision, recall, F1-score and average cost. The fine-tuned RoBERTa model obtained an improved F1-score across all classes, specifically presenting vast improvements for Class 5 comments. These results highlight the effectiveness of DL techniques in enhancing comment quality classification within AVW-Space. Future work will focus on refining classification strategies, improving evaluation metrics, and developing efficient alternatives to support deployment.

Acknowledgments

We thank the members of the Intelligent Tutoring Group for their support

References

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. International Conference on Learning Representations.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), Proc. Conf. the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol 1, (pp. 4171–4186). Association for Computational Linguistics.

Dimitrova, V., & Mitrovic, A. (2022). Choice Architecture for Nudges to Support Constructive Learning in Active Video Watching. Int. Journal of Artificial Intelligence in Education, 32(4), 892–930.

Dimitrova, V., Mitrovic, A., Piotrkowicz, A., Lau, L., & Weerasinghe, A. (2017). Using Learning Analytics to Devise Interactive Personalised Nudges for Active Video Watching. 25th Conference on User Modeling, Adaptation and Personalization, 22–31.

Hugging Face. (2025). Hugging Face Hub Documentation. https://huggingface.co/docs/hub/index

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations. Proc. 8th Int. Conf. Learning Representations.

Li, Y., Raković, M., Dai, W., Lin, J., Khosravi, H., Galbraith, K., Lyons, K., Gašević, D., & Chen, G. (2023). Are deeper reflectors better goal-setters? AI-empowered analytics of reflective writing in pharmaceutical education. Computers and Education: Artificial Intelligence, 5, 100157.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv-1907.

Mitrovic, A., Dimitrova, V., Lau, L., Weerasinghe, A., & Mathews, M. (2017). Supporting Constructive Video-Based Learning: Requirements Elicitation from Exploratory Studies. In E. André, R. Baker, X. Hu, Ma. M. T. Rodrigo, & B. du Boulay (Eds.), Artificial Intelligence in Education (pp. 224–237). Springer International Publishing.

Mitrovic, A., Dimitrova, V., Weerasinghe, A., & Lau, L. (2016). Reflexive Experiential Learning: Using Active Video Watching for Soft Skills Training. In W. Chen & others (Eds.), Proc. 24th Int. Conf. on Computers in Education (pp. 192–201). Asia-Pacific Society for Computers in Education.

Mitrovic, A., Gordon, M., Piotrkowicz, A., & Dimitrova, V. (2019). Investigating the Effect of Adding Nudges to Increase Engagement in Active Video Watching. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), Artificial Intelligence in Education (pp. 320–332).

Mohammadhassan, N., Mitrovic, A., & Neshatian, K. (2022). Investigating the effect of nudges for improving comment quality in active video watching. Computers & Education, 176, 104340.

Mohammadhassan, N., Mitrovic, A., Neshatian, K., & Dunn, J. (2020). Automatic assessment of comment quality in active video watching. In H.-J. So, Ma. M. Rodrigo, J. Mason, A. Mitrovic, D. Bodemer, W. Chen, Z.-H. Chen, B. Flanagan, M. Jansen, R. Nkambou, & L. Wu (Eds.), 28th International Conference on Computers in Education, (pp. 1–10). APSCE.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. 1–12.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., & others. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8), 9.

Sy, C. Y., Maceda, L. L., Canon, M. J. P., & Flores, N. M. (2024). Beyond BERT: Exploring the Efficacy of RoBERTa and ALBERT in Supervised Multiclass Text Classification. Int. Journal of Advanced Computer Science & Applications, 15(3).

Wang, Z., Hsu, C.-Y., Horikoshi, I., Li, H., Majumdar, R., & Ogata, H. (2024). Classifying Self-Reflection Notes: Automation Approaches for GOAL System. In A. Kashihara & others (Eds.),32nd Int. Conf. Computers in Education. APSCE.

Wulff, P., Mientus, L., Nowak, A., & Borowski, A. (2023). Utilizing a Pretrained Language Model (BERT) to Classify Preservice Physics Teachers' Written Reflections. Artificial Intelligence in Education, 33, 439–466.

Yousef, A. M. F., Chatti, M., & Schroeder, U. (2014). The State of Video-Based Learning: A Review and Future Perspectives. International Journal on Advances in Life Sciences, 6, 122–135.

Zhang, D., Zhou, L., Briggs, R. O., & Nunamaker, J. F. (2006). Instructional video in e-learning: Assessing the impact of interactive video on learning effectiveness. Information & Management, 43(1), 15–27.

Exploring the Relationship Between System Operation Behaviors and Learning Achievement in Agricultural Education

Haiqiao LIU[a]*, Tsubasa MINEMATSU[b], Chengjiu YIN[c],Sijie XIONG[a] & Atsushi SHIMADA[a]

[a]*Faculty of Information Science and Electrical Engineering, Kyushu University, Japan*
[b]*Faculty of Computer Science and Systems, Kyushu Institute of Technology, Japan*
[c]*Department of Institute for Information Technology, Kyushu University, Japan*
**liu.haiqiao.966@s.kyushu-u.ac.jp*

Abstract: With the growing integration of digital technologies into agricultural education, learning support systems play a pivotal role in enhancing students' understanding of environmental and field-based concepts. While previous studies have shown their effectiveness, less attention is given to the specific student behaviors that contribute to improvements. This study investigates which system operation behaviors are strongly associated with learning achievements in a high school agricultural context. Behavioral log data are collected from a class using the Field Environment Digest System, a web-based platform visualizing real-time field data from sensor networks. Students are grouped into quartiles based on test performance, and their usage behaviors are analyzed comparatively. Results show that high-performing students engage more in interactive operations such as switching chart views, exploring legends, and adjusting sensor variables. These patterns suggest that deliberate, exploratory interactions are linked to better outcomes and can inform instructional strategies and system designs that promote deeper understanding.

Keywords: Agricultural education, Learning support system, System interaction behavior, Student engagement, Educational technology

1. Introduction

With the rapid advancement of digital technologies in education, learning support systems are increasingly used across disciplines, including agricultural education. These systems often integrate real-time environmental monitoring, data visualization, and mobile interfaces to provide contextualized, interactive learning experiences (Subbarao et al., 2024). In agricultural high schools, they help bridge the gap between theory and practice, enabling students to develop a concrete understanding of environmental and farming concepts (Lin et al., 2017).

Despite their growing adoption, many studies focus on overall learning effectiveness without examining how students interact with specific system features (Cao, 2021). Understanding not only whether a system improves performance, but also how it is used to achieve better outcomes, remains a key gap. Identifying behaviors that contribute most to learning can guide pedagogy and system design.

In our previous study, we introduced the Field Environment Digest System in high school agricultural classes, finding that students who used it achieved higher test scores. However, the mechanisms behind this impact remain unclear. This study analyzes students' log-recorded actions in the system—such as switching sensors, adjusting chart views, and interacting with visualizations—to determine which operations are significantly associated with learning achievement. By correlating behavioral logs with test performance, we aim to identify effective patterns of use. The findings can inform technology design and provide actionable feedback for instructors. Therefore, this study seeks to answer the following research question:1. Does the use of the agricultural learning system lead to improved student

academic performance?2. Which system operation behaviors contribute most to differences in student academic performance?

2. Related Work

Learning analytics and educational data mining are widely used to analyze students' interaction behaviors in digital learning environments, aiming to uncover patterns that improve outcomes and guide instruction (Kaur & Dahiya, 2023). Research shows that behavioral features such as content access frequency, revisiting materials, and engagement in assessments often correlate positively with grades or test performance (Liu et al., 2022). Machine learning models predict success from interaction logs (Zeng, 2023), while temporal analyses examine how action sequences influence outcomes (Prasertisirikul et al., 2022).

However, most studies are in higher education or large-scale online systems such as MOOCs and LMS courses (Sun et al., 2018), limiting applicability to secondary agricultural education. Fine-grained behavioral analysis is particularly important here because tasks involve multi-step decision-making, real-time interpretation of environmental data, and integration of theory with practice (Zachariou et al., 2025). Such analysis can reveal behaviors missed in overall usage counts and help teachers provide targeted support (Klerkx, 2022).

Agricultural education often occurs in authentic field settings, where students manage both conceptual understanding and hands-on practice (Nur Dinie & Enio, 2023). These contexts create unique interaction patterns, such as interpreting sensor data to adjust cultivation parameters (Wan Jusoh et al., 2023). Yet, research on high school agricultural education rarely examines detailed usage behaviors and their impact on achievement (Baideldinova et al., 2022). This study analyzes system operation logs from high school students using a mobile-based field environment monitoring system. By linking specific interactive operations with test performance, it aims to identify behavior patterns that influence learning outcomes and inform the design of agricultural learning technologies.

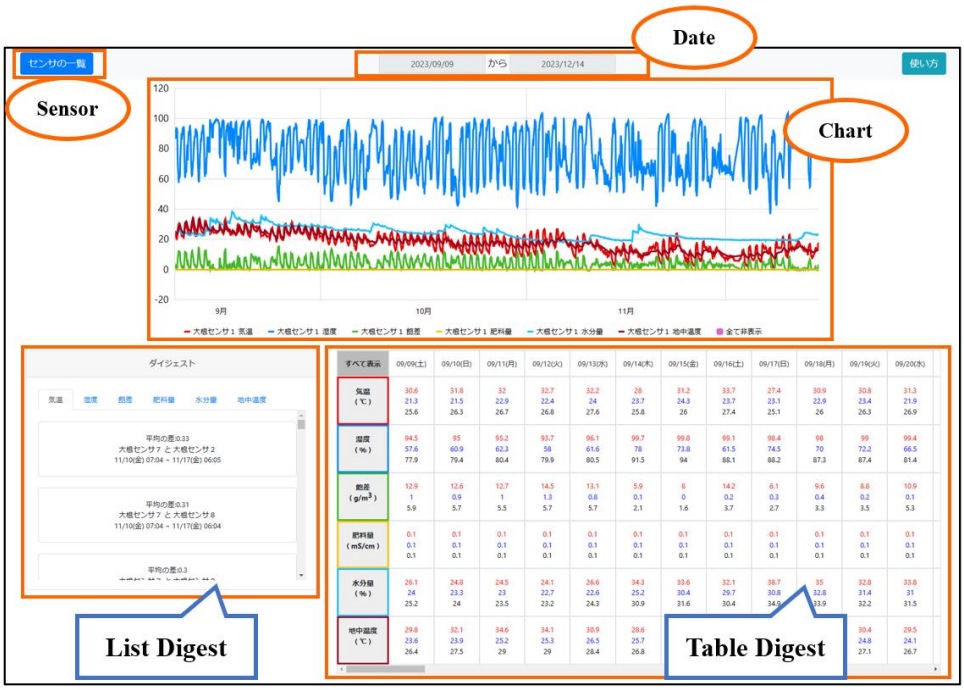


Figure 1. Field Environment Digest System Interface

3. System Description

In our research, we introduce the "Field Environment Digest System" to enhance student learning. The system aggregates sensor-derived agricultural field data and presents it in a simplified format (Shiga et al., 2023), providing time-series data from sensors such as

temperature and humidity. Students explore this data through interactive line graphs and statistical summaries, helping them connect theoretical knowledge with real-world conditions. A core function is the logging module, which records all user operations in real time. This study analyzes specific log events to identify behavioral patterns linked to learning performance, counting each user action once per occurrence to calculate frequency.

The stored log data is shown in Table 1. The id column indicates the ID of each log entry; user_id assigns a unique ID to each student; created_at records the timestamp; action identifies the type of operation performed; time_from and time_to record the period viewed; sensor specifies the sensor involved; and category shows the type of data explored. For example, in the first and second rows, user 235 adjusts the displayed time range of sensor xx-yy and maintains it for a certain duration; in the second row, the user examines the period of change related to moisture content. In the third row, the user checks moisture content via the aggregate summary, with the viewed period indicated by time_from and time_to. In the fourth row, the user operates the saturation content chart, with the viewed period similarly derived from the same columns.

Table 1. *Data to Manage Operation Logs*

id	user_id	created_at	action	time_from	time_to	sensor	category
1	235	1697761492	chart_change	1694703600	1697727600	xx-yy	temp
2	235	1697761503	chart_legend	1696724399	1696724419	xx-yy	humidity
3	235	1697761506	date_from	1697284929	1697306428	xx-yy	moisture
4	235	1697761509	date_to	1695131013	1695134613	xx-yy	saturation
5	235	1697761529	sensor_select	1695975878	1696001721	xx-yy	electric

4. Method

This study is conducted at an agricultural high school in Japan with 34 first-year students from the same intact class enrolled in an agricultural experiment course. The same teacher delivers all lessons to ensure identical classroom conditions. To minimize confounding variables, all students follow the same curriculum before the experiment, report no prior experience with the system in a pre-test survey, and use identical iPads after a short orientation session.

During the experimental phase, students use the Field Environment Digest System in the last 20 minutes of class to explore sensor-derived data (e.g., humidity, temperature, soil moisture) from their managed plots. A post-test, based on validated items from previous examinations, assesses factual recall, conceptual understanding, and application, with emphasis on conceptual understanding. System logs record six types of operations—Chart_change, Chart_legend, Date_from, Date_to, Digest_table, and Sensor_select—representing distinct forms of data exploration and visualization.

Data analysis uses Pearson correlation and linear regression to examine the relationship between operation frequency and test scores, with regression slope and coefficient of determination indicating predictive strength. Correlations between each behavior type and scores identify high-impact interactions. Students are grouped into quartiles by performance, and the Kruskal–Wallis H test detects group differences, followed by Bonferroni-adjusted Mann–Whitney U tests for post hoc comparisons.

5. Results

5.1 Analysis of the Relationship Between System Operations and Learning Outcomes

Analysis for RQ1 shows a strong positive relationship between overall system use and academic performance. Figure 2 plots test scores against the number of system operations, showing a clear upward trend: students performing more than 30 operations often score above 90, while those with fewer than 20 rarely exceed 80. Operation frequency correlates highly

with test scores (r = 0.89, p < 0.001), with a regression slope of 1.03 and R^2 of 0.79, indicating that 79% of score variance is explained by usage. Frequent interaction is linked to higher scores, indicating that consistent engagement enhances understanding of agricultural concepts (Salama & Aly, 2024; Lin, 2025).

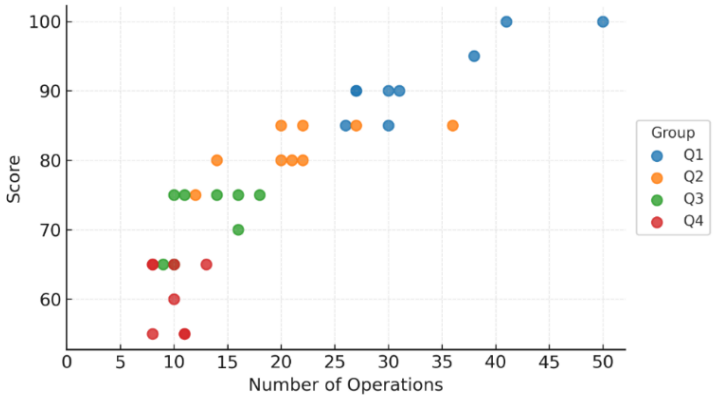


Figure 2. Relationship Between Score and Number of Operations

Table 2. *Correlation Analysis Between Test Score and System Operation Count*

Metric	Pearson r	p	Slope	R^2
Value	0.89	0.001***	+1.03	0.79

***p < .001

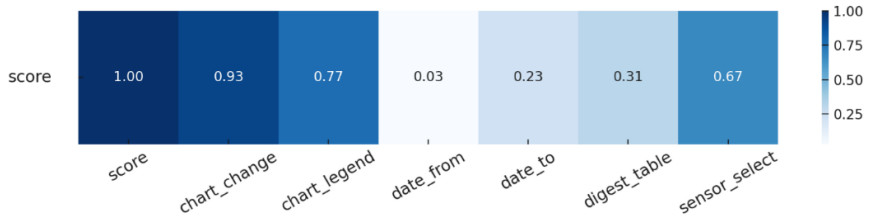


Figure 3. Pearson correlation coefficients between six types of system operations and students' test scores.

Further analysis (Figure 3) identifies chart_change (r = 0.93), chart_legend (r = 0.77), and sensor_select (r = 0.67) as the behaviors most associated with performance, while date_from (r = 0.03), date_to (r = 0.23), and digest_table (r = 0.31) show weaker associations. These findings suggest that active exploration and comparison of environmental data are more strongly linked to higher performance than procedural actions, underscoring the importance of meaningful interaction in digital learning tools (Klerkx, 2022). Teachers can integrate short, targeted system-based tasks into coursework to encourage these behaviors, while designers should ensure easy access to high-value features, such as chart comparisons and multi-sensor views, to promote sustained, high-quality engagement (Salama & Aly, 2024; Lin, 2025).

5.2 Differences in System Operation Behaviors

Students are divided into four quartiles by test scores (Figure 2) to compare behavior frequencies. A Kruskal–Wallis H test reveals significant differences for chart_change (H = 31.016, p < .001), chart_legend (H = 20.459, p < .001), and sensor_select (H = 13.342, p = .004), while date_from, date_to, and digest_table show no differences (Table 3).

Post-hoc Bonferroni-adjusted Mann–Whitney U tests (Table 4) show chart_change differs in nearly all pairwise comparisons (p < .001), making it a robust performance indicator. For chart_legend, differences appear between Q1 and Q3/Q4 (p = .003), suggesting it distinguishes top performers. Sensor_select differs only between Q1 and Q4 (p = .015),

indicating its variation is most evident when contrasting the highest- and lowest-performing groups.

Table 3. *Kruskal–Wallis Test Results for Operation Behavior Differences Among Performance Groups*

Behavior	chart_ change	chart_ legend	date_ from	date_ to	digest_ table	sensor_ select
H	31.016	20.459	1.979	2.631	3.223	13.342
p	0.001***	0.001***	0.577	0.452	0.358	0.004**

** p <.01, *** p < .001

Table 4. *Bonferroni-Adjusted Pairwise Comparisons of Operation Behaviors Between Quartile Groups*

Behavior	Q1 vs Q2	Q1 vs Q3	Q1 vs Q4	Q2 vs Q3	Q2 vs Q4	Q3 vs Q4
chart_change	0.002**	0.003**	0.003**	0.003**	0.003**	0.006**
chart_legend	0.076	0.003**	0.003 **	0.383	0.131	1.000
sensor_select	1.000	0.130	0.015 *	0.871	0.078	1.000

* p <.05, ** p < .01

For RQ2, correlations identify chart_change (r = 0.93), chart_legend (r = 0.77), and sensor_select (r = 0.67) as the most performance-related behaviors. Non-parametric tests confirm chart_change as a consistent differentiator, chart_legend as moderately distinctive for top performers, and sensor_select as significant only between the highest- and lowest-performing groups. High achievers engage more in exploratory actions—switching chart views, toggling variables, and comparing sensor outputs—that integrate theoretical concepts with real-time field data (Reddy & Chandu, 2025).

6. Discussion and Conclusion

This study relates system-operation logs to test performance in high school agricultural learning. Grouping students by scores shows that *chart_change* and *chart_legend* are most strongly linked to higher achievement, indicating that visually guided exploration supports deeper understanding. For design, prioritizing features that elicit meaningful comparisons rather than simply increasing clicks can enhance learning value. For instruction, behavior logs can serve as early indicators of engagement, enabling timely support for students who underuse core visualization tools; encouraging low-performing students to practice these features may help close performance gaps.

Across analyses, data-comparison and visualization behaviors show greater learning value than procedural operations. Actions such as chart_change and chart_legend—which prompt students to juxtapose variables and interpret trends—align strongly with achievement, whereas date_from, date_to, and digest_table display weak correlations and little variation across groups. Unlike LMS contexts where navigation often dominates, impactful actions here require synthesizing multiple environmental factors in real time (Issa et al., 2024). This domain-specific demand supports designs and teaching strategies that foreground cognitively rich, exploratory visual features.

Limitations include analysis of a single session within one subject context, which may constrain generalizability. Future work should analyze longer deployments across varied settings and experimentally manipulate interaction types to test causality. Overall, the type—not merely the amount—of system interaction matters for learning in agricultural education.

Acknowledgments

This work was supported by JST CREST Grant Number JPMJCR22D1, JSPS KAKENHI Grant Number JP22H00551, and MEXT "Innovation Platform for Society 5.0" Program Grant Number JPMXP0518071489, Japan.

References

Cao, Y. X. (2021). An Analysis of Application of Interact Function in Higher Education Internet System—A Case Study of Canvas Learning Management System. DEStech Transactions on Economics, Business and Management.

Field, A. (2024). Discovering statistics using IBM SPSS statistics. Sage publications limited.

Guanin-Fajardo, J. H., Guaña-Moya, J., & Casillas, J. (2024). Predicting Academic Success of College Students Using Machine Learning Techniques. Data, 9(4), 60.

Issa, A., Odedeyi, T., & Darwazeh, I. (2024). IoT-Driven Precision Agriculture using Communication Technologies for Crop Quality and Real-time Environmental Monitoring. 2024 IEEE 30th International Conference on Telecommunications (ICT), 1-5.

Jusoh, W., Daud, M., Azizan, M., & Zakaria, S. (2023). Sensor Learning Application for Precision Agriculture. Journal of Technology and Humanities.

Kaur, K., & Dahiya, O. (2023, February). Role of educational data mining and learning analytics techniques used for predictive modeling. In 2023 3rd International Conference on Innovative Practices in Technology and Management (ICIPTM) (pp. 1-6). IEEE

Lin, Y. (2025). The Effectiveness of Developing Cloud-Based Agricultural Environmental Sensing System to Support Food and Agriculture Education in Elementary School. International Journal of Information and Education Technology.

Lin, Y. P., Chang, T. K., Fan, C., Anthony, J., Petway, J. R., Lien, W. Y., ... & Ho, Y. F. (2017). Applications of information and communication technology for improvements of water and soil monitoring and assessments in agricultural areas—A case study in the taoyuan irrigation district. Environments, 4(1), 6.

Liu, Y., Fan, S., Xu, S., Sajjanhar, A., Yeom, S., & Wei, Y. (2022). Predicting student performance using clickstream data and machine learning. Education Sciences, 13(1), 17.

M.I., N., & M.S.K., E. (2023). The Influence of Infrastructure in Agricultural Subjects' Practical Learning Environment on ASPer UPM Students' Interest and Field of Study Selection. Journal of Technical Education and Training.

Prasertisirikul, P., Laohakiat, S., Trakunphutthirak, R., & Sukaphat, S. (2022, March). A predictive model for student academic performance in online learning system. In 2022 international conference on digital government technology and innovation (dgti-con) (pp. 76-79). IEEE.

Reddy, M. B., Chandu, M., Narayana, M. S., & Anand, M. (2025). Smart Agriculture & System Advancements in Crop Field Intigrating Machine Learning and Agriculture Data. IJSAT-International Journal on Science and Technology, 16(2).

Rohani, N., Rohani, B., & Manataki, A. (2024). ClickTree: A Tree-based Method for Predicting Math Students' Performance Based on Clickstream Data. arXiv preprint arXiv:2403.14664.

Salama, W., & Aly, M. (2024). Web-Based Integrated School Management System for Enhanced Student-Teacher and Resource Accessibility. *2024 34th International Conference on Computer Theory and Applications (ICCTA)*, 115-122.

Shiga, K., Minematsu, T., Taniguchi, Y., Okubo, F., Shimada, A., & Taniguchi, R. I. (2022, August). Development and Evaluation of a Field Environment Digest System for Agricultural Education. In IFIP World Conference on Computers in Education (pp. 87-99). Cham: Springer Nature Switzerland.

Subbarao, S. C., Patel, M., Lee, A., & Lee, H. (2024, April). Cloud-enabled mobile app architecture for water quality monitoring utilizing machine learning algorithms. In 2024 IEEE Opportunity Research Scholars Symposium (ORSS) (pp. 66-69). IEEE

Sun, S., Qian, X., Mu, L., Zan, H., & Zhang, Q. (2018). Performance Prediction Based on Analysis of Learning Behavior. In Data Science: 4th International Conference of Pioneering Computer Scientists, Engineers and Educators, ICPCSEE 2018, Zhengzhou, China, September 21-23, 2018, Proceedings, Part II (pp. 632-644). Springer Singapore.

Togaibayeva, A., Ramazanova, D., Yessengulova, M., Yergazina, A., Nurlin, A., & Shokanov, R. (2022, September). Effect of mobile learning on students' satisfaction, perceived usefulness, and academic performance when learning a foreign language. In Frontiers in Education (Vol. 7, p. 946102). Frontiers Media SA.

Wang, F., Yuizono, T., Wang, T. Y., Kim, E., & Lu, Y. (2023, January). Integrating reflection into a mobile-assisted reading program for learning English as a second language in China. In Frontiers in Education (Vol. 7, p. 1067523). Frontiers Media SA.

Zachariou, M., Adenuga, A., & Jack, C. (2025). Education and Farmers' Environmental Performance: A Systematic Literature Review. Sustainability.

CodeRunner Agent: Integrating AI Feedback and Self-Regulated Learning to Support Programming Education

Huiyong LI[a] & Boxuan MA[b*]
[a]*Research Institute for Information Technology, Kyushu University, Japan*
[b]*Faculty of Arts and Science, Kyushu University, Japan*
*boxuan@artsci.kyushu-u.ac.jp

Abstract: The emergence of Large Language Model (LLM) tools has revolutionized programming education, demonstrating considerable efficacy in providing instant, personalized feedback. However, most existing LLM-based tools focused on direct coding assistance, and neglected the cultivation of self-regulation skills. Moreover, many of these tools operate independently from institutional learning management systems, which creates a significant pedagogical disconnect that limits the ability to leverage contextual learning materials and exercises for generating tailored, context-aware feedback. To address these challenges, we developed CodeRunner Agent, an LLM-based programming assistant. CodeRunner Agent enhances students' self-regulated learning by providing strategy-based AI feedback and embedding within Moodle system. Additionally, it empowers educators to customize AI-generated feedback by incorporating detailed context from lecture materials, programming questions, student answers, and execution results. This integrated approach, emphasizing self-regulation skill development with contextual awareness, offers promising avenues for data-driven enhancements in programming education with AI.

Keywords: LLM-powered tools, programming education, self-regulated learning, personalized feedback, CodeRunner

1. Introduction

Innovation within higher education has been accelerated dramatically by digital transformation strategies and evolving pedagogical approaches. Programming education, in particular, has attracted an ever-growing number of students with wider recognition of the critical importance of computer science literacy. This growth has created significant pedagogical challenges in providing timely, individualized assistance to each student (Yilmaz & Yilmaz, 2023). Large Language Models (LLMs) have emerged as a promising technological intervention. A variety of LLM-powered programming assistants have been developed to provide timely coding guidance and suggestions, potentially transforming traditional instructional approaches and feedback mechanisms in programming education (Pankiewicz & Baker, 2023).

While these advancements present exciting opportunities for personalized learning, they also raise concerns about over-reliance and reduced development of self-regulated learning (SRL) and problem-solving skills (Humble et al., 2024; Prasad & Sane, 2024). Easy access to answers may encourage surface-level understanding, yet many LLM tools lack pedagogical guardrails to promote active engagement and SRL strategies (Sun et al., 2024). A further limitation is the lack of integration with institutional Learning Management Systems (LMS) like Moodle. Effective feedback is highly context-dependent, relying on alignment with lecture materials and assignment specifications (Kaur & Chahal, 2024; Kazemitabaar et al., 2024). Without LMS integration, tools cannot leverage rich contextual data or provide educators with insights into student interactions and learning outcomes (Ma et al., 2024a, 2024b). Therefore, it's essential to ensure that the AI-generated feedback is both relevant and aligned with course objectives by integrating LLM-based tools into the LMS environment.

To address these limitations, we developed CodeRunner Agent (Li & Ma, 2025), an LLM-powered programming assistant embedded within CodeRunner, an open-source Moodle

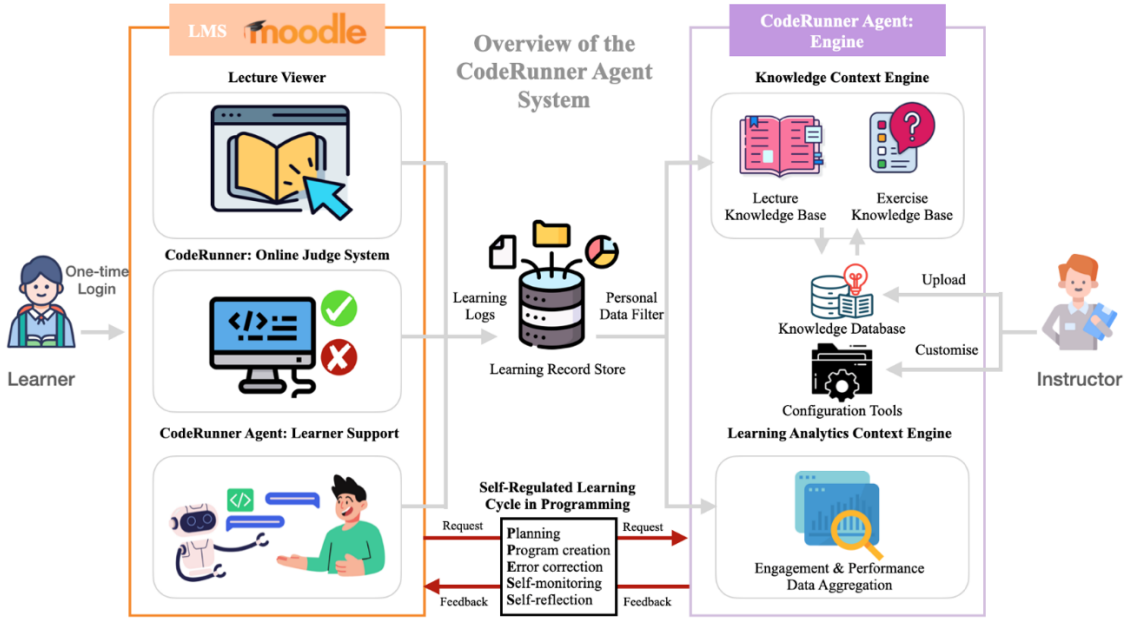


Figure 1. Overview of the CodeRunner Agent Programming Support Environment.

plugin for executing and assessing student-submitted code. CodeRunner Agent leverages the comprehensive context available within an LMS environment from learning logs and enhances students' self-regulated learning in programming education. Specifically, it enhances the delivery of individualized feedback by combining students' knowledge level, self-regulation skills, and strategy-focused LLM-based scaffoldings. It also tracks AI feedback and subsequent actions in students' coding behaviors. This provides valuable insights into learning behaviors and the effectiveness of AI-assisted feedback. By addressing the challenges of scalability, contextual feedback, and skill enhancement, our approach contributes to a deeper understanding of AI's role in modern education.

2. CodeRunner Agent

2.1 Overview of CodeRunner Agent

The framework of CodeRunner Agent is shown in Figure 1. It is developed and integrated to scaffold programming education with an LLM-based assistant in the Moodle LMS. The framework contains a Lecture Viewer, a CodeRunner plugin, and a CodeRunner Agent.

Lecture Viewer is provided to deliver the lecture slides by instructors and access the lecture slides by learners. The operations of the lecture viewer are recorded in the form of Experience API (or xAPI) statements. Then, the xAPI statements are stored in the Learning Record Store (LRS). *CodeRunner* is an open-source plugin that can be imported to Moodle (Lobb & Harlow, 2016). Learners can write programming codes and receive automated grades by running it in a series of tests. A logging plugin named "Logstore xAPI" is used to record the results and send them to the LRS. For example, the code for the test, got output, correct status, and mark reward will be logged in LRS if learners run the test code once. *CodeRunner Agent* is an AI-powered tool to support learners' SRL and teachers' customized designs for LLM-based feedback in programming education. It can be embedded into the Moodle LMS and executed with the CodeRunner plugin. Learners can receive general-purpose and programming-specific regulatory strategy feedback based on OpenAI API capabilities and strategy-focused prompt design. Specifically, the agent provides scaffolding hints while not directly generating code solutions by code guardrails design. A specific LLM prompt is used to rewrite the response by giving salient features of the code when code blocks are detected in the scaffolding hints. The interactions with the agent are automatically tracked as xAPI statements and stored in the LRS (e.g., the request type, request time, exercise ID).

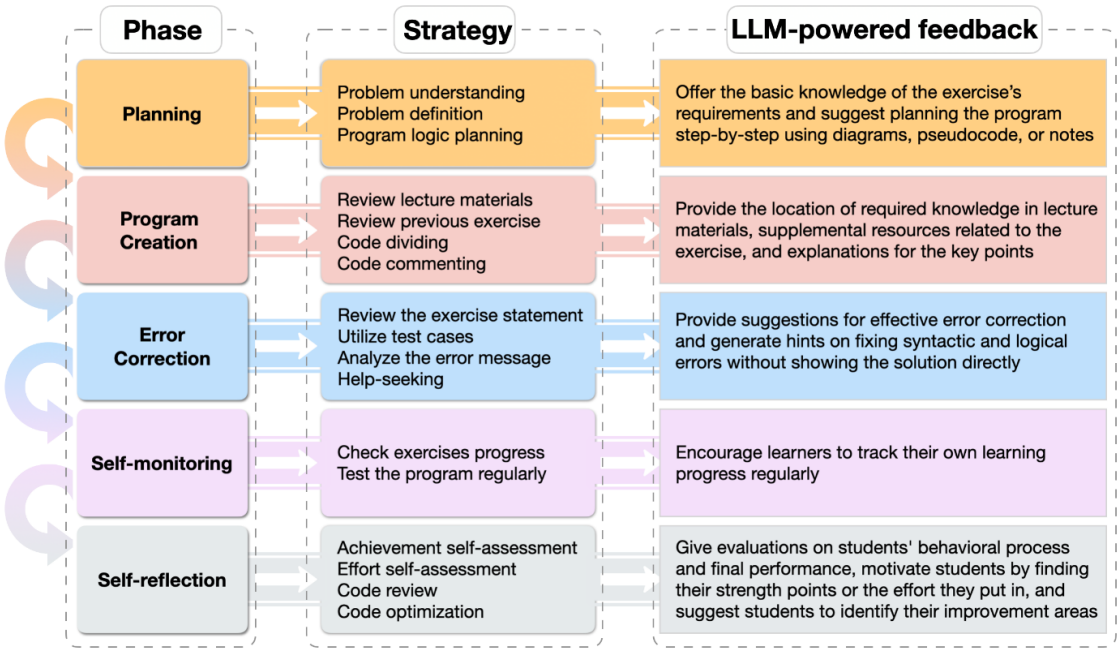


Figure 2. Phase, Target SRL strategy, and LLMs Feedback in the CodeRunner Agent.

The CodeRunner Agent's core intelligence contains two context engines: Learning Analytics Context Engine (LACE) and Knowledge Context Engine (KCE). LACE calculates and aggregates learners' engagement metrics (i.e., time spent, attempt frequency) and performance metrics (i.e., success rates, error patterns) from LRS data after filtering personal data (i.e., name, gender, and email). LACE is integrated as contextual embedding of self-regulation behavioral strategies with the feedback of the agent. KCE handles both lecture and exercise knowledge bases. The lecture knowledge base manages the key programming concepts and the dependencies between the concepts from lecture materials. The exercise knowledge base categorizes exercises by concept, difficulty, solution, and typical mistakes. KCE is used as learner knowledge embedding in the agent.

Instructors can upload lecture materials (i.e., concept definition, concept illustration, and annotated examples) and exercises (i.e., problem statement, solution, and test cases) to the context engine. They will be converted to textual knowledge and stored in the knowledge database for future retrieval by the Retrieval Augmented Generation (RAG) technique. More importantly, instructors can update the knowledge bases and customize the parameters of the context engine by using the configuration tools.

2.2 SRL Support Model in CodeRunner Agent

To enhance strategy-based programming learning for novice learners, LLM-powered feedback is integrated into a five-phase model grounded in the Zimmerman's SRL theory (Zimmerman, 2008) and adopts a conceptual framework for regulating learning in programming (Silva et al., 2024). The model adds two CS domain-specific elements to the general three-phase SRL cycle (forethought, performance, and self-reflection). The model contains five phases, named PPESS: *Planning*, *Program creation, Error correction, Self-monitoring,* and *Self-reflection*.

The phase, target SRL strategy, and LLMs support in the CodeRunner Agent are shown in Figure 2. Both general-purpose and programming-specific regulatory strategies are included for SRL scaffoldings in the five-phase PPESS model. *Planning* supports problem analysis and decomposition. *Program creation* assists the implementation of coding solutions using both declarative and procedural knowledge. *Error correction* helps cognitive diagnosis and metacognitive regulation for addressing coding errors. *Self-monitoring* supports ongoing evaluation and adjustment toward task completion during coding and debugging. Finally, *Self-reflection* helps learners assess their performance and improve future learning strategies.

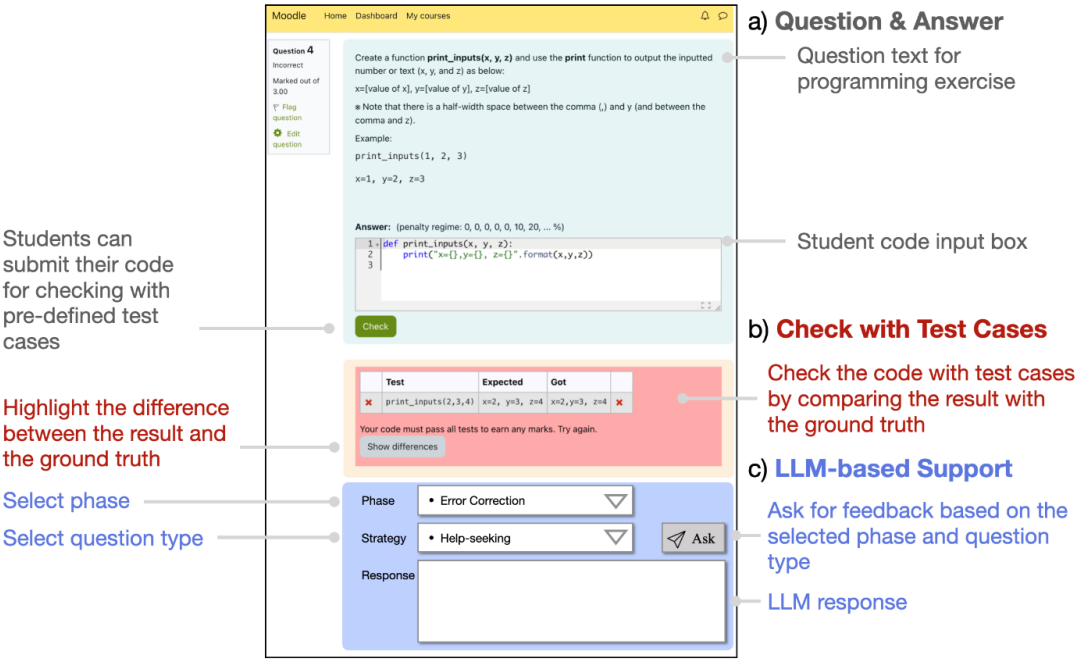


Figure 3. User Interface of Integrated CodeRunner Agent for Learners.

Figure 3 illustrates the user interface of the integrated CodeRunner Agent, which consists of three main components: (a) Question & Answer, (b) Check with Test Cases, and (c) LLM-based Support. The LLM-based support box for the CodeRunner Agent appears below the test results provided by CodeRunner. Learners have the option to select one of five SRL phases like error correction. Additionally, they can select either the question type in general level as default or the question type from one of the specific strategies in the phase like help-seeking. Upon submitting a request, the LLM-based strategy feedback is displayed within the response box.

3. User Experience Interview with Students and Educators

To gain deeper insights into students' and educators' perspectives on the CodeRunner Agent, we conducted semi-structured interviews with four students and two educators after introducing the agent and discussing its capabilities and potential use cases. The student participants included an undergraduate with one month of programming experience (S1) and three graduate students with programming experiences of two years (S2), three years (S3) and six years (S4), respectively. Both educators involved were actively teaching undergraduate programming courses, with teaching experiences of four years (T1) and two years (T2), respectively. Each interview, conducted via Zoom with informed consent, lasted approximately one hour.

3.1 Students Perceptions

- **General Impressions** Students generally held favorable impressions of the agent. S2 and S4 expected the agent to support computational thinking, problem requirement understanding, and code quality improvement more than just code validators. S3 was positive about the multi-phase support provided by the agent, stating that it "quickly helps me understand problems and errors."
- **Over-Reliance Risk Control** As a novice programming learner, S1 appreciated the design of the hint-like feedback and expressed that "hiding the direct answers will help my own thinking". S1 also wanted detailed explanations and visual feedback for error messages. S2 liked the way it "helps develop my thinking skills rather than override them." As advanced programming learners, S4 emphasized the "assistant role" of the agent and preferred the agent "helps me understand the target problem more concisely and why my

understanding was wrong." However, S2 raised concerns that some novice students may struggle even with hints and suggested providing pseudo-code or partial code to ease the learning curve.

- **Contextual Assistance** S4 appreciated the design of the contextual assistant feature since it can provide "more relevant knowledge and suggestions" for the specific learning tasks. Also, the seamless integration of the agent within LMS was highlighted as a key advantage. S3 valued the ability to support the entire learning workflow, from reading materials to coding and checking codes and receiving LLM-assisted feedback.
- **Self-regulated Learning Support** S4 highlighted the critical role of planning support in the agent and wanted to use it regularly. S4 also mentioned that "self-reflection support is highly related to the learning tasks" and expected to use it for code optimization. S2 highlighted that "students can use the agent in different SRL stages, sometimes in a cycle or non-linear manner." S3 expressed that "self-reflection on programming tasks can be widely practiced."
- **Trust and Reliability** Although many students frequently use AI tools, they do not fully trust them. S1 used them regularly for both learning and everyday tasks but viewed them more as virtual freshman peers rather than trustworthy sources, citing issues with complex questions. S4 expressed a more critical perspective, stating that ChatGPT had "about 60%-70% accuracy during my advanced learning tasks" and exhibited weaknesses in logical reasoning. S3 warned that "LLM feedback may not always be accurate or applicable. It may even be misleading in some cases." Across responses, they raised concerns about the low accuracy in advanced tasks, lack of transparency, and the potential for misleading outputs.

3.2 Educators Opinions

Educators provided valuable feedback on the CodeRunner Agent, emphasizing the importance of customization, instructor control over response types, and its pedagogical benefits. T2 noted that CodeRunner Agent is "more beneficial than ChatGPT," explaining that "many students struggle with crafting effective prompts when using ChatGPT" and that "open-ended interactions may raise privacy concerns." In contrast, "the customized feedback from CodeRunner Agent, integrated with SRL, offers contextually relevant support, facilitating actual knowledge acquisition rather than merely solving tasks." T1 further emphasized that "helping students understand the question and concept is more effective for student learning compared to directly offering answers." Additionally, T2 highlighted CodeRunner Agent's potential in a real classroom setting, envisioning its role in facilitating student self-reflection on lecture material and supporting assignment completion more effectively.

Educators expressed concerns about students becoming overly dependent on repeated LLM feedback requests. T2 suggested "setting limits on how frequently students can access such feedback." He further recommended "structuring feedback progressively, beginning with simpler guidance and providing more detailed assistance only when students continue to struggle." Lastly, both educators highlighted the need for a monitoring dashboard to analyze student interactions and LLM responses, aiding instructors in identifying instructional gaps and enhancing teaching strategies.

4. Discussion and Future work

This study aims to develop an LLM-based programming assistant, CodeRunner Agent, that seamlessly integrates with a lecture viewer and CodeRunner plugin in the Moodle LMS. The design is grounded within Zimmerman's SRL theory and targets the freshmen students' programming-specific knowledge acquisition as well as general self-regulation skill development. This study can fill the gap between AI, SRL, and LA by combining LLM-based scaffolding, self-regulation in programming, and learning log-based contextual feedback.

Future work will involve updating and enhancing the CodeRunner Agent by focusing on four key design considerations for LLM-based tools (Kazemitabaar et al., 2024): *D1. Exploiting unique advantages of CodeRunner Agent* (e.g., for deciding when to use the tool,

what are the role and unique advantages of CodeRunner Agent compared to other available resources within the learning ecosystem?); *D2. Refining the CodeRunner Agent interface* (e.g., what are the considerations for CodeRunner Agent to allow users to formulate requests in a way that balances user-friendliness and understandingness with meta-cognitive engagement and SRL skills?); *D3. Balancing the directness of CodeRunner Agent responses* (e.g., how direct should CodeRunner Agent's responses be so that it balances directness and learning engagement, and who should control this balance?); *D4. Supporting trust, transparency, and control* (e.g., once a response from CodeRunner Agent is received, what considerations are needed to ensure accuracy, trust, transparency, and control?).

Moreover, we will enhance educator-agent collaboration by enhancing usability and streamlining the integration of instructor-facing tools like the configuration tool and monitoring dashboard. In addition, we will evaluate the effectiveness of the CodeRunner Agent through short-term pilot studies in actual classroom settings and semester-long experiments conducted across multiple university classes.

Acknowledgements

This work is supported by JSPS KAKENHI Grant Numbers JP24K20903 and JP25K17078.

References

Humble, N., Boustedt, J., Holmgren, H., Milutinovic, G., Seipel, S., & Östberg, A. S. (2024). Cheaters or AI-enhanced learners: Consequences of ChatGPT for programming education. *Electronic journal of e-Learning*, *22*(2), 16-29.

Kaur, A., & Chahal, K. K. (2024). A learning analytics dashboard for data-driven recommendations on influences of non-cognitive factors in introductory programming. *Education and Information Technologies*, *29*(8), 9221-9256.

Kazemitabaar, M., Ye, R., Wang, X., Henley, A. Z., Denny, P., Craig, M., & Grossman, T. (2024). Codeaid: Evaluating a classroom deployment of an llm-based programming assistant that balances student and educator needs. In *Proceedings of the 2024 chi conference on human factors in computing systems, 650,* 1-20.

Li, H., & Ma, B. (2025). Design of AI-Powered Tool for Self-Regulation Support in Programming Education. *arXiv preprint arXiv:2504.03068*.

Lobb, R., & Harlow, J. (2016). Coderunner: A tool for assessing computer programming skills. *ACM Inroads*, *7*(1), 47-51.

Ma, B., Chen, L., & Konomi, S. I. (2024a). Enhancing programming education with ChatGPT: a case study on student perceptions and interactions in a Python course. In *International Conference on Artificial Intelligence in Education* (pp. 113-126). Cham: Springer Nature Switzerland.

Ma, B., Chen, L., & Konomi, S. I. (2024b). Exploring Student Perception and Interaction Using ChatGPT in Programming Education. In *21st International Conference on Cognition and Exploratory Learning in the Digital Age, CELDA 2024* (pp. 35-42). IADIS Press.

Pankiewicz, M., & Baker, R. S. (2023). Large Language Models (GPT) for automating feedback on programming assignments. In *Proceedings of the 31st International Conference on Computers in Education* (pp. 68-77).

Prasad, P., & Sane, A. (2024). A self-regulated learning framework using generative AI and its application in CS educational intervention design. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1* (pp. 1070-1076).

Silva, L., Mendes, A., Gomes, A., & Fortes, G. (2024). What Learning Strategies are Used by Programming Students? A Qualitative Study Grounded on the Self-regulation of Learning Theory. *ACM Transactions on Computing Education, 24*(1), 1-26.

Sun, D., Boudouaia, A., Zhu, C., & Li, Y. (2024). Would ChatGPT-facilitated programming mode impact college students' programming behaviors, performances, and perceptions? An empirical study. *International Journal of Educational Technology in Higher Education*, *21*(1), 14.

Yilmaz, R., & Yilmaz, F. G. K. (2023). The effect of generative artificial intelligence (AI)-based tool use on students' computational thinking skills, programming self-efficacy and motivation. *Computers and Education: Artificial Intelligence*, *4*, 100147.

Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American educational research journal, 45*(1), 166-183.

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Trial Analysis on Behavioral Logs to Identify Risk Tactics and Infer Non-Cognitive Skills in Elementary Students

Shiang-Yi CHOU[a*], Taku KAWADA[b] & Yasuhiro SATO[c] & Tadashi INAGAKI[d] & Toru NAGAHAMA[a]
[a]*Tohoku University, Japan*
[b]*Sendai Shirayuri Gakuen, Japan*
[c]*Future Institute Corporation, Japan*
[d]*Tohoku Gakuin University, Japan*
*chou.shiang.yi.p5@dc.tohoku.ac.jp

Abstract: This study explores whether risk-oriented behaviors—such as rapid, browse-like clicking—can be identified in students' learning tactics, and whether non-cognitive skills can be inferred from learning strategies that include such behaviors. Data from 42 students aged 7 to 11 were analyzed using Cronbach's α, the Kruskal–Wallis test, and the Steel–Dwass post hoc test. Results show that students employing the risk-oriented tactic display distinct learning patterns, and that learning strategies are statistically associated with meta-cognition and social competencies. These associations offer insights into how behavioral logs may reflect non-cognitive skills in K–12 education and provide a foundation for future predictive models and targeted interventions.

Keywords: Learning tactic, Learning strategy, Non-cognitive skills, Risk behavior, K-12

1. Introduction

Recent advances in educational analytics have enabled researchers to examine students' learning tactics and strategies through behavioral log data collected from digital platforms such as Moodle and LEAF (Matcha et al., 2019; Ogata et al., 2022).

Learning tactics are specific actions for task completion (e.g., summarizing), while learning strategies are broader patterns formed by combining these tactics to achieve learning goals (Malmberg et al., 2010; Weinstein et al., 2000). While the use of diverse tactics from learning strategy is often associated with improved self-regulation and academic performance (Fan et al., 2021; Matcha et al., 2019), certain tactics—such as rapid and frequent clicking—may reflect uncertainty or shallow engagement (Huber & Bannert, 2023). However, existing studies predominantly focus on university students, leaving open questions about how elementary school students interact with digital learning tools.

Non-cognitive skills—including motivation, meta-cognition, and social competencies—are also increasingly recognized as critical to student development and academic success in K–12 education (Higgins et al., 2005; Kato, 2020). Deficiencies in these areas have been linked to disengagement, ineffective learning strategies, and difficulties in social interaction (Zynuddin et al., 2023). For instance, students with low motivation are more likely to exhibit signs of academic disengagement, experience burnout, or even drop out of school.

These non-cognitive constructs play a vital role in sustaining learning engagement, supporting emotional regulation, and fostering interpersonal communication. However, their

connection to cognitive behaviors—especially in digital learning environments—remains underexplored. While prior research has leveraged behavioral log data to infer learning strategies associated with academic performance (Matcha et al., 2019), relatively few studies have investigated whether non-cognitive skills themselves can be inferred from such behavioral patterns.

Previous findings suggest that certain behaviors may reflect risk or instability. For example, (Huber & Bannert, 2023) reported that rapid page transitions may be associated with shallow engagement and low self-regulation, while (Malmberg et al., 2013) found that elementary students who used fewer tactics performed better, implying that tactic overuse may signal strategic indecision. These insights raise the question of whether behavioral log data might offer early indicators of deeper cognitive or non-cognitive issues in younger learners. To address these gaps, this study explores two research questions:

- RQ1: Can elementary students' behavioral log data during digital learning reveal specific risk-oriented tactics, such as rapid, browse-like clicking, and identify learning strategies characterized by the overuse of such tactics?
- RQ2: If such risk-oriented tactics and strategies are identified, can we infer specific non-cognitive skill constructs—such as meta-cognition or social competencies—from the learning strategies that incorporate them?

By investigating these questions, we aim to uncover actionable insights that support early detection of shallow engagement and guide the development of more responsive learning systems for K–12 students.

2. Methods

2.1 Procedure

The methodological procedure involved 42 elementary school students (aged 7-11) engaged in a summarization task appropriate for their age. First, they completed a 33-item non-cognitive skills questionnaire using a 6-point Likert scale. Then, they engaged in a digital learning activity using BookRoll from the LEAF system (Ogata et al., 2022), which recorded their interaction logs. All activities were conducted during regular classes.

After excluding 7 incomplete responses, data from 35 students and 51 learning sequences were analyzed. We evaluated the questionnaire's internal reliability and computed dissimilarities between sequences using Euclidean distance. Learning tactics were identified via clustering of action sequences, and learning strategies were determined by clustering students based on tactic usage. Finally, we examined group differences in four non-cognitive constructs.

2.2 Questionnaire

The questionnaire, adapted from Kato (2020), comprised 33 items across four theoretical constructs: Motivation, Expectancy-Value Theory, Meta-cognition, and Social Competencies.

Each item was rated on a 6-point Likert scale. Motivation was assessed via 12 items aligned with self- determination theory's regulatory styles. Expectancy-value theory was measured with 8 items covering ability beliefs and task values. Meta-cognition included 6 items related to planning and self-monitoring strategies, while Social Competencies were assessed through 7 items addressing interpersonal and prosocial skills. We used Cronbach's α to assess the internal consistency of the four constructs.

2.3 Analysis

We analyzed LEAF learning logs, defining sequences as actions occurring within 5 minutes of each other, which covers about 95% of the observed time gaps (51 sequences from 35 students).

To identify learning tactics, we used the TraMineR package in R to analyze action sequences. Dissimilarities were computed using Euclidean distance based on action frequencies, and hierarchical clustering was performed using Ward's method. The optimal number of clusters was determined using the WeightedCluster package.

To identify learning strategy groups, we calculated three variables for each student: the proportion of sequences in each tactic cluster and the total number of sequences, normalized these to the [0,1] range, and again applied hierarchical clustering with Ward's method based on Euclidean distance. The results were visualized through a dendrogram, which was also used to determine the optimal number of clusters.

Finally, to examine group differences in non-cognitive skills, we used the Kruskal–Wallis test and the Steel–Dwass post hoc test, which are appropriate for small, non-normally distributed samples.

3. Result and Discussion

3.1 RQ1: Detection of Learning Tactics and Strategies

3.1.1 Detection of Learning Tactics from Students' Interaction Logs

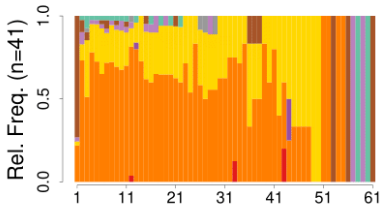
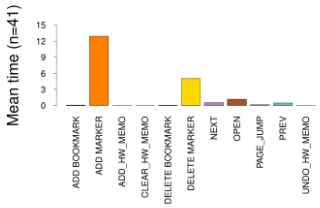
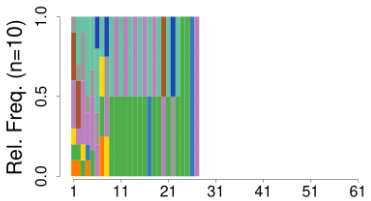
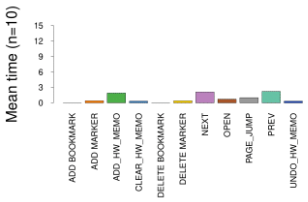
Table 1 presents the characteristics of the two detected tactics (ASW = 0.538).

Tactic 1 – Marker oriented (N = 41, 80.39% of all sequences) featured the longest sequences (median = 16 actions), primarily composed of marker additions and deletions following initial page-opening actions.

Tactic 2 – Browse oriented (N = 10, 19.61% of all sequences) involved short sequences (median = 6.5 actions), mostly consisting of frequent rapid page transitions, and may indicate shallow engagement risk-oriented (Huber & Bannert, 2023).

While both involved browsing actions (e.g., PREV, NEXT), Tactic 1 reflected more targeted activity, whereas Tactic 2 showed diverse but less focused interaction.

Table 1. State distribution within sequences and action frequency for the two learning tactics

Tactics	State Distribution	Frequency of Action
1		
2		

3.1.2 Detection of Learning Strategies Through Students' Use of Tactics

Table 2 summarizes key statistics for the clustering input variables across the three identified learning strategy groups, including the frequency of each learning tactic used and the total number of tactics employed by each student.

Strategy Group 1: Rapid-clicking Strategy (N = 10, 28.57%) uniquely adopted Browse-oriented tactics, operationalized as frequent transitions between distinct actions, and was the only group using both tactics.

Strategy Group 2: Marker Focused Strategy: This was the largest cluster (N=17, 48.57%). Students primarily used the marker tactic and did not engage in browsing, showing the fewest total tactics.

Strategy Group 3: Intensive Marker Strategy: This was the smallest group (N = 8, 22.86%). Students showed the highest use of the marker tactic, with no browsing behavior.

Table 2. Summary statistics for each strategy group (median, 1st and 3rd quartile)

Tactics	Strategy 1	Strategy 2	Strategy 3
1	1.0 (1.0-1.0)	1.0 (1.0-1.0)	2.0 (2.0-2.0)
2	1.0 (1.0-1.0)	0.0 (0.0-0.0)	0.0 (0.0-0.0)
Total	2.0 (2.0-2.0)	1.0 (1.0-1.0)	2.0 (2.0-2.0)

3.2 RQ2: Associations between Learning Strategies and Non-cognitive Constructs

3.2.1 Reliability of the Questionnaire

The internal reliability of the non-cognitive skills questionnaire was evaluated using Cronbach's α.

The Motivation (MOT) scale, with α of 0.7388, met the acceptable threshold, indicating reliable responses across its 12 items. Both the Expectancy-Value Theory (EVT) (α = 0.8590) and Meta-cognition (MC) (α = 0.8582) scales showed high internal consistency, suggesting that the items within each construct measured cohesive aspects of students' learning beliefs and cognitive self-regulation, respectively. Notably, the Social Competencies (SC) scale yielded the highest reliability (α = 0.8677), indicating that participants responded consistently to items related to interpersonal and emotional skills. These results support the appropriateness of using these scales for further analyses.

3.2.2 Group Differences in Non-cognitive Skills Across Learning Strategies

To examine differences in non-cognitive skill constructs across learning strategy groups, we used the combined statistical summary and significance tests, as shown in Table 3.

Figure 1 shows significant differences in learning strategies and non-cognitive constructs. For Meta-cognition, the Marker Focused Strategy showed higher scores than the Rapid- clicking Strategy, suggesting that students in the Marker Focused group tend to have a clearer understanding of their learning methods, while the Rapid-clicking Strategy group's use of both marker and browse tactics may reflect uncertainty or an exploratory stage in their strategy development. The browse tactic, often associated with surface-level learning, could indicate a less stable approach to learning.

For Social competencies, the Intensive Marker Strategy group scored higher than the Rapid-clicking Strategy group, indicating a possible association between the Intensive Marker group's consistent and goal-oriented learning behaviors and stronger social competencies. The Rapid-clicking Strategy group, characterized by lower social interaction and a less focused approach, may need support to improve social engagement and self-regulation.

Overall, the effect sizes for Meta-cognition ($\eta^2 = 0.284$) and Social Competencies ($\eta^2 = 0.218$) are large, indicating that differences between strategy groups are practically meaningful and reinforce the relevance of the statistical findings.

Table 3. Non-cognitive scores across three strategy groups with nonparametric tests

Skills	Strategy Groups Median (Q1-Q3)			Kruskal-Wallis Test		Steel-Dwass Test
	Strategy 1	Strategy 2	Strategy 3	p-value	η^2	
MOT	3.4 (3.1-3.6)	3.8 (3.6-4.2)	3.6 (3.4-4.7)	0.1440	0.059	–
EVT	4.4 (3.8-5.3)	5.4 (5.1-5.8)	5.1 (4.6-5.8)	0.0677	0.106	–
MC	3.2 (3.0-3.6)	4.7 (4.3-5.0)	4.7 (3.6-5.3)	0.0039**	0.284	S1<S2: p = 0.0012**
SC	4.4 (3.6-5.4)	5.3 (4.6-5.7)	5.7 (5.6-5.8)	0.0113*	0.218	S1<S3: p = 0.0116*
Score	9.5 (8.2-18.0)	13.0 (9.0-20.0)	11.0 (7.5-17.0)	–	–	–

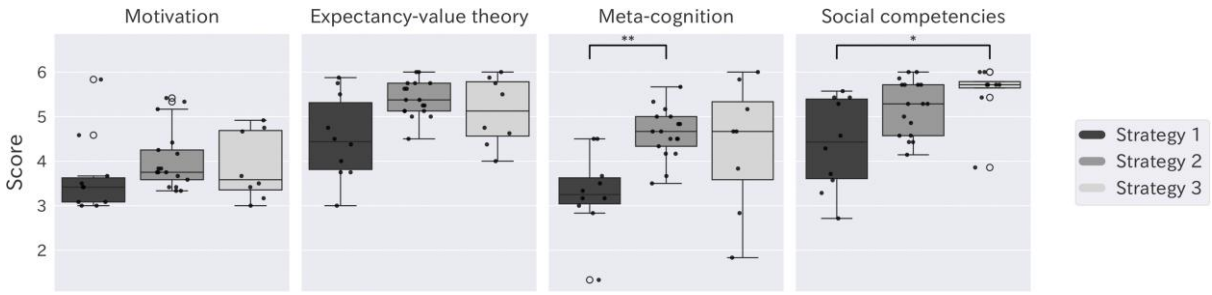


Figure 1. Distribution of strategy groups with respect to the non-cognitive four constructs

4. Conclusion, Limitation and Implication

4.1 Conclusion

This study explored two research questions related to elementary students' cognitive behaviors and non-cognitive skills. We observed that students had developed preferences for certain learning tactics and strategies. Notably, in learning tactics, we identified the marker-based highlighting tactic and the browse-oriented tactic. The rapid and frequent actions in the browse-oriented tactic align with shallow engagement patterns described by (Huber & Bannert, 2023), with frequent page transitions reflecting a less stable or goal-directed approach. In learning strategies, the Rapid-clicking Strategy involved more tactics, aligning with (Malmberg et al., 2013), who suggested excessive or unfocused tactic use may reflect uncertainty rather than effective learning.

Importantly, this study revealed certain non-cognitive skill constructs, such as meta-cognition and social competencies, are related to learning strategies including risk-oriented tactics. The Rapid-clicking Strategy group showed lower scores in both meta-cognition and social competencies compared to other groups, indicating a notable relationship between learning behaviors and non-cognitive skills.

Overall, these associations provide insights into how behavioral logs reflect non-cognitive skills in K–12 education and lay a foundation for future predictive models and interventions supporting students' learning and non-cognitive development.

4.2 Implication and Limitation

This study identifies risk-oriented tactics like rapid, browse-like clicking and highlights learning strategies that use multiple tactics, sometimes to a high degree. These findings can help educators detect shallow engagement and inform adaptive learning system developers. However, some interpretations—such as the link between browse-like behavior and lower meta- cognition—remain speculative and require cautious interpretation and further validation.

Due to the small sample size and case-study design, the generalizability of results is limited. The specific context and participants may not represent broader populations or different educational settings. As a result, this study does not yet produce actionable prediction models or interventions but lays important and incremental groundwork by identifying relevant behavioral patterns associated with non-cognitive skills.

Additionally, this study relies on Kato's questionnaire (Kato, 2020), which covers some non-cognitive skills but not the full range. Future research should use larger samples, add academic data, include qualitative methods (e.g., interviews), and explore more non-cognition skill constructs across frameworks (e.g., Big Five, CASEL) to validate results.

Acknowledgements

This study was funded by MEXT/JSPS KAKENHI Grant Number JP23H00505.

References

Fan, Y., Saint, J., Singh, S., Jovanovic, J., & Gašević D. (2021). A learning analytic approach to unveiling self-regulatory processes in learning tactics. LAK21: 11th International Learning Analytics and Knowledge Conference, 184–195. https://doi.org/10.1145/3448139.3448211

Higgins, S., Hall, E., Baumfield, V., & Moseley, D. (2005). A meta-analysis of the impact of the implementation of thinking skills approaches on pupils (tech. rep.) (EPrint Processing Status: Full text deposited in DRO). https://durham-repository.worktribe.com/output/1638893

Huber, K., & Bannert, M. (2023). Investigating learning processes through analysis of navigation behavior using log files. Journal of Computing in Higher Education, 36(3), 683–701. https://doi.org/10.1007/s12528-023-09372-3

Kato, S. (2020). A fundamental study on measurement of non-cognitive skills of elementary school students. journal of Teaching Course, (15), 17–28. https://aska-r.repo.nii.ac.jp/records/8236

Malmberg, J., Järvelä, S., & Kirschner, P. A. (2013). Elementary school students' strategic learning: Does task-type matter? Metacognition and Learning, 9(2), 113–136. https: //doi.org/10.1007/s11409-013-9108-5

Malmberg, J., Järvenoja, H., & Järvelä, S. (2010). Tracing elementary school students' study tactic use in gstudy by examining a strategic and self-regulated learning [Advancing Educational Research on Computer-supported Collaborative Learning (CSCL) through the use of gStudy CSCL Tools]. Computers in Human Behavior, 26(5), 1034–1042. https://doi.org/10.1016/j.chb.2010.03.004

Matcha, W., Gašević, D., Uzir, N. A., Jovanović, J., & Pardo, A. (2019). Analytics of learning strategies: Associations with academic performance and feedback. Proceedings of the 9th International Conference on Learning Analytics & Knowledge, 461–470. https: //doi.org/10.1145/3303772.3303787

Ogata, H., Majumdar, R., Yang, S. J. H., & Warriem, J. M. (2022). Learning and evidence analytics framework (leaf): Research and practice in international collaboration. Information and Technology in Education and Learning, 2(1), Inv-p001-Inv–p001. https://doi.org/10.12937/itel.2.1.Inv.p001

Weinstein, C. E., Husman, J., & Dierking, D. R. (2000). Chapter 22 - self-regulation interventions with a focus on learning strategies. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), Handbook of self-regulation (pp. 727–747). Academic Press. https://doi.org/10.1016/B978-012109890-2/50051-2

Zynuddin, S. N., Kenayathulla, H. B., & Sumintono, B. (2023). The relationship between school climate and students' non-cognitive skills: A systematic literature review. Heliyon, 9(4), e14773. https://doi.org/10.1016/j.heliyon.2023.e14773

Learning from Others' Codes in Game-based Robot Programming: Behavioral Patterns and Outcomes

Shintaro MAEDA[a*], Kento KOIKE[b], & Takahito TOMOTO[c]
[a] *Graduate School of Information and Computer Science, Chiba Institute of Technology, Japan*
[b] *Faculty of Engineering, Tokyo University of Science, Japan*
[c] *Faculty of Innovative Information Science, Chiba Institute of Technology, Japan*
*front4.shintaro@gmail.com

Abstract: This study introduces a game-based programming environment for developing algorithmic thinking through iterative refinement. The system features "Score-based Gradual Worked Examples," allowing learners to view others' codes that achieved slightly higher scores. An analysis of behavioral logs from 49 university students using K-means clustering revealed seven distinct patterns. Learners who frequently referenced others' codes while making incremental changes showed greater learning gains. Expert evaluation suggests that presenting codes based on algorithmic similarity may be more effective than score-based selection alone. These findings can inform the design of adaptive support in programming education.

Keywords: Computational thinking, algorithmic thinking, worked examples, trial-and-error

1. Introduction

In programming education, students must develop not only syntax knowledge but also the ability to design and execute problem-solving procedures through computational thinking and algorithmic thinking (AT) (Lockwood et al., 2016; Wing, 2006). Visual programming environments like Code.org have shown promise in promoting AT through iterative trial-and-error processes (Choi, 2022; Papert, 1980). However, traditional puzzle-type tasks with single correct answers may limit continuous improvement once a working solution is found.

To address this limitation, we designed a game-based programming task that encourages score maximization rather than finding a single correct answer (Maeda et al., 2024a). This open-ended approach motivates learners to continuously refine their algorithms for better performance. However, novice learners require support, such as worked examples (WE) to progress effectively. Although WE can reduce cognitive load and facilitate learning (Sweller, 2020; Zhi et al., 2019), presenting complete solutions may lead to mere imitation without meaningful trial-and-error learning.

We propose "Score-based Gradual Worked Examples," where learners can view peer-generated codes that achieved slightly higher scores. This approach addresses two key challenges: (1) satisfaction with suboptimal solutions, by presenting incremental improvements rather than complete answers, and (2) limited support, by scaling support through peer-generated examples rather than instructor-authored ones. While previous classroom implementation showed overall score improvements (Maeda et al., 2024a), individual differences in system usage and learning outcomes remain unexplored. Therefore, this study analyzes behavioral patterns to address the following research questions (RQs). RQ1: In a class using the proposed step-by-step "Score-based Gradual Worked Examples," what trial-and-error characteristics are observed among learners, and how are these differences related to learning effectiveness? RQ2: For learners who do not achieve

satisfactory learning outcomes even after viewing the WE, what factors might have influenced their learning?

2. Literature Review

AT—the ability to construct new algorithms to solve problems (Futschek, 2006)—requires learning environments that support repeated trial-and-error (Grover & Pea, 2013; Papert, 1980). Game-based platforms like Code.org and Scratch effectively promote AT through iterative challenges (Hsu & Wang, 2018), encouraging natural skill development through repeated practice. However, without clear guidance, learners may abandon trial-and-error when unable to understand or modify their programs. WE addresses this by reducing cognitive load through solution demonstrations (Sweller, 2020), which are proven to be effective in programming education (Zhi et al., 2019). Yet conventional WE remain static and uniform, limiting adaptation to individual needs. Learners also tend to skip high-quality examples without meaningful engagement (Zhi et al., 2019). Recent research has explored adaptive WE presentation methods. Toukiloglou and Xinogalos (2023) used fuzzy reasoning to adapt WE to learning progress; however, utilization was inconsistent and examples were fixed to specific problems rather than learner-constructed algorithms. Effective WE must be perceived as "helpful" (Muldner et al., 2023), yet learners often find them unhelpful when the "distance from the task" is too great (Wang et al., 2021). To address these limitations, our system automatically presents peer-generated code examples incrementally, matching examples to learners' current performance and facilitating continuous trial-and-error through contextually relevant support.

3. Method

3.1 Participants and Procedure

This study analyzed log data from a classroom implementation that involved 49 second-year engineering students who used the system during a 40-minute session (10-min pre-test, 20-min system use with code viewing features, 10-min post-test). This approach has shown significant score improvements (Wilcoxon signed-rank test, $p < 0.005$, Cohen's $d > 0.8$) as reported in our previous work (Maeda et al., 2024a). To understand how different usage patterns of the "Score-based Gradual Worked Examples" feature influenced these learning outcomes, we extracted and analyzed behavioral indicators from the system logs.

3.2 System Description

The learning environment requires learners to construct algorithms for a virtual robot that performs seeding and harvesting tasks. To maximize crop yield, learners must design effective action sequences and conditional branching, fostering AT development. The system scores and ranks algorithms, allowing learners to view others' codes that was "one rank higher"—a restriction ensuring examples remain sufficiently aligned with learners' current performance. Figure 1 shows the system features (Learners write code in C#).

Game-based programming task designed to encourage score acquisition: The system is structured as a harvesting game with a virtual robot. Evaluation metrics like productivity, cost, and efficiency are used to assign scores. Productivity is indicated by the number of crops harvested, and cost is indicated by the number of robotic actions. Efficiency is indicated by the value obtained by subtracting cost from productivity (Maeda et al., 2024a). Improving one's score thus becomes equivalent to engaging in trial-and-error to develop more effective algorithms.

Ranking feature: The system displays each learner's score in relation to their peers. This approach not only promotes self-evaluation but also encourages the setting of new goals

and increases motivation by emphasizing relative achievements rather than absolute performance.

Code viewing function: To foster continuous trial-and-error learning, learners are only allowed to view the code of another learner who holds a rank just above theirs. This “one-rank-higher” restriction ensures that the code being referenced is similar in complexity, making it more accessible and relevant.

Since the learners’ code is used, differences in quality and syntax are expected to arise. Thus, beginners who are learning programming are expected to learn from others’ codes who are at a similar level. Additionally, learners who already have some ability to write code are also expected to learn from the codes of others who are at a similar level. It should be noted that this system focuses only on the behavior of robots, or algorithmic evaluation (scores), and therefore does not evaluate syntactic style.

Figure 1 (left) shows the robot planting and harvesting interface, where harvest scores depend on crop growth. Since each command incurs costs, efficient algorithm design is essential. The total score equals harvest score minus operational costs. The ranking panel (right) displays peer scores and allows viewing code one rank above.

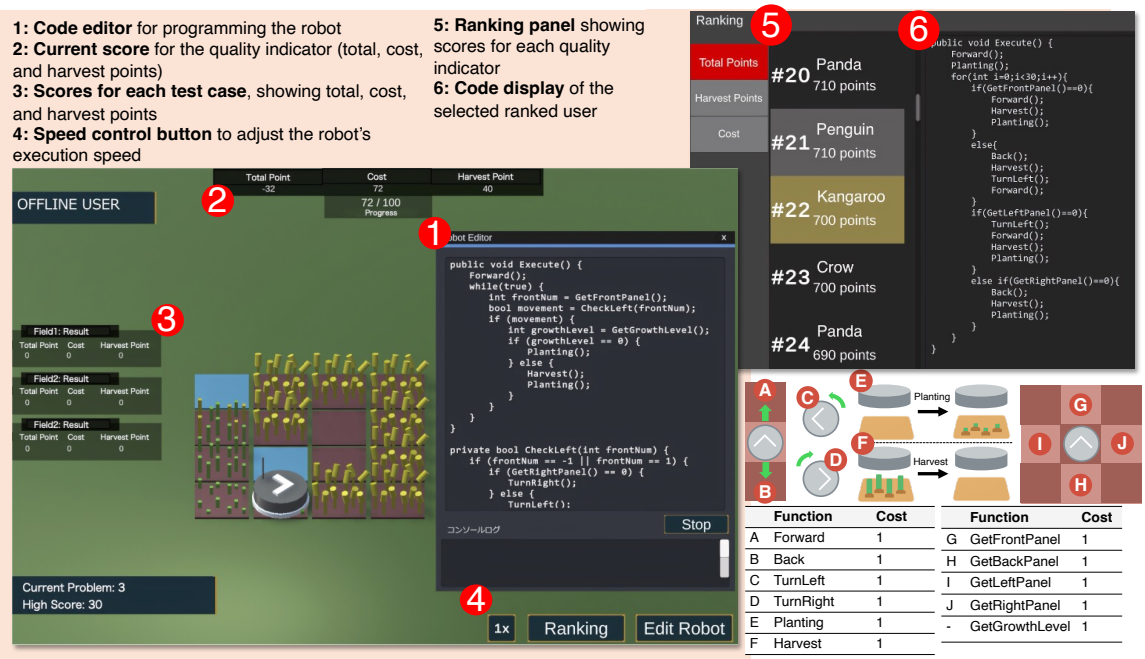


Figure 1. Example of the Code Visualization Interface and Ranking System.

3.3 Overview and Selection of Behavioral Indicators

To understand system usage patterns and their impact on learning, we analyzed behavioral log data using seven key indicators: (1) #Exec: Code execution count after modifications (excluding repeated runs) (2) ΔTotal: Cumulative behavioral changes across executions (3) Δ/Exec: Average behavioral change per execution (ΔTotal/#Exec) (4) Maximum score: Highest total score achieved (5) #Ref: Frequency of viewing others' codes during learning (6) Ref/Exec: Reference rate per execution (#Ref/#Exec) (7) Learning gain: Normalized gain (Hake, 1998): g = (post−pre-score)/(max−pre-score), where max = 3284 points. These metrics collectively assess trial-and-error authenticity and the utilization of others’ code examples from multiple perspectives.

4. Results

Table 1 presents the basic statistics for each indicator across three phases: pre (Pre), during system-use learning (Learn), and post (Post). While learners cannot access WE features in

the Pre and Post phases, these features were enabled in the Learn phase. Additionally, the same questions were used for Pre, Post, and Learn. #Exec increased from 13.2 (Pre) to 24.2 (Learn) and 15.0 (Post), with the Learn phase showing higher activity likely due to having double the exercise time (20 vs. 10 minutes).

Reference frequency (#Ref) varied widely (mean 3.8, SD 3.4), but reference rate per execution (Ref/Exec) remained stable at approximately 20%. While average scores improved significantly from 79.4 to 1,279.1, suggesting learning effectiveness, the relationship between behavioral indicators and achievement remained unclear from basic statistics alone, necessitating further analysis.

Table 1. *Basic Statistics*

	Pre		**Post**		**Learn**		**Pre-Post**	
	Mean (SD)	**Min-Max**	**Mean (SD)**	**Min-Max**	**Mean (SD)**	**Min-Max**	**Mean (SD)**	**Min-Max**
#Exec	13.2 (10.3)	0.0-48.0	24.2 (14.8)	4.0-68.0	15.0 (8.8)	4.0-43.0	-	-
ΔTotal	4.0 (3.2)	0.0-17.4	8.0 (5.3)	0.3-27.4	4.5 (3.1)	0.0-16.7	-	-
Δ/Exec	0.3 (0.2)	0.0-0.6	0.3 (0.1)	0.1-0.6	0.3 (0.2)	0.0-0.7	-	-
Max Score	79.4 (435.5)	−676.0 -1568.0	815.4 (721.7)	−500.0 -2760.0	1279.1 (924.2)	−370.0 -3284.0	--	-
Learning Gain	-	-	-	-	-	-	0.4 (0.3)	-0.1 (-1.0)

Note: Negative scores may occur when harvesting fails or when movement costs exceed the harvest value

K-means clustering was performed using five normalized indicators (#Exec, ΔTotal, Δ/Exec, Score, #Ref), yielding seven clusters based on silhouette score optimization. Table 2 presents the behavioral characteristics of each cluster.

Table 2. *Behavioral Indicators in Each Cluster*

Cl.	**Name**	**N**	**#Exec**	**ΔTotal**	**Δ/Exec**	**Max Score**	**#Ref**
0	HTR	2	52.5(0.7)	12.5(0.5)	0.2(0.0)	669.0(420.0)	10.0(1.4)
1	MAT	17	17.3(5.2)	6.4(1.9)	0.4(0.1)	566.8(366.3)	2.4(1.9)
2	PAS	6	9.7(5.1)	2.7(1.4)	0.2(0.1)	29.3(319.4)	0.8(1.0)
3	SIR	9	22.0(5.2)	10.3(2.8)	0.5(0.1)	1154.9(504.2)	7.3(2.8)
4	EFS	4	13.3(6.5)	2.8(2.1)	0.2(0.1)	2194.0(517.1)	6.0(2.9)
5	PTM	7	44.3(12.3)	7.6(2.5)	0.2(0.1)	1004.6(913.6)	2.0(2.0)
6	BRV	4	42.3(11.1)	20.9(5.3)	0.5(0.1)	650.5(418.2)	4.3(4.8)

Three high-performing clusters emerged: Self-improver & Reference (SIR) showed high Ref/Exec and Δ/Exec, making incremental code modifications while frequently referencing others' codes. Efficient-Strategist (EFS) achieved the highest scores despite low #Exec, suggesting effective use of others' codes within limited trials. Heavy-trial & Reference (HTR) had high #Exec and #Ref but low reference density, limiting its effectiveness.

Four clusters showed limited success: Passive (PAS) exhibited minimal engagement across all metrics. Persistent-trial Mixed (PTM) and Moderate-activity (MAT) both showed high trial activity but low reference usage, with MAT (n=17) displaying conscious algorithmic modifications but variable outcomes. Bulk-revision (BRV) made large code changes (high ΔTotal and Δ/Exec) but with moderate reference usage, potentially losing focus through excessive modifications.

A one-way ANOVA revealed significant differences in learning outcomes between clusters ($F(6,42)=4.781$, $p=0.0008$, $\eta^2=0.4058$). Post-hoc analysis (Tukey's HSD) showed EFS significantly outperformed PAS ($p=0.0005$), while MAT outperformed EFS ($p=0.0016$). Despite lower #Exec and ΔTotal, EFS achieved the highest scores through high Ref/Exec—frequently referencing others' codes within limited executions while making incremental modifications. These findings suggest that reference density (Ref/Exec) rather than trial quantity (#Exec) is the key factor for improving learning outcomes, though causal relationships require further investigation given the small cluster sizes.

5. Discussion and Future work

Cluster analysis revealed that learning effectiveness depended more on the quality of the trial-and-error than quantity. EFS and SIR clusters demonstrated that high reference density (Ref/Exec) combined with incremental modifications led to superior outcomes, despite fewer overall trials. Conversely, HTR showed that frequent references without meaningful integration yielded limited gains, while PAS's minimal engagement and MAT/BRV's low reference usage resulted in poor performance.

These findings address both RQs. For RQ1, reference density emerged as more critical than trial frequency for effective learning. For RQ2, HTR's case demonstrated that reference volume alone is insufficient—the quality of engagement matters. This aligns with Wang et al.'s (2021) identified barriers: comprehension difficulties and modification challenges, which intensify when the presented code significantly differs from learners' own solutions. These barriers likely influenced the observed behavioral patterns and outcomes, suggesting that code examples should be closely aligned with learners' current approaches to facilitate effective adoption.

To reduce learning barriers, we propose presenting codes based on two similarity types: (A) syntactic format and (B) algorithmic proximity—aligning with code clone classifications where Types I-III represent textual similarity and Type IV represents functional similarity (Roy & Cordy, 2007). Our findings suggest that algorithmic similarity facilitates better understanding and integration into learners' trial-and-error processes.

In a preliminary study (Maeda et al., 2024b), we quantified algorithmic similarity using F-values based on execution behaviors and presented both algorithm-based and conventional score-based rankings to three expert programmers (a professor specializing in programming education and two industry professionals). Expert evaluation strongly favored algorithm-based ranking (mean: 5.67/6) over conventional ranking (mean: 3.67/6). Experts noted that algorithmically similar codes were "helpful for development" and "used the same strategy," while conventional rankings showed "differences too large for beginners" risking "code copying rather than learning." Thus, algorithmic similarity-based presentation could effectively replace the current "one-rank-higher" restriction.

Recent advances in deep learning, such as SANN for generating code vectors (Hoq et al., 2025), enable presentation of syntactically similar programs that match learners' coding levels rather than expert syntactic patterns. Combining both similarity dimensions could reduce code quality as a confounding factor and enable clearer behavioral pattern interpretation. This understanding would support personalized strategies, such as presenting examples aligned with learners' behavioral tendencies or offering alternative approaches when learners stagnate—ultimately enabling more adaptive, individualized learning support.

This observational study has some limitations. First, it cannot establish causal relationships between code viewing and learning improvements. While supplementary analysis (Maeda et al., 2025) showed score improvements when learners incorporated elements from others' codes, it remains unclear which code types are most "adoptable" and "effective." Therefore, the hypothesis that algorithmically and syntactically similar codes facilitate learning requires further empirical validation. Further, some clusters, such as HTR and BRV, remain an issue for the future due to their small sample sizes. Second, this study evaluated a single robot programming task, limiting generalizability to other programming domains or diverse learner populations, which warrants future investigation. The algorithmic similarity approach requires validation through actual learner testing beyond expert evaluation.

This study has implications for future research. First, it is not possible to evaluate how learners read and understand the code of others. In contrast, in evaluations based on previous research, a survey was conducted on learners who viewed the codes of others to investigate how their own code changed, and it was confirmed that a certain number of learners incorporated parts of others codes into their own, in a way that was not imitation (Maeda et al., 2025). Furthermore, this led to an improvement in scores. Another avenue for future

research is investigating the learning effects of sharing others' codes with large score differences.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP25K21362.

References

Choi, W. C. (2022). The influence of code.org on computational thinking and learning attitude in block-based programming education. *Proceedings of the 2022 6th International Conference on Education and E-Learning*, 235–241. https://doi.org/10.1145/3578837.3578871

Futschek, G. (2006). Algorithmic thinking: The key for understanding computer science. In R. T. Mittermeir (Ed), *Informatics Education – The Bridge between Using and Understanding Computers (Vol.* 4226, pp. 159–168*). Springer Berlin Heidelberg*. https://doi.org/10.1007/11915355_15

Grover, S., & Pea, R. (2013). Computational thinking in K–12: A review of the state of the field. *Educational Researcher*, *42*(1), 38–43. https://doi.org/10.3102/0013189X12463051

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, *66*(1), 64–74. https://doi.org/10.1119/1.18809

Hoq, M., Patil, A., Akhuseyinoglu, K., Brusilovsky, P., & Akram, B. (2025). An automated approach to recommending relevant worked examples for programming problems. *Proceedings of the 56th ACM Technical Symposium on Computer Science Education*, *1*, 527–533. https://doi.org/10.1145/3641554.3701951

Hsu, C.C., & Wang, T.I. (2018). Applying game mechanics and student-generated questions to an online puzzle-based game learning system to promote algorithmic thinking skills. *Computers & Education*, *121*, 73–88. https://doi.org/10.1016/j.compedu.2018.02.002

Lockwood, E., & Asay, A. (2016). Algorithmic thinking: An initial characterization of computational thinking in mathematics. *North American Chapter of the International Group for the Psychology of Mathematics Education*.

Maeda, S., Koike, K., & Tomoto, T. (2024a). Code visualization system for writing better code through trial and error in programming learning: Classroom implementation and practice. *International Conference on Computers* in *Education*. https://doi.org/10.58459/icce.2024.5006

Maeda, S., Koike, K., & Tomoto, T. (2024b). Code-sharing platform in programming learning: A proposal for a strategy-aware code-sharing methodology. *International Conference on Human-Computer Interaction*, 338–348.

Maeda S., Koike K., & Tomoto T. (2025). Proposal and classroom practice of game task type virtual robot programming to promote writing “better code” through trial and error [Manuscript submitted for publication].

Muldner, K., Jennings, J., & Chiarelli, V. (2023). A review of worked examples in programming activities. *ACM Transactions on Computing Education*, *23*(1), 1–35. https://doi.org/10.1145/3560266

Papert, S. (1980). *Mindstorms: Children, computers, and powerful ideas*. Basic Books, Inc.

Roy, C. K., & Cordy, J. R. (2007). A survey on software clone detection research. *Queen’s School of Computing Technical Report* 2007-541. Queen’s University, Canada.

Sweller, J. (2020). Cognitive load theory and educational technology. *Educational Technology Research and Development*, *68*(1), 1–16. https://doi.org/10.1007/s11423-019-09701-3

Toukiloglou, P., & Xinogalos, S. (2023). Adaptive support with working examples in serious games about programming. *Journal of Educational Computing Research*, *61*(4), 766–789. https://doi.org/10.1177/07356331231151393

Wang, W., Kwatra, A., Skripchuk, J., Gomes, N., Milliken, A., Martens, C., Barnes, T., & Price, T. (2021). Novices’ learning barriers when using code examples in open-ended programming. *Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education*, *Vol.* 1, 394–400. https://doi.org/10.1145/3430665.3456370

Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, *49*(3), 33–35. https://doi.org/10.1145/1118178.1118215

Zhi, R., Price, T. W., Marwan, S., Milliken, A., Barnes, T., & Chi, M. (2019). Exploring the impact of worked examples in a novice programming environment. *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, 98–104. https://doi.org/10.1145/3287324.3287385

Teacher-Involved Automatic Characteristics Classification in Handwritten Math Answer Process

Shunsuke TONOSAKI[a*], Taito KANO[a], Chia-Yu HSU[b] & Hiroaki OGATA[b]
[a]*Graduate School of Informatics, Kyoto University, Japan*
[b]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
*tonosaki.shunsuke.75i@st.kyoto-u.ac.jp

Abstract: In Japanese junior high schools, while digital logs of students' handwritten math answers offer utilization opportunities for classroom sharing, manually reviewing them is burdensome for teachers, and existing automatic classification methods often fail to meet their pedagogical needs. This study, co-designed with teachers, defined four pedagogically classification labels, identified their associated handwriting features, and subsequently evaluated classification models. Among the models tested, XGBoost was most effective, notably meeting our success criterion (precision > 0.5) for teacher support on the diagram problem type (0.532), thus demonstrating the practical feasibility of this approach. Our findings highlight the importance of a teacher-involved approach to designing learning analytics systems for practical classroom use.

Keywords: learning analytics, learning logs, handwriting process

1. Introduction

In Japanese elementary and junior high schools, students are provided with a digital device such as PC, tablets and pen devices for educational purposes through the GIGA School Program, and these devices are now used instead of paper for math problem-solving (MEXT, 2020). This enables the accumulation of students' handwriting processes as learning logs (Ogata et al., 2018), and visualization through the conversion and replay of these logs as videos has been achieved (Yoshitake et al., 2020). Teachers can use the replay function to observe students' thought processes regarding their difficulties and diverse ways of thinking, share these with the entire class, and create opportunities for mutual learning.

However, checking each student's log requires significant time and effort, placing a heavy burden on teachers. To reduce the burden on teachers and enable them to quickly access answers that match their purpose for sharing answers, it is important to provide support for summarizing and classifying the characteristics of each answer.

While various studies have classified handwritten answers, their criteria are often defined by researchers. These have included rule-based models focused on student stumbles (Asai et al., 2012), machine-learning for clustering (Yoshitake et al., 2020) or predicting performance (Stahovich & Lin, 2016), and LLMs for content recognition (Caraeni et al., 2025). However, these approaches did not consider labels required by teachers, even though incorporating user perspectives is crucial in Learning Analytics (Mavrikis et al., 2019).

This study proposed a teacher-involved co-design process for defining classification labels and criteria for handwritten math answers. To verify the feasibility of this co-design approach, several basic classification models were developed and evaluated. To guide this process, we address the following research questions:
RQ1: Which categories of handwriting process characteristics can be derived from teachers' observations?
RQ2: Which features of the handwriting process affect the teacher-focused characteristics?
RQ3: Is it feasible to classify handwritten answers based on teacher-defined labels for real-world practice?

2. Context

This study analyzed the handwriting problem-solving processes of 226 first- and second-year junior high school students in Japan using the Learning and Evidence Analytics Framework (LEAF) system (Ogata et al., 2018). The problems were answered using a pen device on the eBook tool (Book Roll) within the LEAF system, and two teachers—one proficient in using its analysis tool (Log Palette) for student assessment, and the other an experienced math teacher skilled in interpreting answer characteristics—reviewed replay videos of each student's answer using Log Palette (Yoshitake et al., 2020). The dataset comprised 292 description-type (law of exponents) and 222 diagram-type (inscribed angle theorem) answers, derived from three problems of each type presented one problem per page (Figure 1, left). We used two problem types because prior research suggests that both student handwriting behavior and the characteristics teachers focus on can differ by problem type (Tonosaki et al., 2024).

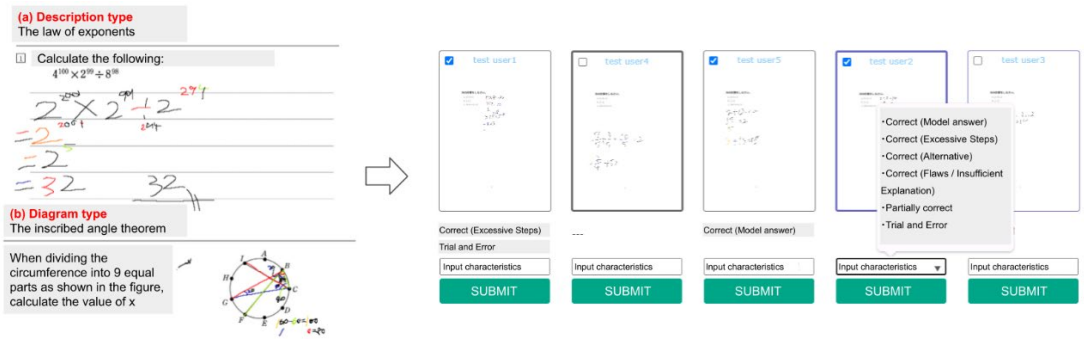


Figure 1. Examples and the Labeling Function of Handwriting Answering Process

To collect data for analysis, we first asked two teachers to label answers suitable for sharing in class, using a Log Palette function that allowed predefined or custom labels to each answer (Figure1, right). For the subsequent automatic classification, we used two data types from the logs: the performance score (*PFM*) and 13 handwriting process features (Table 1), all of which were standardized per problem type.

Table 1. *Handwriting Process Features*

Features	Description
TAT	Total answering time (second)
TNS	Total number of strokes
TNE	Total number of erasers
Speed	Whole answering speed (*TNS* / *TAT*)
AD	Average of duration
TSD	Total number of short durations (< *AD*) from each previous stroke
TLD	Total number of long durations (≥ *AD*) from each previous stroke
VD	Variance of duration
AST	Average of stroke time (second)
TSL	Total number of stroke lengths
ASL	Average of stroke lengths
ASLS	Average of each stroke speed (stroke length / stroke time)
VSLS	Variance of each stroke speed

3. Analysis & Results

3.1 Categorized Labels (RQ1)

We unified and categorized teacher labels through semi-structured interviews and the Card Sorting method (Upchurch et al., 2001). In this process, teachers first discussed the meaning

and criteria for each label and then grouped them based on their practical classroom use. We refined these provisional categories, and the final labels and their definitions were confirmed by the teachers via email, at which point data saturation was reached.

The initial 12 labels provided by teachers, reflecting criteria such as performance, content, and stroke movement, were grouped into the four final classification labels. These labels, whose definitions are shown in Table 2, had the following distribution: "*Standard Model Answer*" (n=33 Description, n=14 Diagram), "*Creative Alternative Answer*" (n=15, n=12), "*Trial-and-Error or Mistake Answer*" (n=6, n=68), and "*Unclear Process Answer*" (n=1, n=37).

Table 2. *Categorized Labels and Their Definitions by Teachers (translated)*

Categorized Labels	Definition
Standard Model Answer	Uses appropriate mathematical notation and terminology, proceeds smoothly with sufficient detail, and reaches an expected answer.
Creative Alternative Answer	An answer reached through an original process, different from the model answer.
Trial-and-Error or Mistake Answer	Regardless of correctness, the answer shows visible trial and error or mistakes recognizable by others.
Unclear Process Answer	Regardless of correctness, the answer includes unclear steps that require additional explanation from the student.

3.2 Feature Selection & Analysis (RQ2)

To identify characteristic features, we examined the relationship between the features in Table 1 and categorized labels using two approaches: a teacher-driven questionnaire and a data-driven analysis. First, in the questionnaire, two teachers rated the relationship between each feature and label as 'Positive,' 'Negative,' or 'No Relationship.' Second, for the data-driven analysis, we used SHAP to quantify the explanatory power of each feature independently of the model (Lundberg & Lee, 2017). We trained an XGBoost model and considered the top five features ranked by their SHAP values as highly explanatory for each label.

Comparing the questionnaire and data analysis (Table 3), several features were commonly identified as significant, such as *PFM* for "*Standard Model Answer*" (description-type), *TAT* for the same label (diagram-type). In contrast, feature selection for "*Unclear Process Answer*" was challenging, as teachers reported 'No Relationship' for the diagram-type, and insufficient data prevented data-driven analysis for the description-type.

Table 3. *Results of each Feature Selection Method (Positive: +, Negative: -)*

Labels	Problem type	Questionnaire	Data Analysis (SHAP)
Standard Model Answer	*Description*	*PFM*+, *TAT*+, *TNS*+, *Speed*+	*TNE*+, *ASLS*-, *PFM*+, *VSLS*-, *Speed*-
	Diagram	*TAT*+, *TNS*-, *TNE*-, *Speed*-, *ASLS*+, *AD*-, *TSD*-	*PFM*+, *TAT*+, *TSD*+, *AST*+, *VSLS*+
Creative Alternative Answer	*Description*	*TAT*-, *TSL*+	*ASLS*-, *PFM*+, *AD*-, *TAT*+, *Speed*+
	Diagram	*TAT*+, *TNS*+, *TNE*+, *Speed*+, *ASLS*+, *AD*+, *TSD*-	*TNS*+, *VD*-, *AST*+, *PFM*+, *VSLS*+
Trial-and-Error or Mistake Answer	*Description*	*TAT*+, *TNE*+, *AD*+	*Speed*+, *AST*+, *AD*-, *PFM*-, *ASLS*-
	Diagram	*TNE*+, *Speed*+, *ASLS*-, *VSLS*+, *TSL*+, *ASL*+, *AST*-	*Speed*-, *VD*+, *TLD*+, *TSD*+, *TNS*-
Unclear Process Answer	*Description*	*TAT*-, *TNS*-, *Speed*-, *AD*-	(Nothing)
	Diagram	(Nothing)	*VSLS*+, *TNS*-, *ASL*+, *TSD*+, *VD*-

3.3 Multi-Label Classification (RQ3)

To verify the feasibility of automatically classifying answers using the teacher-defined labels, we evaluated three multi-label classification models, each offering different practical benefits: a transparent teacher-driven (rule-based) model, a high-performance data-driven (machine-learning) model, and a content-aware LLM-driven model. For evaluation, we prioritized macro-average precision to ensure the reliability of labels suggested to teachers. To define the effectiveness criterion for RQ3, we set a success threshold of a precision of at least 0.5. The threshold was chosen because it implies that if a teacher reviews two answers flagged by the models with a certain label at least one will be correctly classified, which we considered a sufficient level for supporting teachers. We also calculated Recall, F1 score, and Accuracy as the overall evaluations of the models. All models used the 13 handwriting features (Table 1) and performance score (*PFM*) as input.

First, for the teacher-driven classification, we implemented a rule-based model. We tested four feature selection strategies based on two sources—teacher questionnaires and data analysis (SHAP): (1) using only features from the questionnaire, (2) using only features from data analysis, (3) using features common to both (AND), and (4) using features from either source (OR). The classification rule was based on whether a feature's value exceeded its median, with the threshold adjusted to maximize precision.

Second, we implement machine-learning using handwriting process features for data-driven classification. The XGBoost, Support Vector Machine (SVM), and Random Forest were used in this study. For the three models, we conducted classification tasks using the features selected by SHAP, followed by parameter tuning using Grid Search. The models' performances were evaluated using 5-fold cross-validation, and the differences between the machine-learning models were tested for statistical significance using paired t-tests.

Third, for the LLM-driven classification, we used OpenAI's GPT-4o for multi-label image recognition. The prompt provided the model with the problem context, the student's answer image, label definitions, the desired output format, and few-shot examples, and instructed it as follows: "Look at the answer below and, based on the definitions, assign all applicable labels (all independent) accurately and return them in an array." We then tested three prompt variations: one with no features (None), one with all 14 features (All), and one with a pre-selected subset of features (Selected). For the 'Selected' condition, we identified the top five features that most improved precision when added to the prompt individually in a preliminary test. All features provided to the LLM were expressed in natural language on a five-point scale (e.g., "Answering Time: Top").

As shown in Table 4, the machine-learning models significantly outperformed the other approaches. XGBoost achieved the highest precision for the description-type (0.496) and the diagram-type (0.532), with the latter meeting our success criterion. Paired t-tests confirmed that XGBoost's precision was statistically significantly higher than the other models for both types ($p < .05$). In contrast, both the rule-based and LLM-based models resulted in low precision.

Table 4. *Evaluations of Each Multi-Label Classification Model*

	Method	Precision	Recall	F1 Score	Accuracy
(a) Rule-Based (comparison by feature selections)					
Description	Only Questionnaire	0.106	0.438	0.161	0.716
Diagram	OR	0.200	0.551	0.268	0.650
(b) Machine-Learning (comparison by models)					
Description	XGBoost	0.496	0.501	0.496	0.880
Diagram	XGBoost	0.532	0.525	0.515	0.747
(c) LLM (comparison by feature selections)					
Description	Selected	0.128	0.348	0.182	0.107
Diagram	None	0.149	0.283	0.187	0.107

4. Discussion

4.1 Co-Designing Pedagogical Labels (RQ1)

Through our labeling process with teachers, we identified four types of labels intended for classroom use. Based on the interview responses and our interpretation, each label has a distinct usage scenario. For example, "*Standard Model Answer*" can be used as exemplary and "*Creative Alternative Answer*" to spark student interest. "*Trial-and-Error or Mistake Answer*" is seen as opportunities to share common mistakes, while "*Unclear Process Answer"* serves as a reminder for students to show their work.

Unlike labels from prior work aimed at detecting when students stumble (Asai et al., 2012), our labels were defined for the pedagogical purpose of sharing answers in class. This highlights the need to design classification labels flexibly according to their educational use.

4.2 Analysis of Teacher-Focused Features (RQ2)

Our analysis revealed a difference in focus between the teacher questionnaires and the data analysis (Table 3). The teachers tended to select visually distinguishable, time-level features like total answering time (*TAT*), while the data-driven approach identified fine-grained, stroke-level features (e.g., *ASLS*, *AD*). These features reflected different student states depending on the problem type. For description-type problems, stroke timing features were prominent, suggesting that a student's calm or rushed state was a key indicator. For instance, a "*Standard Model Answer*" was associated with calm, consistent work. For diagram-type problems, features related to process accumulation like the number of strokes (*TNS*) were more significant, reflecting statuses such as having a clear goal despite many steps ("*Creative Alternative Answer*") or making repeated corrections ("*Trial-and-Error or Mistake Answer*").

Notably, performance (*PFM*) was rarely a top feature, distinguishing our study from those focused on performance prediction (Caraeni et al., 2025; Stahovich & Lin, 2016) and highlighting the value of behavioral data in capturing teacher-defined characteristics.

4.3 Effectiveness of Classification for Teacher Support (RQ3)

Our results show that the XGBoost machine-learning model performed significantly better than the other models. Notably, it achieved effective classification for the diagram-type (0.532) by meeting our success criterion (precision > 0.5), though it fell just short for the description-type (0.496). This demonstrates the practical feasibility of building a teacher-support tool, at least for diagram-type problems. Such a tool can effectively surface relevant student answers for review, thereby reducing teachers' search effort. In contrast, the rule-based and LLM-based approaches showed limited effectiveness.

The differing results among the models can be attributed to the unique complexity of the handwriting process. Machine-learning models likely succeeded by identifying complex patterns within this behavioral data. In contrast, the rule-based model's low precision suggests that simple rules are insufficient. This difficulty implies that the process's complexity may be a factor in the burden teachers face when identifying answer characteristics. Similarly, the LLM likely struggled as it analyzed only the final static image, lacking access to the crucial time-series data of the writing process.

4.4 Limitation & Future Work

This study has several limitations that suggest directions for future work, including: (1) Data Scale and Imbalance: The data was labeled by only two teachers, and the inter-rater reliability was not calculated, resulting in a small, imbalanced dataset that could lead to overfitting. Future work should involve more teachers and problems to create a more robust data set. (2) Feature Selection: As our use of SHAP may not be optimal for all models (e.g., SVM). Employing model-specific feature selection techniques could further improve performance. (3)

LLM Application: The LLM approach showed low precision and high computational costs. Performance could be enhanced through prompt optimization and fine-tuning, while costs could be reduced by using open-source models. (4) Hybrid Model Development: A key future direction is to combine the teacher-driven, data-driven, and LLM approaches. Integrating rule-based pedagogical intent with machine-learning's behavioral pattern recognition and the LLM's content understanding could create a more accurate and educationally relevant hybrid system. (5) Practical System Implementation: To ensure any resulting tool effectively reduces teacher burden, it is crucial to visualize the classification results and continue the co-design process by including teacher feedback from the prototype stage (Mavrikis et al., 2019).

5. Conclusion

This study, co-designed with teachers, established four pedagogical labels for classifying handwritten math answers intended for classroom sharing. We identified the process features influencing these labels and found that among the models tested, the XGBoost machine-learning model was the most effective at automatic classification. While not for full automation, these models offer practical support by helping teachers efficiently select relevant answers for classroom discussion. This study's primary contribution is its teacher-involved methodology, demonstrating the importance of co-designing learning analytics systems to reflect classroom needs, especially when dealing with abstract data like handwriting processes.

Acknowledgements

This work was partly supported by the Council for Science, 3rd SIP JPJ012347, and JSPS KAKENHI, Grant Number 23H00505.

References

Asai, H., Nozawa, A., Sonoda, S., & Hayato, Y. (2012). *Onrain tegaki dēta o mochiita gakushū-sha no tsumazuki kenshutsu [Student's stumble detection using online handwriting data]*. DEIM Forum A8-4. http://db-event.jpn.org/deim2012/proceedings/final-pdf/a8-4.pdf

Caraeni, A., Scarlatos, A., & Lan, A. (2025). Evaluating GPT-4 at grading handwritten solutions in math exams. *15th International Conference on Learning Analytics & Knowledge (LAK25)*, 126–128. https://www.solaresearch.org/wp-content/uploads/2025/02/LAK25_CompanionProceedings-Final.pdf

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/1705.07874

Mavrikis, M., Geraniou, E., Gutierrez Santos, S., & Poulovassilis, A. (2019). Intelligent analysis and data visualisation for teacher assistance tools: The case of exploratory learning. *British Journal of Educational Technology: Journal of the Council for Educational Technology*, *50*(6), 2920–2942. https://doi.org/10.1111/bjet.12876

MEXT. (2020). *The image of the transformation of learning brought by "1 device for 1 student with a high-speed network."* https://www.mext.go.jp/en/content/20200716-mxt_kokusai-000005414_04.pdf

Ogata, H., Majumdar, R., Akçapınar, G., Hasnine, M., & Flanagan, B. (2018). *Beyond learning analytics: Framework for technology-enhanced evidence-based Education and learning*. http://hdl.handle.net/2433/237322

Stahovich, T. F., & Lin, H. (2016). Enabling data mining of handwritten coursework. *Computers & Graphics*, *57*, 31–45. https://doi.org/10.1016/j.cag.2016.01.002

Tonosaki, S., Kano, T., Hamada, S., Horikoshi, I., & Ogata, H. (2024). Toward contextualized handwriting process analysis: Comparison between problem types in math. *The 32nd International Conference on Computers in Education*. https://doi.org/10.58459/icce.2024.5063

Upchurch, L., Rugg, G., & Kitchenham, B. (2001). Using card sorts to elicit Web page quality attributes. *IEEE Software*, *18*(4), 84–89. https://doi.org/10.1109/MS.2001.936222

Yoshitake, D., Flanagan, B., & Ogata, H. (2020). Supporting group learning using pen stroke data analytics. *28th International Conference on Computers in Education Conference Proceedings*, *1*, 634–639. https://repository.kulib.kyoto-u.ac.jp/dspace/handle/2433/259798

Recomposition Based Learning for Promoting Structural Understanding
- From Reconstruction of External Representations to Recomposition of Internal Representation-

Tsukasa Hirashima[a*], Kodai WATANABE[a]
[a]*Hiroshima University*
*tsukasa@lel.hiroshima-u.ac.jp

Abstract: In educational settings, structured external representations, such as diagrams, examples, or concept maps, are often assumed to ensure learners' understanding. However, recognition of a representation does not guarantee that its underlying meaning has been internalized. Understanding involves not only the elemental meaning conveyed by individual components but also the structural meaning that emerges from their organization; the additional meaning constructed through this organization can be defined as constructed meaning ("constructed meaning = structural meaning – elemental meaning"). To foster this constructed meaning, we propose Recomposition Based Learning, a framework in which learners reconstruct a target structure from predefined components. This process externalizes, compares, and refines learners' internal representations, thereby promoting structural understanding and metacognitive reflection. By focusing on the manipulation of external representations to refine internal ones, the framework reduces cognitive load while preserving the need for interpretive reasoning. It is formalized as a four-stage cycle: (1) hypothetical recomposition, (2) difference detection, (3) conceptual clarification and completion, and (4) reflective recomposition. This paper outlines the theoretical basis of the framework.
Keywords: Recomposition Based Learning, Structural Understanding, External and Internal Representation, Kit-Build Concept Map

1. Introduction

In educational settings, learning content is often presented through structured external representations, visible and manipulable formats such as diagrams, tables, or concept maps. While these representations may appear to convey knowledge clearly and objectively, a learner's understanding depends not only on what is presented but also on how it is internally interpreted. External representations, provided by teachers, must be mentally reconstructed into internal representations, namely, the learner's own conceptual framework of meaning.

Understanding involves both the elemental meaning carried by individual components and the structural meaning arising from their organization. The difference between these, constructed meaning, is the additional meaning generated through organizing components into a coherent structure ("constructed meaning = structural meaning – elemental meaning"). A mismatch in constructed meaning often arises between what is shown and what is actually understood.

Despite this, instruction frequently proceeds under the implicit assumption that recognition of a representation implies understanding. Learners are often not required to reconstruct or articulate the underlying meaning themselves, and practical constraints such as limited class time, curriculum coverage, and standardized assessments reinforce a reliance on recognition rather than reconstruction. As a result, surface-level recognition is often mistaken for comprehension.

To address this gap between visible structure and internal comprehension, we propose Recomposition Based Learning, a structured learning framework in which learners reconstruct target structures from predefined components. This process supports learners in externalizing,

comparing, and refining their internal representations, thereby fostering constructed meaning. In contrast to open-ended scratch-building, which demands high generative effort and yields highly variable outputs, recomposition reduces cognitive load while still engaging learners in interpretive reasoning. This approach narrows the scope of possible arrangements of the given components, while ensuring that learners must still engage in semantic interpretation.

This study (1) frames the refinement of internal representations through the manipulation of external ones as a core principle for promoting structural understanding, (2) proposes Recomposition Based Learning as a pedagogically grounded framework for fostering constructed meaning, (3) offers design implications for learning environments that align external representations with internal understanding. In this paper, Recomposition Based Learning is introduced as an implementation using concept maps, called the Kit-Build Concept Map (KB map for short) (Hirashima et al. 2015, Hirashima 2024).

2. Learning of Internal Representation through the Manipulation of External Representations

This section outlines the theoretical foundations of our study, focusing on the distinction and interaction between external and internal representations, and on how the manipulation of external representations can refine internal ones.

2.1 External and Internal Representations

A crucial distinction in learning theory lies between external representations—such as diagrams, tables, or concept maps—and internal representations, which refer to the learner's mental models or conceptual structures (Norman, 1993; Zhang, 1997). External representations make knowledge structures visible and manipulable, thereby supporting cognitive processes. However, comprehension ultimately depends on how learners interpret and organize the presented information internally.

Understanding consists of two components: (1) Elemental meaning: the meaning inherent in individual components of a representation. (2) Structural meaning: the meaning arising from the organization and interrelation of these components. The constructed meaning is the additional meaning that emerges from integrating elemental meanings into a coherent structure. Achieving alignment between external and internal representations means ensuring that learners not only recognize the elements but also construct the intended meaning from their relationships. Ainsworth (2006) emphasizes that such alignment is both essential and inherently difficult to achieve because learners bring different prior knowledge and interpretive frameworks. Even when an external representation appears unambiguous, the constructed meaning derived by different learners can vary widely.

Scaife and Rogers (1996) further articulate how graphical representations function within external cognition, highlighting the role of design in facilitating the transition from perception to conceptual understanding. Therefore, effective learning design should not merely present well-structured external content but also engage learners in actively manipulating and reconstructing these structures. This process externalizes internal interpretations, making them visible for comparison, reflection, and refinement. Through such interaction, learners progressively sophisticate their internal representations, strengthening both their structural understanding and metacognitive awareness.

2.2 Constructive Constructivism: A Theoretical Possibility

While constructivist learning theories emphasize the active role of learners in meaning-making, our framework highlights a specific dimension of this process: the generation of constructed meaning as the difference between elemental and structural meanings. We term this perspective Constructive Constructivism. From this view, learning is understood not merely as acquiring elemental or structural meaning, but as the iterative reconstruction of their differences. By explicitly treating these differences as objects of learning, Constructive Constructivism extends conventional constructivist perspectives, offering a more precise theoretical grounding for how external representations can be employed to refine internal ones.

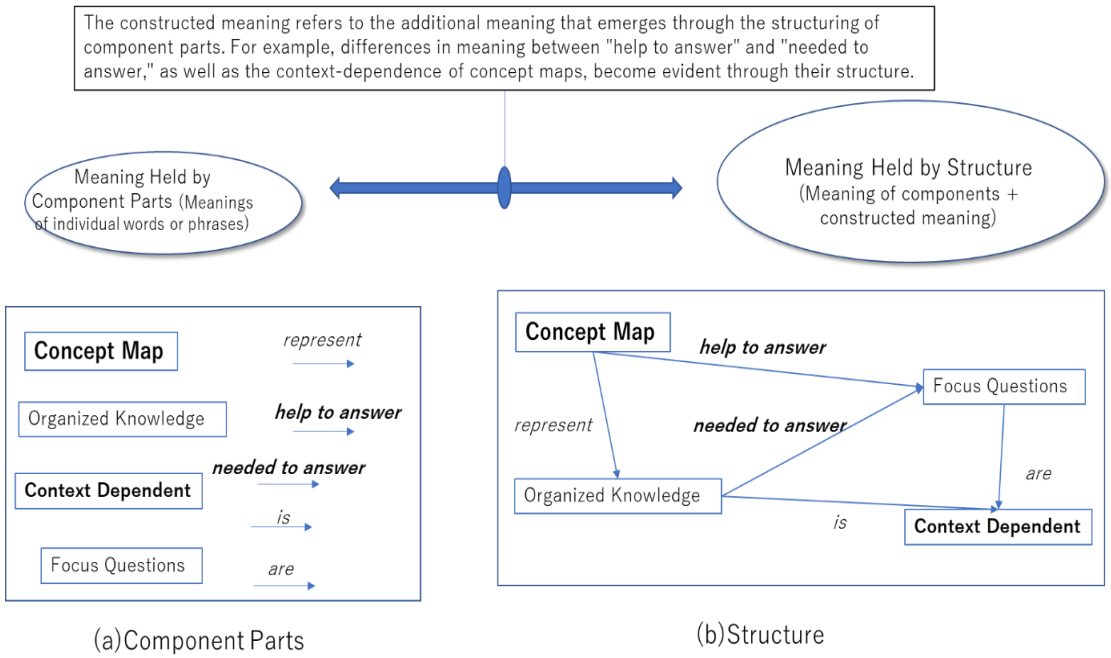


Figure 1. Relationship between External and Internal Representations in Conceptual Understanding: A Case of a Concept Map about Concept Maps

3. Illustrative Case: Recomposition of a Part of Concept of Concept Map

As defined by Novak and Cañas (2008), a concept map is a graphical tool for organizing and representing knowledge through nodes and labeled links, explicitly encoding the relationships between concepts. This characteristic makes it a suitable medium for examining both elemental and constructed meanings in our framework. Figure 1 shows the relationship between meanings held by individual components and those held by the overall structure. In panel (a), “Component Parts” represent the elemental meaning of individual words or phrases, such as Concept Map, Organized Knowledge, Context Dependent, and Focus Questions. These elements, along with linking phrases like represent, help to answer, needed to answer, is, and are, each carry meaning on their own.

However, as shown in panel (b), when these components are connected into a coherent structure, they convey structural meaning, that is, integrated meaning derived from their relationships. The difference between the structural meaning and the sum of the elemental meanings is the constructed meaning: the additional meaning that emerges only through the organization of components.

For example, in the structure, the distinction between help to answer and needed to answer becomes evident, as does the way Context Dependent relates to Focus Questions through are. These nuances are not apparent when looking at individual components in isolation but emerge through their arrangement. This example demonstrates that learning from structured representations requires learners to construct meaning beyond simply recognizing components. In Recomposition Based Learning, learners reconstruct such structures from provided components, making their internal understanding of both elemental and constructed meaning visible for reflection and refinement.

4. Recomposition Based Learning Framework

4.1 Overview

This framework is grounded in the premise that understanding emerges through a dynamic interplay between internal and external representations, and that the refinement of internal representations can be effectively supported through the manipulation of external ones.

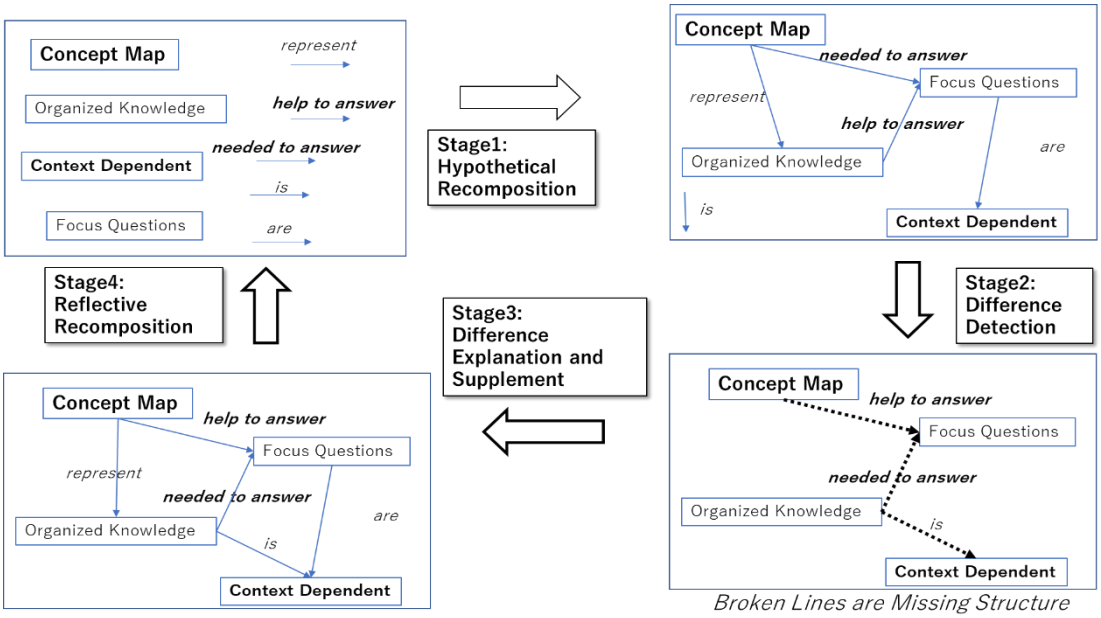


Figure 2. Four Stages of Recomposition Based Learning

In this framework, learners reconstruct structured knowledge using predefined components, thereby externalizing their internal understanding. Unlike open-ended construction tasks, which often impose high cognitive demands by requiring learners to generate components independently, recomposition enables learners to focus on interpreting and organizing meaning, rather than generating content from scratch. This design reduces cognitive load and facilitates clearer comparisons across learners, making external representations more diagnosable and sharable. These features support both metacognitive engagement and collaborative learning.

Recomposition tasks explicitly engage learners' internal representations by requiring them to assemble a coherent structure from given components. Through the processes of comparison, explanation, and revision, learners are encouraged to reflect not only on the structural configurations but also on their associated meanings. Even when a reconstruction, based merely on recall, appears structurally accurate, the act of recomposition can reveal conceptual uncertainty, prompting metacognitive awareness. This contrast between surface-level correctness and semantic ambiguity provides a powerful opportunity for engaging learners in conceptual inquiry and reflective understanding.

4.2 Four Stages of Recomposition Based Learning

4.2.1 Stage1: Hypothetical Recomposition

In the first stage, learners are provided with a set of components—such as concept nodes and relational links—and are asked to reconstruct a target structure. This reconstruction is guided by the learner's current internal representations and may result in a structure that aligns or misaligns with the expected model. The act of recomposing is inherently hypothesis-driven, as learners must interpret and infer the intended relationships. This stage makes visible the learner's initial conceptual model.

As shown in Figure 2, a set of nodes and links are provided as components (top left). In Stage 1, the learner connected four of the links, while one remains unconnected(top right). The resulting concept map (learner map) offers valuable diagnostic information about the learner's understanding.

4.2.2 Stage 2: Difference Detection

After the initial recomposition, learners compare their structure with a reference model or with peer reconstructions. This comparison enables them to identify differences not only in configuration but also in interpretive assumptions. These differences highlight discrepancies

in meaning, not merely form, and draw attention to mismatches between external correctness and internal understanding.

Importantly, difference detection is not limited to identifying overt mismatches. Even when learners appear to reproduce a structure successfully based on surface-level memory, the act of comparing and justifying their composition can evoke subtle discomfort or uncertainty. Learners may become aware that, despite their structure appearing correct, they lack confidence in the underlying rationale. This realization initiates metacognitive reflection and can stimulate self-generated questioning, transforming apparent success into a springboard for deeper learning.

In the lower-right part of Figure 2, the missing structure is represented where three missing links (dashed lines) are identified. This difference map (difference learner map) is generated by overlapping the learner map and the original map.

4.2.3 Stage 3: Conceptual Clarification and Completion

In this stage, learners engage in conceptual clarification and map completion by explaining why such differences exist, by reasoning through their own decisions or interpreting the intended meaning of the target structure and supplement the differences to complete the map. This phase supports metacognitive articulation and helps learners identify the assumptions underlying their interpretations. It encourages conceptual elaboration and the refinement of internal models through clarification. The map in the lower-left part of Figure 2 shows a revised map in which the learner has explained and supplemented the differences to complete the structure.

4.2.4 Stage 4: Reflective Recomposition

Based on the insights gained, learners proceed to revise their structure. This revision is not simply a correction but a reconstruction informed by deeper understanding. It enables learners to realign their internal representations with the meaningful structure underlying the original model. This final stage represents a transformation of understanding through iterative clarification and metacognitive awareness.

4.3 Related Work on Kit-Build Concept Map and Recomposition

Previous studies have reported that the KB map system effectively supports learning in classroom settings. For instance, Pailai et al. (2017) demonstrated that using the KB map system in elementary school science classes improved learning outcomes through formative assessment. The system supports individual and collaborative learning as well as the development of higher-order thinking skills. Wunnasri et al. (2018), Sadita et al. (2020), and Pinandito et al. (2021) showed that collaborative KB map construction enhanced reading comprehension and facilitated meaningful dialogue in a second-language learning context. Compared to traditional concept mapping, learners were more inclined to ask questions about each other's maps because the system made comparison easy and structured. Nurmaya et al. (2023) and Nurmaya et al. (2025) further reported that learners using the KB map system achieved higher scores on higher-order thinking questions than those using conventional scratch-built maps. The KB map system has been successfully applied across various educational contexts, including elementary science classes, high school science instruction, university-level reading comprehension, and undergraduate information science courses. These applications demonstrate the system's flexibility and effectiveness across different subjects and learner levels. However, the fourth stage, Reflective Recomposition, has not yet been implemented as a system function and remains an important subject for future work.

5. Conclusion

In this paper, we proposed Recomposition Based Learning, a four-stage framework that externalizes and refines learners' internal representations by manipulating provided external representations. The framework is grounded in constructivist principles: learners actively build meaning by interpreting and manipulating external structures. When multiple learners work with the same set of components, compare their reconstructions, and engage in mutual

explanation, the process reflects social constructivist learning. Such sharing of structures promotes meaning negotiation and facilitates learning within the Zone of Proximal Development (ZPD). Although the final stage, Reflective Recomposition, is not yet system-guided, many learners voluntarily engaged in this phase, suggesting its perceived value and indicating promising directions for future system design. Recomposition Based Learning thus offers a practical and theoretically grounded approach to supporting both individual understanding and collaborative knowledge construction through technology-enhanced learning.

Moreover, by encouraging learners to question apparent correctness, detect inconsistencies, and reconstruct meanings beyond surface recognition, the framework also provides a foundation for higher-order educational goals such as critical and creative thinking. Recent studies support this direction; for instance, Hasani et al. (in press) demonstrated that Kit-Build concept mapping as a preparatory activity enhanced Community of Inquiry (CoI)-based asynchronous online discussions by improving learners' conceptual knowledge. Such evidence indicates that Recomposition Based Learning can serve not only as a method for refining internal representations, but also as a stepping stone toward collaborative inquiry and the cultivation of critical and creative capacities. Future work will address the implementation and empirical validation of the Reflective Recomposition stage.

References

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction, 16(3), 183–198.

Hasani, L. M., Junus, K., Sadita, L., Ohsaki, A., Hirashima, T., & Hayashi, Y. (in press). Improving Community of Inquiry-based asynchronous online discussion through improving conceptual knowledge by information-organizing preparatory activity with Kit-Build concept mapping. Online Learning.

Hirashima, T., Yamasaki, K., Fukuda, Y., & Funaoi, T. (2015). Kit-Build concept map for automatic diagnosis of conceptual understanding. Research and Practice in Technology Enhanced Learning, 10*(1), 1–23.

Hirashima, T. (2024). Formative assessment and meaningful learning with concept mapping through recomposition. Information and Technology in Education and Learning, 4(1), pp.1-14.

Norman, D. A. (1993). Things That Make Us Smart: Defending Human Attributes in the Age of the Machine. Addison-Wesley.

Novak, J. D., & Cañas, A. J. (2008). The theory underlying concept maps and how to construct them. *Technical Report IHMC CmapTools 2006-01 Rev 01-2008*. Institute for Human and Machine Cognition.

Nurmaya, N., Pinandito, A., Hayashi, Y., & Hirashima, T. (2023). Promoting students' higher order thinking with concept map recomposition. *IEICE Transactions on Information and Systems, E106-D*(8), 1262–1271.

Nurmaya, N., Pinandito, A., Hayashi, Y., & Hirashima, T. (2025). Promoting Higher-Order Thinking through Online Collaborative Concept Map Recomposition. Research and Practice in Technology Enhanced Learning(accepted).

Pailai, J., Wunnasri, W., Yoshida, K., Hayashi, Y., & Hirashima, T. (2017). The practical use of Kit-Build concept map on formative assessment. Research and Practice in Technology Enhanced Learning, 12, 1-23.

Pinandito, A., Hayashi, Y., & Hirashima, T. (2021). Online collaborative Kit-Build concept map: Learning effect and conversation analysis in collaborative learning of English as a foreign language reading comprehension. *IEICE Transactions on Information and Systems, E104-D*(7), 981–990.

Sadita, L., Hirashima, T., Hayashi, Y., Wunnasri, W., Pailai, J., Junus, K., & Santoso, H. B. (2020). Collaborative concept mapping with reciprocal kit-build: a practical use in linear algebra course. Research and Practice in Technology Enhanced Learning, 15, 1-22.

Scaife, M., & Rogers, Y. (1996). External cognition: How do graphical representations work? International Journal of Human-Computer Studies, 45*(2), 185–213.

Wunnasri, Warunya, Jaruwat Pailai, Yusuke Hayashi, and Tsukasa Hirashima. "Reciprocal Kit-Build concept map: an approach for encouraging pair discussion to share each other's understanding." IEICE Transactions on Information and Systems 101, no. 9 (2018): 2356-2367.

Zhang, J. (1997). The nature of external representations in problem solving. Cognitive Science, 21(2), 179–217.

Behavior-Aware Argumentation: Integrating Toulmin's Model for Adaptive Learning in Digital Environments

Banni Satria ANDOKO[a*], Vivin Ayu LESTARI[a], Usman NURHASAN[a], Rosa Andrie ASMARA[a], Galih Putra RIATMA[b], Ananda Galih PRATIWI[a], Tsukasa HIRASHIMA[c]
[a]*Information Technology Dept, State Polytechnic of Malang, Indonesia*
[b]*Electrical Engineering Dept, State Polytechnic of Malang, Indonesia*
[c] *Graduate School of Advanced Science and Engineering, Hiroshima University, Japan*
*ando@polinema.ac.id

Abstract: This study investigates an adaptive learning mechanism within Viat-map, a learning tool based on Toulmin's Argument Model, to address engagement challenges among struggling and gaming students. In addition to scoring correctness, the system monitors two key behavioral indicators: the number of steps taken and the time spent on tasks. These indicators are used to tailor the number of multiple-choice options, reducing them for overloaded learners and increasing them for those who may be guessing strategically. In a controlled experiment with sophomore EFL students, pre- and post-tests, ANCOVA, and effect size analyses showed significant gains for struggling learners, while high performers benefited less. Findings highlight the value of behavior-aware analytics in optimizing educational technologies, and offer insight into refining personalized learning algorithms.

Keywords: Adaptive Learning, Toulmin Argument, Viat-map, EFL, Constructivism

1. Introduction

Understanding how students interact with digital learning environments is essential to creating effective interventions. Viat-map implements Toulmin's Argument Model, a framework built on claim, ground, and warrant, to help learners develop logical reasoning and critical thinking (Andoko et al., 2023; Andoko, Mubarok, et al., 2022). The Toulmin model, which emphasizes the construction of arguments through Claims, Grounds, and Warrants, offers a structured approach to developing logical thinking skills, a core component of critical reasoning (Al-Ajm & Ambusaidi, 2022; Majidi et al., 2021). Toulmin's structure is especially relevant here because it provides clear scaffolding for reasoning. This makes it easier for learners to focus their working memory on processing ideas rather than holding unstated connections (Le Cunff et al., 2024; Tzafilkou et al., 2021).

Previous evaluations of Viat-map have focused mostly on whether students reached correct answers. This approach risks missing the underlying learning process, particularly for two types of learners (Andoko et al., 2024).:

- Struggling learners, who spend a long time working but make little progress, often because of cognitive overload
- Gaming learners, who exploit system rules to score highly without fully understanding the material

Additionally, phenomena such as "gaming the system"—in which students attempt to exploit platform mechanics to achieve high scores without genuine understanding—highlight the limitations of performance-only assessment methods and underscore the need for behavior-aware analytics (R. Baker et al., 2007; R. S. J. D. Baker et al., 2006).

To address this, Viat-map now incorporates two indicators of engagement: the time spent and the sequence of steps taken. Together, these indicators capture both pacing and exploration patterns. They are widely supported in learning analytics research as useful measures of cognitive engagement and strategic behavior.

A prior clustering study identified both gaming and struggling learner profiles. These informed new adaptive rules: choice complexity is decreased for overloaded learners and increased for rapid guessers. This study tests the effect of those rules and examines how prior knowledge, material type, and learner profile interact with the adaptive mechanisms. The research questions for this study are:

1. How does the adaptive learning intervention impact student learning outcomes after controlling prior knowledge, and does its effect vary across different types of instructional materials?
2. To what extent does prior knowledge influence the effectiveness of adaptive learning, and how do learning gains differ between low-performing and high-performing students?

2. Literature Review

2.1 Cognitive Load

Cognitive Load Theory distinguishes between intrinsic load, which relates to the complexity of the task and the learner's prior knowledge; extraneous load, which comes from irrelevant demands; and germane load, which is the mental effort devoted to learning (Leppink et al., 2013). In EFL reading and argumentation, structured scaffolds such as Toulmin's clear claim-evidence-warrant layout can reduce extraneous load and free mental resources for deeper processing. Research shows that such structures help mitigate working-memory strain, which improves retention and transfer (Wang et al., 2020; Le Cunff et al., 2024).

2.2 Gaming The System

Gaming occurs when learners bypass the intended cognitive process to achieve success, for example by cycling through options or taking advantage of predictable patterns in multiple-choice questions. This behavior can hide the true level of understanding and distort assessment data. Adaptive strategies such as varying complexity or giving targeted feedback have been shown to redirect engagement back to meaningful learning (Petre et al., 2019; R. S. J. D. Baker, 2006).

2.3 Toulmin Argument

Toulmin's model breaks argumentation into components that can be taught explicitly. For adaptive systems, this is valuable for two reasons: it has proven benefits for reasoning skills (Majidi et al., 2021), and its component-based design fits naturally with step-by-step interaction data (Rismanto et al., 2021), making it possible to analyse behaviour in detail (Magalhães, n.d.).

2.4 Viat Map

Viat-map is designed to structure learning using Toulmin's basic argument model, where In Viat-map, claims are set by the teacher and learners choose the grounds and warrants. This keeps the activity structured while giving students a chance to evaluate the quality and relevance of different arguments. The system records each selection step and the time taken, which makes it possible to distinguish deliberate reasoning from random trial-and-error. This information feeds directly into the adaptive mechanisms.

3. Method

3.1 Adaptive Learning Function in Viat-map

Viat-map adjusted the number of answer options in real time, responding to how students interacted with each question. When a student took more than seven seconds to choose, the system treated this as a sign of difficulty and reduced the options from three to two, easing cognitive load. If a student changed their answer more than twice within ten seconds, the system assumed they might be guessing and increased the options to four, prompting more careful consideration. Most students remained at the default of three options unless their behavior matched one of these patterns. In the control group, the number of options never changed.

3.2 Experimental Setting

The study involved 41 second-year students from two intact Information Technology classes at the State Polytechnic of Malang. The classes followed their regular timetable, with one class assigned to the control condition and the other to the adaptive condition. Two reading passages from the standard English syllabus were used: Computer and iPod Nano. Each was paired with a ten-item multiple-choice pre-test and a post-test, and the answer keys were only revealed once the study ended. The experiment was designed following the flow shown in Figure 1

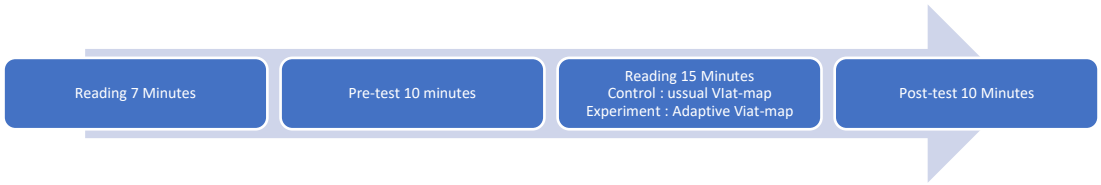


Figure 1. Experimental setting flow

4. Result

Post-test scores were compared between groups using ANCOVA, with pre-test scores as the covariate to control for prior knowledge. Hedges's g was calculated to gauge the practical significance of the results. The first analysis should focus on establishing baseline comparability between the experimental and control groups before measuring the impact of the intervention. Descriptive Statistics for Pre-Test Scores using Welch two-sample t-test was conducted for both materials and here is the result:

To examine differences in prior knowledge before the intervention, we analyzed pre-test scores using Welch's two-sample t-test. As shown in Table 1, the control group had a slightly higher average score (5.375) than the experimental group (4.235), but the difference was not statistically significant ($p = 0.0769$). A second analysis for the second material (Table 2) showed similar results, with the control group scoring 5.416 and the experimental group 4.764 ($p = 0.2408$). In both cases, the p-values exceeded 0.05 and the confidence intervals included zero, indicating no significant baseline differences between the groups.

Table 1. Summary of Pre-Test Score Comparison Between Control and Experimental Groups for iPod Nano Material

Statistic	Control Group	Experimental Group	Test Value
Mean Pre-Test Score	5.375	4.235	—
t-test Value (t)	—	—	1.8172
Degrees of Freedom (df)	—	—	38.859
p-value	—	—	0.0769
95% Confidence Interval	—	—	-0.129 to 2.408

Table 2. Summary of Pre-Test Score Comparison Between Control and Experimental Groups for Computer Material

Statistic	Control Group	Experimental Group	Test Value
Mean Pre-Test Score	5.416667	4.764706	—
t-test Value (t)	—	—	1.1919
Degrees of Freedom (df)	—	—	37.619
p-value	—	—	0.2408
95% Confidence Interval	—	—	-0.455 to 1.760

After ensuring that the baseline of each group are comparable, the analysis now is divided into two sections to show a detailed analysis of each material within groups.

4.1 ANCOVA Analysis for Computer Material

Students' prior knowledge clearly shaped their post-test outcomes, as shown in Table 3. Still, the adaptive learning intervention made a meaningful difference, boosting performance with a statistically significant result ($p = 0.0229$). The model's low residual variance adds confidence that these findings genuinely reflect the intervention's impact.

Table 3. ANCOVA Results for Adaptive Learning Intervention

Factor	Degreesof Freedom (Df)	Sum of Squares (SS)	Mean Square (MS)	F-Value	p-Value
Pre-Test	1	54.35	54.35	64.486	1.04e-09*
Group (Intervention Impact)	1	4.74	4.74	5.627	0.0229*
Residuals	38	32.03	0.84	—	—

*Significance codes: *** (p < 0.001), ** (p < 0.01), * (p < 0.05)*

The adaptive learning program clearly improved student performance, with a strong effect size shown in Table 4. The negative indicator suggests a shift in how students learned. The confidence interval confirms that the improvement was meaningful and not just by chance..

Table 4. Effect Size Calculation (Hedges's g)

Effect Size Measure	Estimate	95% Confidence Interval
Hedges's g	-0.836 (large)	Lower: -1.491, Upper: -0.181

4.2 ANCOVA Analysis for iPod Nano Material

Table 5 shows that pre-test scores had a strong influence on post-test performance, underlining the key role of prior knowledge in shaping learning outcomes. Still, the adaptive learning intervention led to a clear improvement (p = 0.00673), confirming its effectiveness. The moderate residual variance suggests the model was well-fitted and dependable, meaning the results reliably reflect the intervention's impact.

Table 5. ANCOVA Results for Adaptive Learning Intervention

Factor	Degrees of Freedom (Df)	Sum of Squares (SS)	Mean Square (MS)	F-Value	p-Value
Pre-Test	1	134.04	134.04	104.090	1.95e-12*
Group (Intervention Impact)	1	10.58	10.58	8.218	0.00673
Residuals	38	48.94	1.29	—	—

*Significance codes: *** (p < 0.001), ** (p < 0.01), * (p < 0.05)*

Table 6 confirms that the adaptive learning function had a strong impact on student performance, supported by a large effect size. The negative value reflects a shift in performance patterns, indicating potential underlying changes in how students engaged with and benefited from the intervention. The confidence interval reinforces the reliability of this effect, showing that the program produced meaningful improvements in outcomes. Complementing this, Table 7 presents ANCOVA and effect size results for low- and high-performing students across materials, revealing that the strongest and most significant gains occurred among lower performers, particularly for certain content areas. This convergence of findings underscores both the overall effectiveness of the adaptive function and its differential benefits across learner profiles.

Table 6. Effect Size Calculation (Hedges's g)

Effect Size Measure	Estimate	95% Confidence Interval
Hedges's g	-1.001 (large)	Lower: -1.667, Upper: -0.335

Table 7. ANCOVA and Effect Size Results for Low vs. High Performers Across Materials

Material	Performance Group	p-value (ANCOVA, Group Effect)	Hedges's g	95% CI for g
Computer	Low performers	0.0239 *	–1.415 (large)	–2.644 to –0.186

Computer	High performers	0.4003	–0.611 (medium)	–1.411 to 0.188
iPod Nano	Low performers	0.01010 *	–1.276 (large)	–2.291 to –0.260
iPod Nano	High performers	0.3420	–0.600 (medium)	–1.508 to 0.307

5. Discussion and Conclusion

Research Question 1 examined how adaptive learning influenced student outcomes and whether its effect varied with different materials. After accounting for prior knowledge, post-test scores rose significantly ($p = 0.0229$; $p = 0.00673$) with large effect sizes (–0.836 to –1.001). Prior knowledge strongly predicted performance ($p = 0.00281$; $p = 0.000203$). High achievers relied more on existing knowledge, while lower-performing students benefited more from the program. The effect also varied by topic, with both Computer and iPod Nano materials improving results but to different extents, underscoring the value of matching adaptive strategies to the subject.

Research Question 2 showed that low-performing students gained the most from the adaptive learning intervention, with significant results ($p = 0.0239$; $p = 0.01010$) and large effect sizes ($g = –1.415$) compared to smaller, non-significant effects for high performers ($p > 0.34$; $g = –0.600$; CI included zero), whose outcomes were largely driven by prior knowledge. Effectiveness also varied by material, suggesting that subject-specific tailoring can further optimize impact. These findings align with prior work linking domain-specific prior knowledge (Simonsmeier et al., 2022) and cognitive load considerations (Dong et al., 2020) to adaptive learning benefits, and with evidence of domain-dependent effects (Mirari, 2022). They highlight the importance of refining adaptive algorithms to meet diverse learner needs and expanding research to address long-term retention and engagement (Martin et al., 2020).

While these findings support adaptive learning, especially for lower-performing students, they come with limitations. The study involved a small sample from one institution, limiting generalizability. Learning gains were measured only immediately after the intervention, so long-term effects are unknown. The adaptation relied on a narrow set of behavioral indicators and fixed thresholds, which may not fully reflect individual differences. Future studies should use larger, more diverse samples, track long-term outcomes, and draw on richer learning analytics for more flexible, learner-specific adaptations. Including motivation and engagement measures could also reveal how adaptive systems shape the learning experience beyond grades.

6. Acknowledgement

We would like to thank the State Polytechnic of Malang for funding this research under Research grant number: SP DIPA-139.03.2.693474/2025.

References

Al-Ajm, B., & Ambusaidi, A. (2022). The Level of Scientific Argumentation Skills in Chemistry Subject among Grade 11th Students: The Role of Logical Thinking. *Science Education International*, *33*(1), 66–74. https://doi.org/10.33828/sei.v33.i1.7

Andoko, B. S., Asmara, R. A., Lestari, V. A., Ikawati, D. S. E., Prasetyo, A., Hirashima, T., & Hayashi, Y. (2023). Experimental Comparison of Promotion Effect for EFL Reading Comprehension between Conventional Summarization and Toulmin Argument Reconstruction. *Proceedings of the 31st International Conference on Computers in Education, Volume 1*, 732–741.

Andoko, B. S., Lestari, V. A., Pramudhita, N., Nuraini, A., Nisak, K., & Hirashima, T. (2024). Analyzing Student Behavior in Viat-map: Steps and Time as Performance Indicators. In Kashihara (Ed.), *Proceedings of the 32nd International Conference on Computers in Education.* Asia-Pacific Society for Computers in Education.

Andoko, B. S., Mubarok, F. U., Hirashima, T., Arhandi, P. P., Astiningrum, M., & Najib, M. F. (2022). Constructing Toulmin's Logical Structure Through Viat-map Application for

Reading Comprehension of EFL Students. *2022 Fifth International Conference on Vocational Education and Electrical Engineering (ICVEE)*, 196–200. https://doi.org/10.1109/ICVEE57061.2022.9930471

Baker, R., Mercedes, M., Rodrigo, T., Baker, R. S. J. D., Lagud, M. C. V, Lim, S. A. L., Macapanpan, A. F., Pascua, S. A. M. S., Santillano, J. Q., Sevilla, L. R. S., Sugay, J. O., Tep, S., & Viehland, N. J. B. (2007). *Affect and Usage Choices in Simulation Problem-Solving Environments. Affect and Usage Choices in Simulation Problem-Solving Environments*. https://www.researchgate.net/publication/221297694

Baker, R. S. J. D. (2006). *Detecting When Students Game the System 1 Developing a Generalizable Detector of When Students Game the System*.

Baker, R. S. J. D., Corbett, A. T., Koedinger, K. R., & Roll, I. (2006). Generalizing detection of gaming the system across a tutoring curriculum. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4053 LNCS*, 402–411. https://doi.org/10.1007/11774303_40

Dong, A., Jong, M. S.-Y., & King, R. B. (2020). How Does Prior Knowledge Influence Learning Engagement? The Mediating Roles of Cognitive Load and Help-Seeking. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.591203

Le Cunff, A. L., Giampietro, V., & Dommett, E. (2024). Neurodiversity and cognitive load in online learning: A systematic review with narrative synthesis. *Educational Research Review*, *43*. https://doi.org/10.1016/j.edurev.2024.100604

Leppink, J., Paas, F., Van der Vleuten, C. P. M., Van Gog, T., & Van Merriënboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods*, *45*(4), 1058–1072. https://doi.org/10.3758/s13428-013-0334-1

Magalhães, A. L. (n.d.). Teaching How to Develop an Argument Using the Toulmin Model. *International Journal of Multidisciplinary and Current Educational Research*. www.ijmcer.com

Majidi, A. el, Janssen, D., & de Graaff, R. (2021). The effects of in-class debates on argumentation skills in second language education. *System*, *101*. https://doi.org/10.1016/j.system.2021.102576

Martin, F., Chen, Y., Moore, R. L., & Westine, C. D. (2020). Systematic review of adaptive learning research designs, context, strategies, and technologies from 2009 to 2018. *Educational Technology Research and Development*, *68*(4), 1903–1929. https://doi.org/10.1007/s11423-020-09793-2

Mirari, K. (2022). The Effectiveness of Adaptive Learning Systems in Personalized Education. *Journal of Education Review Provision*, *2*(3), 70–76. https://doi.org/10.55885/jerp.v2i3.194

Petre, C., Duffy, B. E., & Hund, E. (2019). "Gaming the System": Platform Paternalism and the Politics of Algorithmic Visibility. *Social Media + Society*, *5*(4). https://doi.org/10.1177/2056305119879995

Rismanto, R., Nurul Asri, A., Satria Andoko, B., Hirashima, T., & Arifiandi Leonanta, A. (2021). A Preliminary Study: Toulmin Arguments in English Reading Comprehension for English as Foreign Language Students. *Proceedings - IEIT 2021: 1st International Conference on Electrical and Information Technology*, 6–10. https://doi.org/10.1109/IEIT53149.2021.9587398

Simonsmeier, B. A., Flaig, M., Deiglmayr, A., Schalk, L., & Schneider, M. (2022). Domain-specific prior knowledge and learning: A meta-analysis. *Educational Psychologist*, *57*(1), 31–54. https://doi.org/10.1080/00461520.2021.1939700

Tzafilkou, K., Perifanou, M., & Economides, A. A. (2021). Negative emotions, cognitive load, acceptance, and self-perceived learning outcome in emergency remote education during COVID-19. *Education and Information Technologies*, *26*(6), 7497–7521. https://doi.org/10.1007/s10639-021-10604-1

Wang, C., Fang, T., & Gu, Y. (2020). Learning performance and behavioral patterns of online collaborative learning: Impact of cognitive load and affordances of different multimedia. *Computers and Education*, *143*. https://doi.org/10.1016/j.compedu.2019.103683

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Dual Analysis of Literature and Existing Tools for Establishment of Acceptable xAPI Profile for Collaborative Learning

Mitsushiro EZOE[ab*], Yasuhisa TAMURA[c] & Ren KOMATSUZAKI[c]
[a]*The University of Electro-Communications, Japan*
[b]*codeTakt, Inc., Japan*
[c]*Sophia University, Japan*
*mitsushiro.ezoe@codetakt.com

Abstract: The growing use of digital tools in collaborative learning necessitates standardized methods for analyzing learner interactions. While xAPI offers a flexible syntax for logging learning activities, a common vocabulary for collaborative learning is lacking. This hinders cross-platform data integration and analysis. This study investigates the feasibility of applying existing xAPI vocabularies to describe and analyze collaborative learning logs. We conducted a dual analysis: a literature review of 72 Computer-Supported Collaborative Learning (CSCL) studies to identify theoretically important interactions, and an empirical analysis of interaction logs from widely used educational tools in Japanese K-12 settings to identify common real-world actions. Our findings reveal a significant gap between the higher-order interactions studied in research (e.g., assessing, tracking) and the more fundamental actions logged by tools (e.g., editing, submitting). This suggests that while existing vocabularies can cover a majority of current tool-based logs, they fall short of capturing the full complexity of collaborative processes. This study clarifies the current state and challenges of standardizing collaborative learning data, proposing that establishing a baseline for fundamental action logging is a crucial first step toward more advanced, cross-platform learning analytics.

Keywords: Collaborative Learning, Learning Analytics, xAPI, Data Standardization

1. Introduction

Learning Analytics (LA) aims to improve education by analyzing data from various learning environments. As LA expands from individual learning to complex, multi-tool environments like Computer-Supported Collaborative Learning (CSCL), a fundamental challenge arises: data interoperability. Diverse platforms log interactions in proprietary formats, making it difficult to integrate data, reproduce analyses, and scale findings. This fragmentation means that a student's collaborative process, which might span a discussion forum, a shared document, and a presentation tool, is broken into isolated data silos. Consequently, educators and researchers cannot obtain a comprehensive understanding of group dynamics or individual contributions, significantly hindering evidence-based support.

The Experience API (xAPI) is a technical standard designed to address this by providing a common format for learning data. However, xAPI only defines the syntax. For meaningful cross-system analysis, a shared vocabulary and structure—defined in an xAPI Profile—are essential. An xAPI Profile specifies semantic rules for a particular context by defining Concepts (verbs, activity types), Statement Templates (structural patterns for statements), and Patterns (sequences of related events). By enforcing consistency, profiles ensure that records from different tools can be meaningfully compared, enabling the development of reusable dashboards and analytical tools. While profiles exist for domains like self-regulated learning, a widely accepted profile for collaborative learning is still missing.

Collaborative learning involves complex, distributed, and multi-modal interactions such as proposing ideas, reaching consensus, and coordinating actions. These interactions are often spread across multiple digital tools (e.g., LMS, shared whiteboards, messaging apps), each with its own data schema. This fragmentation makes it nearly impossible to gain a holistic view of collaborative processes. A standardized approach to describing these interactions is a critical prerequisite for advancing collaborative learning analytics, especially in contexts like Japan's GIGA School Program, which has rapidly deployed devices nationwide and amplified the need for interoperable data.

However, a significant gap often exists between the complex interactions that learning theories emphasize and the simple, fundamental actions that educational tools actually log. Before a comprehensive new profile can be designed, it is crucial to understand the nature and extent of this gap. This study, therefore, does not aim to propose a new, definitive profile. Instead, it takes a necessary prior step: to assess the feasibility of using existing, widely recognized xAPI vocabularies to describe collaborative learning. We explore the alignment and misalignment between theoretically significant interactions and empirically observed behaviors in real-world educational settings. This leads to our research questions:

- RQ1: What types of learner interactions are frequently studied in CSCL literature?
- RQ2: What types of learner interactions are most frequently logged by common collaborative learning tools in practice?
- RQ3: To what extent can existing xAPI vocabularies bridge the gap between theoretical models and practical tool logs?

2. Methods

This study employs a mixed-method approach, structured in three phases. The scope is focused on the Japanese K-12 educational context, which has seen rapid, nationwide adoption of digital learning tools.

2.1 Theoretical Analysis (RQ1)

We conducted a literature review to identify interaction types central to collaborative learning research. We searched Google Scholar for empirical CSCL studies, resulting in a final set of 72 peer-reviewed articles. For each article, we identified the dataset used and categorized the described learner interactions using verbs from the existing TLA (Teacher-Led Learning Analytics) Profile. This profile was chosen as a reference because it is a well-established, research-oriented vocabulary designed to capture pedagogical interactions. Its focus on describing learning activities from an analytical perspective provided a suitable lens for classifying the objectives of prior empirical research. The process involved mapping the descriptions of datasets and analytical methods in each paper to the most appropriate verbs within the TLA Profile, thereby creating a quantitative overview of research trends in the CSCL field. This process, facilitated by GPT-4o for initial screening and data extraction, allowed us to quantify which collaborative behaviors are most frequently targeted in academic research. A complete list of references cited in this study is provided in the Appendix.

2.2 Empirical Analysis (RQ2)

To understand real-world learner behavior, we analyzed anonymized interaction logs from schoolTakt, a major classroom support system used in Japanese elementary and junior high schools, alongside a functional comparison with two other popular tools, LoiLoNote School and Miraiseed. These tools were selected due to their high market share and representative functionalities in the Japanese K-12 context. The GIGA School Program has led to the widespread adoption of these specific systems, making their logs a valuable source of authentic, large-scale data on student interactions. Our analytical framework (Figure 1) was designed to abstract the specific UI elements of each tool into generalizable action types, enabling a more fundamental comparison of their core functionalities and the types of data

they generate. As illustrated in Figure 1, we classified user actions based on a three-level framework: (a) actions on the content of the learning item, (b) actions on the learning item itself, and (c) actions directed toward others' items. This framework allowed us to identify the most common, fundamental user operations. Figure 2 provides a concrete example of a typical activity flow within schoolTakt, illustrating the sequence of these fundamental actions from the start of a lesson to its completion.

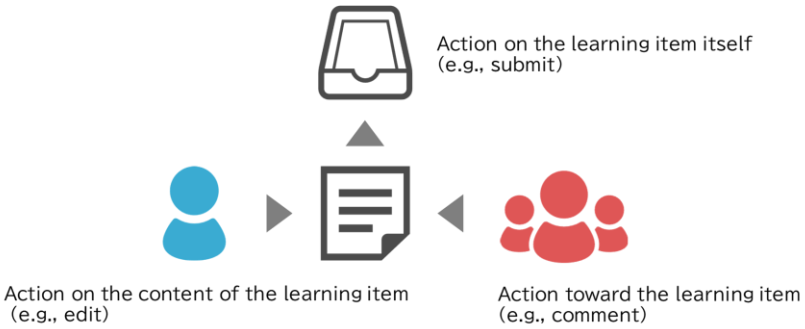


Figure 1. Classification of Actions Related to Learning Items

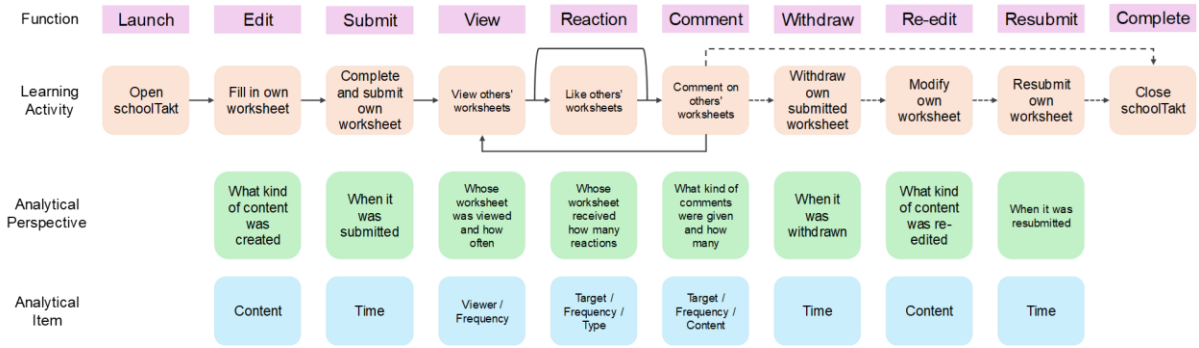


Figure 2. Example of Collaborative Learning Activities Using schoolTakt

2.3 Comparative Analysis (RQ3)

Finally, we compared the findings from the theoretical and empirical analyses. We evaluated the extent to which the interaction types identified in the literature (RQ1) align with the actions logged by the tools (RQ2). This comparative analysis aimed to determine the coverage and limitations of applying a standardized, existing vocabulary to describe both theoretical constructs and practical behaviors.

3. Results

3.1 Theoretical Analysis (RQ1)

The theoretical analysis covered 72 CSCL studies, which together utilized 338 distinct data instances for their analyses. The frequency of TLA Profile verbs used to describe these instances showed a clear and uneven distribution, as summarized in Figure 3. A small number of verbs related to data collection and evaluation were used frequently, while most of the 38 available verbs in the profile were rare. Specifically, analytical verbs like captured (used in 154 instances), tracked (135), and assessed (98) dominated the findings. This strong focus on monitoring and evaluation verbs suggests that the data models in the literature often prioritize

the act of analysis itself over the semantic description of specific collaborative processes. In other words, the research tends to describe the data as something that was "captured for analysis" rather than describing the rich, underlying collaborative action (e.g., "negotiating meaning" or "building consensus") that the data represents.

	All	(%)	Measured Data	(%)	Analyzed Data	(%)
データの個数	338		230		108	
captured	154	45.5621	137	59.5652	17	15.7407
tracked	135	39.9408	105	45.6522	30	27.7778
assessed	98	28.9941	56	24.3478	42	38.8889
evaluated	73	21.5976	39	16.9565	34	31.4815
explored	51	15.0888	29	12.6087	22	20.3704
contextualized	51	15.0888	25	10.8696	26	24.0741
surveyed	47	13.9053	43	18.6957	4	3.7037
socialized	26	7.6923	17	7.3913	9	8.3333
asserted	15	4.4379	15	6.5217	0	0.0000
organized	14	4.1420	10	4.3478	4	3.7037
clarified	10	2.9586	5	2.1739	5	4.6296
located	9	2.6627	9	3.9130	0	0.0000
directed	7	2.0710	2	0.8696	5	4.6296
selected	7	2.0710	2	0.8696	5	4.6296
inferred	6	1.7751	1	0.4348	5	4.6296
screened	5	1.4793	0	0.0000	5	4.6296
validated	3	0.8876	1	0.4348	2	1.8519

Figure 3. Verb Frequency in CSCL Literature

3.2 Empirical Analysis (RQ2)

The analysis of tool logs revealed a focus on more fundamental, content-creation actions. Table 1 provides a functional classification of the three major tools. While core functions like Edit and Submit are common, the implementation of peer-interaction functions like View, Comment, and Like/Stamp varies significantly. For example, while all tools support content creation (Edit), the mechanisms for peer interaction (View, Comment) are not standardized, posing a direct obstacle to cross-platform analysis of collaborative behaviors. The functions are primarily concrete, individual contributions to a shared space, as visualized in the activity flow in Figure 2.

Table 1. *Common Functions in Collaborative Learning Tools*

Action Category	Action Type	LoiLoNote	Okulink Plus	schoolTakt
Actions on the Content of the Learning Item	Edit	○	○	○
	Submit	○	○	○
	Send	○	○	–
	Receive	○	○	–
Actions on the Learning Item Itself	Withdraw	–	–	○
	Delete	○	○	*1
	Duplicate	○	○	○
Actions Toward the Learning Item	View	*2	*2	○
	Like	–	*3	○
	Stamp	–	○	–
	Comment	–	○	○

*1: LoiLoNote and Okulink Plus permit students to delete learning items from their own screens.
*2: In LoiLoNote and Okulink Plus, students must first receive an item and then import it into their own board for viewing.
*3: Okulink Plus offers several reaction stamps, including applause and flower stamps, commonly used in Japanese classrooms to indicate high praise.

3.3 Comparative Analysis (RQ3)

A clear discrepancy emerged between the theoretical and empirical findings. While research emphasizes analytical verbs like assessed and tracked, the tools primarily log foundational actions like edit, submit, and comment. This indicates a significant gap: existing tools are effective at capturing what students produce, but not necessarily how they collaborate in terms of higher-order cognitive and social processes like negotiation or metacognitive monitoring. An existing vocabulary like the TLA profile can describe many of the tool-based actions (e.g., posted for submitting, interacted for commenting). However, it lacks the specific granularity to differentiate the nuanced interactions that CSCL research deems important.

4. Discussion

This study investigated the feasibility of applying a standard xAPI vocabulary to collaborative learning logs, revealing a significant gap between the focus of academic research and the realities of tool-based data collection. This discussion explores the implications of this gap and proposes a pragmatic path forward.

4.1 The Gap Between Collaborative Learning Theory and Practice

Our findings show that while existing educational tools in the Japanese K-12 context log a variety of user actions, these are primarily primitive, task-oriented behaviors (e.g., editing, submitting). In contrast, CSCL research focuses on higher-order processes like meaning negotiation and metacognitive monitoring. This discrepancy is a key challenge for data-driven collaborative learning analytics. Current tools do not explicitly capture the complex cognitive and social interactions that are theoretically most valuable. This may be because tools are designed to support basic classroom management and content delivery, where simple, observable actions are easier to implement and interpret for teachers. Capturing higher-order processes often requires more sophisticated logging mechanisms or even manual coding, which is not practical for everyday classroom use. This suggests that many higher-order collaborative processes may occur implicitly or go unlogged due to these platform limitations.

4.2 Implications for Data Standardization

This limitation, however, does not diminish the importance of standardization. On the contrary, it highlights the need for a foundational layer of interoperability. Before we can analyze complex interactions, we must be able to reliably compare basic actions across different platforms. For educational service providers, agreeing on a common way to describe fundamental actions like "submitting a card" or "commenting on a peer's idea" is a necessary and pragmatic first step. This study demonstrates that existing xAPI vocabularies provide a viable, albeit incomplete, starting point for this endeavor. This directly addresses the practical challenge that even primitive analysis is difficult in the current fragmented data landscape. Without a shared understanding of what submit means across platforms, any analysis of higher-order processes that follow a submission is built on unstable ground.

4.3 A Pragmatic Path Forward

The findings suggest a clear, pragmatic path for advancing collaborative learning analytics. Instead of immediately attempting to create a single, all-encompassing xAPI profile that captures every nuance of collaboration, the community—including researchers, developers, and educators—should first focus on standardizing a "core" set of fundamental, observable actions. This core vocabulary, derived from the common functions identified in our empirical analysis (e.g., edited, submitted, viewed, commented), would establish a baseline of interoperability. Once this foundation is in place, more complex, context-specific layers can be

added to capture higher-order processes. This layered approach ensures that basic cross-platform analysis becomes immediately possible, while providing a stable framework upon which more sophisticated analytics can be built in the future. This approach allows for both immediate practical benefits and long-term research advancement.

4.4 Limitations and Future Work

This study has several limitations that suggest future research directions. First, its focus on Japanese K-12 educational tools limits the generalizability of the findings to other contexts where different pedagogical approaches and tools may be used. Second, by relying on existing log data and pre-defined verbs, we could not capture the full richness of higher-order collaborative behaviors that were not explicitly logged. The use of pre-defined verbs may have constrained the analysis, preventing the identification of more nuanced, context-specific actions.

Future work should follow the pragmatic path outlined above, including broader, multi-national, cross-platform studies to validate and expand the proposed “core” vocabulary. Research should also explore learning activity designs and tool features that make higher-order processes more observable and loggable, such as structured debates or peer-review workflows. Examining teacher dashboards or AI-driven feedback systems using this standardized core data would be valuable, potentially demonstrating the practical benefits of this foundational approach.

5. Conclusion

This study examined the alignment between collaborative learning theory and practice through the lens of data standardization. We found a clear gap between the higher-order interactions valued in research and the fundamental actions logged by widely used educational tools. While a comprehensive analysis of collaboration is not yet feasible with current logs alone, establishing a standardized vocabulary for foundational actions is a critical first step. Providing a common data language enables more robust, cross-platform analytics. This pragmatic approach will improve the utility of learning analytics for educators and pave the way for future research to capture more complex dynamics. Bridging the gap between theory and practice requires building from a solid, interoperable foundation—one action at a time.

References

Benesse Corporation. (2025). Miraiseed. https://bso.benesse.ne.jp/miraiseed/index.html

codeTakt Inc. (2025). schoolTakt. https://schooltakt.com/

Lang, C., Siemens, G., Wise, A., Gasevic, D., & Merceron, A. (Eds.). (2022). Handbook of learning analytics (2nd edition). SOLAR, Society for Learning Analytics and Research New York, NY, USA.

LoiLo Inc. (2025). LoiLoNote School. https://n.loilo.tv/ja/

Manso-Vázquez, M., Llamas-Nistal, M., & Mikic-Fonte, F. A. (2018). An xAPI Application Profile to Monitor Self-Regulated Learning Strategies.

Rosé, C. P., Fischer, F., Järvelä, S., & Weinberger, A. (Eds.). (2021). The international handbook of computer-supported collaborative learning. Springer.

Sawyer, R. K. (Ed.). (2014). The Cambridge handbook of the learning sciences (2nd ed.). Cambridge University Press.

Serrano-Laguna, Á., Martínez-Ortiz, I., Haag, J., Regan, D., Johnson, A., & Fernández-Manjón, B. (2016). Applying standards to systematize learning analytics in serious games. Computer Standards & Interfaces.

Appendix

You can view the bibliographic information of the final set of 72 peer-reviewed articles for Theoretical Analysis (RQ1) at the following URL.
https://drive.google.com/file/d/1v2-fL1AGeRodU24KkJJhDAvF8Q8SMQng/view?usp=sharing

Enhancing Retrospective Think-Aloud Protocols with Eye-Gaze Data

Amit PAIKRAO*, Debarshi NATH, Ramkumar RAJENDRAN, & Ritayan MITRA
Centre for Educational Technology, Indian Institute of Technology Bombay, India
*amitpaikrao@iitb.ac.in

Abstract: Recent advances in cognitive research indicate potential in integrating eye-tracking patterns with traditional retrospective think-aloud protocols. However, the exploration of this approach in learning and education has not taken place yet. This study compares conventional video-stimulated retrospective think-aloud with an eye-gaze-enhanced retrospective think-aloud method, examining their effectiveness in eliciting process-level cognitive data and metacognitive awareness. Preliminary findings indicate integration of eye gaze with retrospective think-aloud enhances verbalization precision of cognitive process-level data, reduces attentional ambiguity, minimizes cognitive reconstruction artifacts, and enables the formulation of targeted probes. These outcomes, aligned with Flavell's metacognitive monitoring framework, highlight multimodal protocol analysis in capturing the dynamic interplay between visual attention, decision-making, and self-regulated learning.

Keywords: Think-aloud, video-stimulated retrospective think-aloud, eye-tracking, eye gaze enhanced retrospective think-aloud, cognitive process level data, metacognition.

1. Introduction and Literature Review

Insights into cognitive processes that take place in computer-based learning environments remain significantly challenging, as the complex interplay of their thoughts, decisions, and metacognitive reflections often remains latent. In such scenarios, think-aloud protocols have emerged as valuable tools that allow researchers to gain deeper insights by prompting learners to verbalize their experiences (Ericsson & Simon, 1980). However, reliance on memory and verbal articulation inherently introduces abstraction that may obscure the fine-grained dynamics (Ericsson & Simon, 1993). Primarily, there are two variants of think-aloud protocols: concurrent think-aloud (CTA) and retrospective think-aloud (RTA) (Ericsson & Simon, 1980). In CTA, participants verbalise their thoughts while simultaneously performing the task. This method provides immediate insight into information held in working memory (W. Rainey Johnson et al. 2023). On the other hand, the retrospective think-aloud (RTA) method involves participants reflecting on their actions, decisions, and reasoning without the cognitive burden of simultaneous task performance and verbalization (Ericsson & Simon, 1993). Nonetheless, RTA's efficacy depends upon participants' memory recall capabilities, with a temporal delay between task completion and reflection introducing potential memory decay effects. When RTA implementation is not proximal to task execution, authentic cognitive processes may weaken significantly. RTA is inherently vulnerable to various response biases. Participants frequently engage in post-hoc cognitive reconstruction rather than accurate recall, thereby introducing threats to internal validity and maturation effects that compromise methodological rigour (Noushad et al., 2023). Additionally, conventional RTA protocols may inadequately capture the dynamic complexity of cognitive mechanisms, particularly in contexts involving visual-attentional processes and graphical interface interaction (Johnson et al., 2023). Given the limitations of traditional RTA methods, integrating RTA with video stimulation, known as video-stimulated recall (VS-RTA) (Lyle, 2003; Alhadreti & Mayhew, 2018), helps to overcome some of these limitations. VS-RTA has been used in teaching, counselling, psychiatry/medical research, nursing, and counselling studies. Video-stimulated think-aloud involves replaying the video of the task interaction during the retrospective interviews (Paikrao

& Mitra, 2023). This approach effectively leverages visual records to cue memory and helps to facilitate detailed reflections on the learning and problem-solving process. However, it is also susceptible to rationalization and reconstruction of events.

The subsequent integration in RTA involves the use of eye gaze data. While eye gaze data information is extremely valuable on its own (Rajendran et al., 2018), eye-tracking integrations in RTA, specifically superimposition of eye gaze with video and gaze plot for webpages. The eye-gaze data is overlaid onto the recorded interaction video in such a method. The resulting merged video is then used to prompt retrospective interviews with participants to elicit responses (Olsen et al., 2010; Elbabour et al., 2017). We refer to this integration of eye-gaze data (including fixations, saccades, and eye-gaze patterns) with retrospective think-aloud as ET-RTA. This methodology has been primarily explored in usability testing to identify more specific usability problems (Dewi et al., 2023; Olsen et al., 2010; Elbabour et al., 2017; Ishammari et al., 2015; Alhadreti & Mayhew, 2018; Pathan et al., 2025). The application of this methodology in educational contexts is minimal, and its potential can only be realised when this approach is investigated in different learning contexts. Existing research has not been able to address how eye-tracking technologies can be used in conjunction with the think-aloud protocol and what advantages such approaches can offer over traditional video-stimulated retrospective think-aloud. Given this predominant focus on usability testing problems, there remains a compelling research imperative to explore this combined methodology in the context of learning to shed light on cognitive and metacognitive processes. This identified research gap has motivated us to compare the video-stimulated retrospective think-aloud (VS-RTA) method with the eye-gaze enhanced retrospective think-aloud (ET-RTA). The pertinent research question is: How eye tracking integration enhances *retrospective think-aloud methodology beyond traditional video stimulation?* It seeks to extend eye-tracking applications beyond traditional usability domain toward a more comprehensive understanding of cognitive processes.

2. Methodology

We employed a single-participant within-subjects design as a proof-of-concept investigation with a single higher education student from electrical engineering. The goal was to trace the epistemic frames of participants as they engage in the role-playing game “The Power to the People” (https://www.rhombicogames.com/game/power-to-the-people/). The task in this game was to design an electrical grid and maintain a continuous power supply to the residents of the city. The study comprised three phases: gameplay (16 minutes), VS-RTA (30 minutes), and ET-RTA (42 minutes). We used Tobii Pro Lab software with a Tobii Pro Nano eye tracker (60 Hz) to record eye movements.

VS-RTA was conducted immediately after gameplay. Participant was shown the gameplay recording and prompted to recall his/her thoughts, decisions, and rationale for their actions. Upon completion of the VS-RTA interview, the same participant participated in the ET-RTA interview. In ET-RTA, the participant viewed the same recording but overlaid with recorded eye-gaze data (fixation points, saccades, and gaze paths). These interviews were recorded using OBS screen recorder. Figure 1 shows the VS-RTA in the left image and the ER-RTA in the right image. Participant provided informed consent for video recording, eye-tracking data collection, and use of anonymized data for research purposes.

All verbalizations were transcribed and analyzed by the first author, focusing on the interviewer’s questions and the participant’s responses. Using content analysis, we mapped the transcribed verbalizations to the corresponding gameplay to identify comparable scenes from both interview conditions’ thematic analysis. Resulting themes were discussed with Authors 2 and 3 to reduce bias and enhance the validity of findings.

3. Result

Our analysis compared verbal responses from VS-RTA and ET-RTA conditions across equivalent gameplay segments, revealing systematic differences in response quality, quantity, and cognitive depth. Quantitative analysis revealed significant differences in verbal productivity. During VS-RTA, the participant concluded the interview early, but continued elaborating when eye gaze patterns were shown. ET-RTA session yielded 1,233 words compared to 562 in VS-RTA, which amounts to more than double the verbal output. While no statistical inference is possible with a single case, this descriptive difference suggests a potential trend worth further investigation. The systematic comparison of equivalent gameplay segments identified four primary themes where ET-RTA demonstrated advantages over VS-RTA.

3.1 Reduction of Speculative Verbalizations

Participants frequently used tentative language in the VS-RTA, demonstrating uncertainty about actual thoughts during gameplay. For instance (scene 1: Figure 1), when discussing diesel generator placement, the participant stated: *"Probably I was thinking that it's too near from the city. It will pollute. So I put it far. Maybe that is the reason I thought."* This excerpt illustrates the speculative nature of recall, with phrases like "probably" and "maybe". However, when eye-gaze patterns were shown for the identical instance, the same participant provided more definitive statements: *"I was looking for the position of the diesel generator that will be appropriate. I was checking the distance from the city...".* This shows that ET-RTA helps eliminate speculative language and replaces uncertainty with concrete descriptions. This demonstrates how eye-gaze patterns provided concrete anchors for more precise recall.

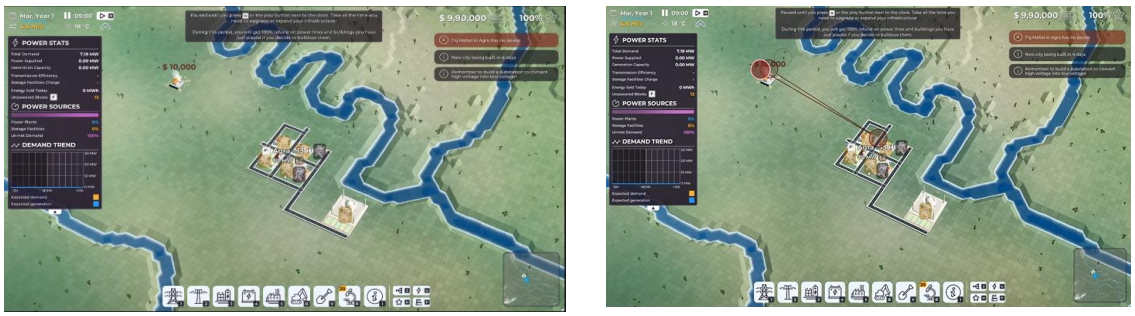


Figure 1. Scene 1 of gameplay for VS-RTA (left) and for ET-RTA (right) with fixations (red circles) and saccades (line connecting fixations)

3.2 Precision in Attentional Mapping

VS-RTA requires researchers to infer attentional focus from verbal descriptions and screen interactions, making this process vulnerable to confirmation bias. In this case, interview questions are primarily based on activities visible on-screen. In Scene 2, during VS-RTA, researchers formed questions based on observed screen selections: *"You selected the diesel generator? What made you choose it?"* To which the answer was: *"I was trying to put it near city. As its generating capacity was more..."*. In contrast, ET-RTA's eye-gaze data revealed the participant's visual attention patterns, enabling researchers to formulate targeted probes tailored to a specific fixation pattern (the participant's gaze hovered only on the selection button without reading the plant name and specifications). This observation led to a more precise, evidence-based question: *"While selecting this diesel generator, you haven't checked the full name and the ratings, you just selected it. What made you choose this option?".* This targeted probe elicited a markedly different and more honest response: *"I have not explored much, I saw a diesel generator. I just selected that..."*. This comparison demonstrates how ET-

RTA enables researchers to move beyond assumptions about participant attention and formulate questions that accurately reflect actual cognitive processes.

3.3 Mitigation of Reconstruction Artefacts

Participants in VS-RTA may describe what they are currently observing on screen, rather than recalling their actual thoughts. These reconstruction artefacts are particularly evident when describing static screens with minimal interaction. During Scene 3, when the participant began simulating grid operation, their verbalization focused exclusively on the visible changing numerical values: *"...transmission efficiency is not that good, the efficiency is 84-85%, very poor. Transmission efficiency should be greater than 95-96..."* This response demonstrates typical reconstruction behavior, describing observable screen elements without providing insight into the underlying decision-making process or attentional shifts. In contrast, ET-RTA with eye-gaze data showcases that the participant's attention pattern was more complex. The gaze data revealed the participant initially fixated on power statistics but then shifted attention to "Agra city." This evidence enabled rich verbalization of actual in-situ thought processes: *"...I was thinking the transmission efficiency should be high and capacity utilization should be high... because of Agra town is nearby, and it has higher capacity... Look this Taj Mahal is not powered. I was thinking to power up that. I was looking at power stats and the city because we have to give supply at higher efficiency.".* This response reveals the participant's actual strategic thinking for connecting transmission efficiency to geographic considerations and customer needs instead of simply describing visible numbers. This comparison illustrates how eye-gaze mitigates reconstruction artefacts by providing evidence of attention patterns, enabling participants to verbalize authentic cognitive processes rather than post-hoc rationalizations based on current screen observations.

3.4 Eliciting Granular Process-level Data

Participant often describe on-screen actions during VS-RTA, providing surface-level descriptions of what they did and basic reasoning. When confronted with eye gaze patterns, it provides a concrete visual representation of their cognitive processes. Reviewing eye movements enables a deeper understanding of attention patterns, decision-making, and problem-solving. In Scene 4, during VS-RTA, the participant provided a general description of the fault-recovery process: *"Probably something low voltage power line has failed as Taj Mahal, Agra has lost power, customer satisfaction has come down. I thought to provide on parallel, as I try to repair it...".* This response demonstrates a typical problem description with speculative language and limited insight into the detailed decision-making process. In contrast, when shown eye-gaze patterns revealing prolonged fixation on the "FAULT" indicator and revisits of repair costs, the participant's verbalization became significantly more granular and process-oriented: *"Here the line fault has occurred on the main line... I was thinking about the reasons of fault occurrence as I saw the customer satisfaction is going down. I repaired the fault for some $500. But it failed again, so again I was looking at how many funds are available... I thought to provide the second parallel transmission line, and share the load, it will reduce the heat, and fault won't reoccur...".* This reveals the complete cognitive sequence, i.e., fault identification, causal reasoning, cost consideration, outcome evaluation, and alternative solution generation. This demonstrates that eye gazes enabled the participant to articulate fine-grained problem-solving steps absent from the VS-RTA. This comparison illustrates how reviewing eye movements leads to increased self-awareness and improved verbalization of granular information about cognitive processes, transforming general action descriptions into detailed process-level insights.

4. Discussion and Conclusion

This exploratory case provides initial evidence that combining eye gaze information with RTA can enrich think-aloud protocols. While the findings are illustrative rather than generalizable, they point to promising directions for larger-scale validation. The four major themes that emerged in this study are enhanced memory precision through gaze cues, reduced attentional ambiguity, minimized cognitive reconstruction artifacts, and improved elicitation of process-level data. These align with established cognitive theories while offering novel protocol analysis for educational research. Previous research indicates that subjects develop fabricated thinking processes during stimulated recall due to incomplete memories and the human tendency to react rather than recall (Lyle, 2003; Güss et al., 2018). Our findings regarding minimized reconstruction highlight how eye gazes pinpoint attentional behavior for precise thought process verbalizations, providing crucial insights for designing effective learning environments and adaptive systems.

Flavell (1979), in his seminal work, highlighted that metacognitive experiences are frequently triggered when individuals are confronted with unfamiliar situations or challenging questions. Upon activation, these metacognitive experiences become a vital component of the cognitive monitoring system, providing crucial feedback and prompting strategic action (both cognitive and metacognitive). Our research indicates that displaying participants' eye-gaze data is **a potent metacognitive trigger** within the ET-RTA protocol.

VS-RTA faces significant challenges with non-veridicality, as granular details are not readily visible to interviewees or interviewers (Johnson et al., 2023; Ericsson & Simon, 1993). Our findings (in section 3.4) demonstrate that ET-RTA directly addresses this issue by externalizing implicit cognitive processes. Crucial cognitive actions that participants omitted were evident in eye-gaze patterns. Enhanced granularity corroborates Ericsson and Simon's (1993) assertion that verbal reports primarily reflect cognitive products unless augmented by externalized cues. Our findings extend this principle by demonstrating how gaze visualization targets the process dimension of cognition, making implicit evaluation criteria explicit and verifiable. This documentation is particularly critical in game-based learning contexts, where iterative strategy refinement constitutes a core learning mechanism.

Eye-gaze patterns trigger metacognitive activation, prompting deeper reflection on attentional processes and cognitive states. Participants observing eye movements on inactive screen elements recalled underlying monitoring behaviors, suggesting externalized gaze data acts as a catalyst for self-reflection and richer verbal data. Our quantitative and qualitative (section 3.4) analyses demonstrate ET-RTA's effectiveness in triggering metacognition and motivating richer verbal responses.

VS-RTA relies on participants' short-term memory during interviews, which organizes information in small chunks, limiting recall accuracy (Ericsson and Simon 1993; Johnson et al., 2023). Our study found VS-RTA's unassisted memory recall resulted in tentative language (section 3.1) as they struggled connecting video cues with fading memories. ET-RTA's gaze replays provided specific visual reference points that reduced tentative language and increased clarity, which helps overcome the limitation of VS-RTA regarding the temporal decay of episodic memory.

5. Conclusion and Limitation

Through a detailed case, this study illustrates how integrating eye-gaze data into retrospective think-aloud may help address some limitations of video-stimulated protocols. Findings overcome the challenges of VS-RTA, such as memory decay and cognitive reconstruction. Two of our major themes highlight the methodological contributions of this approach: first, it demonstrates that eye-gaze can act as a potent **metacognitive trigger**, leading to the elicitation of granular cognitive process data; and second, it streamlines the interview process by enabling the formulation of targeted probes based on specific fixation patterns. These findings have significant pedagogical implications, as ET-RTA can be adapted as a learning

tool to help students develop better self-regulation and strategic thinking skills. By extending the application of ET-RTA beyond its traditional use in usability studies, our research provides new insights into its potential for educational contexts.

The within-subjects, single-participant design constrains generalizability, and the fixed order of protocols introduces possible practice and order effects. As such, the results should be viewed as exploratory. Nevertheless, the analysis demonstrates the feasibility of ET-RTA in educational contexts and generates hypotheses for larger, counterbalanced studies. Future research should explore a broader range of learning contexts and a larger sample size to provide a more comprehensive list of ET-RTA's benefits in learning and instruction.

Acknowledgements

This work is supported by the TIH Foundation for IoT and IoE under grant ID RD/0123-TIH0018-003.
We thank Federico Casares, Founder of Rhombico Games, for allowing us to use “The Power to the People” game for our study.

References

Alhadreti, O., & Mayhew, P. (2018, April). Rethinking thinking aloud: A comparison of three think-aloud protocols. In Proceedings of the 2018 CHI conference on human factors in computing systems (pp. 1-12).

Dewi, G. A. P. A. (2023). *USABILITY TESTING PADA APLIKASI PERSEDIAAN KABUPATEN BADUNG MENGGUNAKAN TEKNIK PERFORMANCE MEASUREMENT, RTA, DAN SUS* (Doctoral dissertation, Universitas Pendidikan Ganesha).

Elbabour, F., Alhadreti, O., & Mayhew, P. (2017). Eye tracking in retrospective think-aloud usability testing: Is there added value?. *Journal of Usability Studies*, *12*(3), 95-110.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological review*, *87*(3), 215.

Ericsson, K. A., & Simon, H. A. (1993). Protocol analysis (revised edition). Overview of methodology of protocol analysis.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. American Psychologist, 34, 906 - 911.

Güss, C. D., Tuason, M. T., Göltenboth, N., and Mironova, A. (2018). Creativity through the eyes of professional artists in Cuba, Germany, and Russia. J. Cross Cult. Psychol. 49, 261–289. doi: 10.1177/00220221177308172.

Johnson, W. R., Artino Jr, A. R., & Durning, S. J. (2023). Using the think-aloud protocol in health professions education: an interview method for exploring thought processes: AMEE Guide No. 151. *Medical teacher*, *45*(9), 937-948.

Lyle, J. (2003). Stimulated recall: A report on its use in naturalistic research. British educational research journal, 29(6), 861-878.

Noushad, B., Van Gerven, P. W. M., & de Bruin, A. B. H. (2023). Twelve tips for applying the think-aloud method to capture cognitive processes. *Medical Teacher*, *46*(7), 892–897. https://doi.org/10.1080/0142159X.2023.2289847

Olsen, A., Smolentzov, L., & Strandvall, T. (2010, September). Comparing different eye tracking cues when using the retrospective think aloud method in usability testing. In Proceedings of HCI 2010. BCS Learning & Development.

Paikrao, A. M., & Mitra, R. (2023). Expert-Novice Differences In Electrical Circuit Analysis Based On The Order Of Attention. European Society for Engineering Education (SEFI). DOI: 10.21427/G9YB-5J15

Pathan, R., Murthy, S., & Rajendran, R. (2025). Identifying cognitive and metacognitive processes using retrospective and concurrent think-aloud protocols. In B. Jiang et al. (Eds.), Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education.

Rajendran, R., Kumar, A., Carter, K. E., Levin, D. T., & Biswas, G. (2018). Predicting Learning by Analyzing Eye-Gaze Data of Reading Behavior. *International Educational Data Mining Society*.

W. Rainey Johnson, Anthony R. Artino Jr & Steven J. Durning (2023), Using the think aloud protocol in health professions education: an interview method for exploring thought processes: AMEE Guide No. 151, Medical Teacher, 45:9, 937-948, DOI: 10.1080/0142159X.2022.2155123

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Improvement of a Japanese Language Learning Support System that Enables Word Accent Learning

Kazuki TOMITA[a*], Satoru KOGURE[b], Yasuhiro NOGUCHI[b], Koichi YAMASHITA[c], Tatsuhiro KONISHI[b] & Makoto KONDO[b]
[a]*Graduate School of Integrated Science and Technology, Shizuoka University, Japan*
[b]*Faculty of Informatics, Shizuoka University, Japan*
[c]*Faculty of Business Administration, Tokoha University, Japan*
*tomita.kazuki.20@shizuoka.ac.jp

Abstract: Prosody is an important element in verbal communication. Accent, a form of prosody, plays a role in smooth communication in Japanese. In a previous study, we developed a system that visualizes Japanese word accents as accent graphs. In this study, we analyzed correct accent data and added algorithms to improve the accuracy of the accent estimation system, as well as an accent correctness judgment system. As a result, the system developed in this study was more accurate than the previous one.

Keywords: Japanese language learning, Japanese word accent recognition, accent graph

1. Introduction

In recent years, the number of foreign residents and workers in Japan has increased. To live in Japan, they must be able to communicate in Japanese. Prosody is a crucial element of spoken communication, and among its components, accent is particularly significant in Japanese because it helps distinguish word meanings. Japanese employs a pitch accent system, which plays a vital role in preventing misunderstandings and ensuring smooth communication. In previous research, we developed a Japanese dictogloss learning support system that provides an environment for acquiring Japanese pitch accents (Kogure et al., 2024). The present study aims to improve the accuracy of accent estimation and correctness judgment in that system.

2. Related Works

Matsuzaki and Takahashi (2015) addressed prosody learning systems for Japanese. In their system, learners' utterances and model utterances are visualized as “prosody graphs.” These graphs display pitch based on the prosodic features of Japanese and allow learners to compare their utterances with the model visually. However, a limitation of prosody graphs is that slight deviations in the slope between the learner's and model's graphs can mislead learners, causing them to overlook significant errors.

Kogure et al. (2024) proposed a system that visualizes pitch using an "accent graph." Like prosody graphs, accent graphs display pitch based on Japanese prosody but differ by representing pitch using only two levels: "High" and "Low." This binary representation avoids the issue of minor slope differences in prosody graphs. Therefore, the present study adopted accent graphs for pitch visualization.

The Online Japanese Accent Dictionary (OJAD) is a dictionary for Japanese learners (Minematsu et al., 2012). We used OJAD's pitch-accent information as ground-truth data, and a subset of its voice samples for verbs was employed in the evaluation experiments.

3. Japanese Pitch Accent in This Study

This study followed the Tokyo dialect accent standard. Figure 1 illustrates the accent types of four-mora words—Type 0, Type 1, Type 3, and Type 4—with each mora represented by a dot. The accent core is shown in white. In Type 4, a pitch drop occurs on the postpositional particle *ga* (the fifth mora), indicating a falling pitch after the noun. Focusing on word-level accent learning, only nouns and verbs with two or more morae are considered. To improve the accuracy of accent estimation and correctness judgment, we analyzed OJAD's reference audio data. Using *Praat*, we visualized pitch graphs for 90 audio samples and manually identified distinguishing characteristics between Type 0 and Type 1 accent words. These features were then used to enhance the system's algorithms.

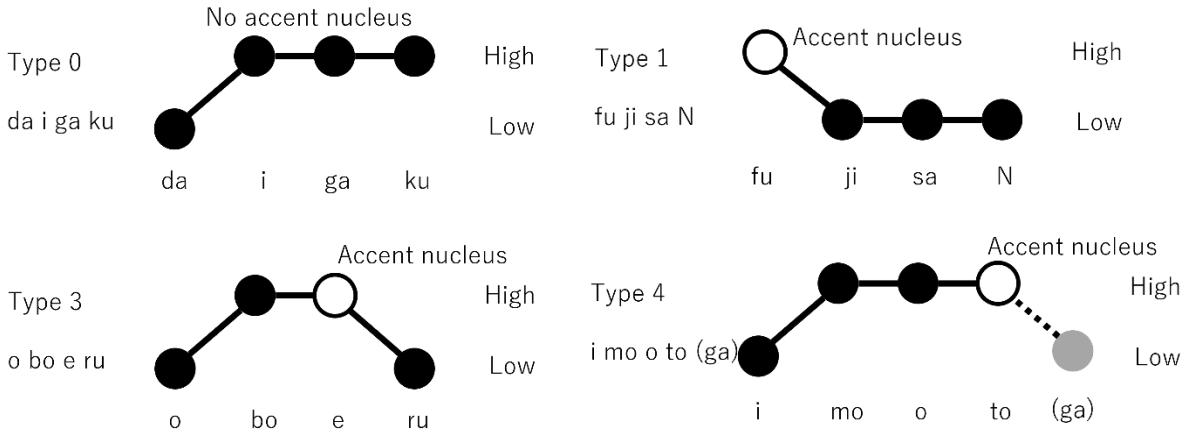


Figure 1. Examples of pitch graphs for each accent type.

4. Accent Learning System

The system estimated accent graphs based on the high and low pitch levels for each mora. The estimation algorithm was implemented using *Praat* scripts. Building on Kogure (2024), we refined the algorithm using the characteristics identified in Chapter 3 for Type 0 and Type 1 accent words. First, a reference pitch value was set. Then, the average fundamental frequency of each mora was measured and compared with the reference pitch. If the average pitch was equal to or higher than the reference, it was labeled "High"; otherwise, "Low." The reference value was adjusted depending on the accent type: for Type 0 and Type 1, the first mora's pitch was used as the reference; for other types, the overall average pitch was used.

Correctness was assessed by comparing the estimated pitch pattern with the correct patterns derived from the OJAD and Japanese accent definitions. If the pitch values matched each mora, the results were considered correct. Additionally, based on the characteristics of Type 0 and Type 1 accent words identified in Chapter 3, we added new rules to the judgment algorithm to improve accuracy.

Feedback was provided in both text and visual formats based on the learner's audio and the system's estimation and judgment results. The feedback text indicated whether the accent was correct. The feedback image included the learner's waveform, pitch graph, mora segmentation, and both the estimated and correct accent graphs. These images were generated using *Praat* scripts. Auditory feedback was also provided: learners could listen to both their pronunciation and the correct pronunciation from OJAD, allowing for auditory comparison.

5. Evaluation Experiments

We evaluated the accuracy of the improved accent estimation and correctness judgment system. For comparison, we used the system developed by Kogure et al. (2024). The evaluation used 370 audio samples: 90 correct samples from OJAD and 280 samples (both correct and incorrect) from native Japanese speakers. Both systems performed accent estimation and correctness judgment on these samples, and we calculated the accuracy and F-measure as a comprehensive accuracy metrics.

As a result, the accuracy of the research system by Kogure et al. (2024) was 0.692 and the F-value was 0.674, whereas the accuracy of the system in this study was 0.816 and the F-value was 0.848. This improvement is attributed to the fact that the previous system often misjudged correctly pronounced accents as incorrect, which the new system successfully addressed.

6. Conclusion

In this study, by analyzing speech uttered with correct accent and enhancing the system with new algorithms, we aimed to improve the accuracy of the accent estimation and correctness judgment system developed by Kogure et al. (2024). As a result, the accuracy of accent estimation and accent correctness judgment system was improved.

We plan to conduct a usability evaluation experiment using the developed system interface. Participants will use the system and complete a questionnaire regarding their experiences. Additionally, we plan to develop a system that can support the learning of Japanese compound accents and sentence intonation by extending the learning targets.

Acknowledgements

This study was supported by JSPS KAKENHI (Grant Number: JP24K15205).

References

Boersma, P., & Weenink, D. (2024). *Praat: doing p honetics by computer* (Version 6.2.06) [Software]. https://www.fon.hum.uva.nl/praat/

Immigration Services Agency. (2024). *Number of foreign residents as of the end of June 2024*. https://www.moj.go.jp/isa/publications/press/13_00047.html (in Japanese)

Kogure, S., Tomita, K., Noguchi, Y., Yamashita, K., Konishi, T., & Kondo, M. (2024). Construction of a Japanese language learning support system that enables word accent learning. *Proceedings of the 31st International Conference on Computers in Education* (ICCE 2023) (pp. 329-334)

Lee, A., & Kawahara, T. (2019). *Julius* (Version 4.5) [Software]. https://doi.org/10.5281/zenodo.2530395

Matsuzaki, H., & Takahashi, E. (2015). Development of a system for automatic output of prosody graphs corresponding to a delayed fundamental frequency fall. *Journal of Japanese Language Education Methods, 22*(1), 48-49 (in Japanese). https://doi.org/10.19022/jlem.22.1_48

Minematsu, N., Kobayashi, S., Shimizu, S., & Hirose, K. (2012). Improved prediction of Japanese word accent sandhi using CRF. *Proceedings of Interspeech 2012 ISCA's 13th Annual Conference* (pp. 2562–2565). https://doi.org/10.21437/interspeech.2012-663

Ministry of Health, Labor, and Welfare. (2024). *Summary of reporting status of "Employment of foreign nationals" (as of the end of October 2023)*. https://www.mhlw.go.jp/stf/newpage_37084.html (in Japanese)

The Japan Foundation. (2009). Japanese language teaching methods series, Vol. 2: Teaching Voice. The Japanese Foundation. (in Japanese)

INSTd: A Computer-Adaptive Assessment of Definition-Based Comprehension Skills

Noriko ARAI[a*], Naoya TODO[b], Teiko ARAI[c], Shingo SUGAWARA[d], Takuya MATSUZAKI[e] & Koken OZAKI[f]
[a]*National Institute of Informatics, Japan*
[b]*Tokyo Metropolitan University, Japan*
[c]*Takasaki City University of Economics, Japan*
[d]*Institute of Science for Education, Japan*
[e]*Tokyo University of Science, Japan*
[f]*University of Tsukuba, Japan*
*arai@nii.ac.jp

Abstract: This paper introduces INSTd, a computer-adaptive test designed to assess learners' ability to interpret dictionary-style definitions and apply them in appropriate contexts. Unlike conventional vocabulary tests that measure already-acquired word knowledge, this test presents definitions and evaluates learners' ability to acquire new words based on those definitions. We analyzed the correlation of the estimated abilities of INSTd with their vocabulary sizes among 1,221 sixth-grade students. Results show a moderate correlation with vocabulary size ($r = 0.343$, $p < .001$), indicating that the test, although related to vocabulary, captures a distinct aspect of reading ability.

Keywords: definition comprehension, vocabulary assessment, digital learning environment, item response theory

1. Introduction

In digital learning environments, such as online textbooks and platforms like Wikipedia, it is common practice to hyperlink unfamiliar terms to their dictionary definitions. It reflects a widely accepted assumption: that learners can interpret these definitions and apply them as needed to acquire new vocabulary. However, is this assumption justified?

Most prior research on vocabulary has focused on size or depth—what learners already know (Anderson, 1981; Begler, 2010; Sternberg, Powell & Kaye, 1987). Fewer studies have investigated whether learners can successfully interpret unfamiliar definitions when presented in context (Arai et al, 2025). This paper introduces INSTd, a new test designed to measure this often-overlooked ability. In each item, examinees are presented with a short dictionary-style definition and asked to select the appropriate examples or applications from multiple-choice options.

Empirical validation suggests that INSTd is psychometrically robust. Results from a study involving 1,221 sixth-grade students showed a moderate positive correlation between estimated ability scores on INSTd and vocabulary size ($r = 0.343$, $p < .001$). This suggests that while INSTd is related to vocabulary knowledge, it measures a different dimension of reading ability.

2. Test Design

INSTd is a computer-adaptive test designed to measure examinees' ability to read short, dictionary-style definitions and apply them accurately across different contexts. Each item presents a concise definition, a simple instruction, and four multiple-choice options, with one or more correct answers. The test interface is deliberately minimal, featuring one item per screen and a clear "Submit" button, which helps examinees concentrate solely on interpreting the definition. Figure 1 illustrates an example of INSTd questions.

Definitions are primarily sourced from widely used Japanese educational materials, including the popular elementary-level dictionary Reikai Shougaku Kokugo Jiten (Sanseido, 2024), as well as Wikipedia and textbooks. Vocabulary terms are selected based on definitional clarity.

Each item is authored by an experienced developer and reviewed by at least two reviewers to verify that the definition is unambiguous, the options are appropriate, and the correct answers logically follow from the definition.

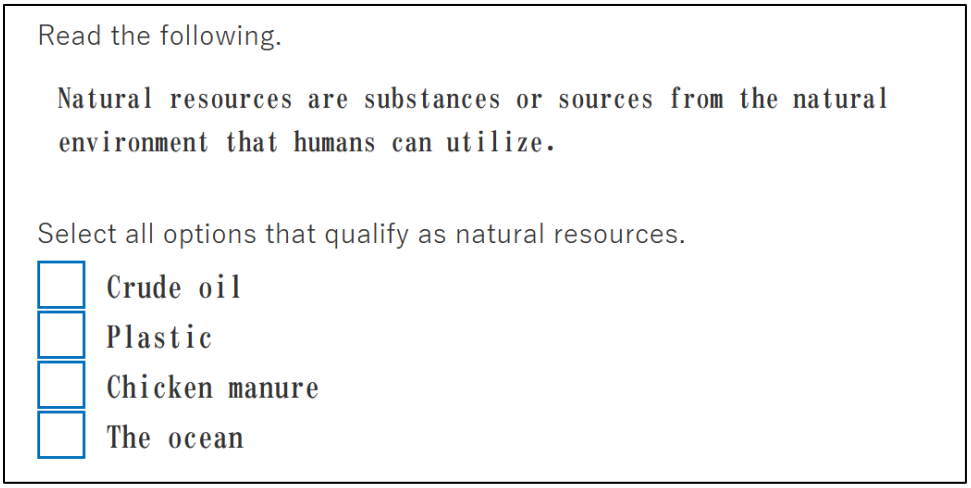


Figure 1. Screenshot of an INSTd question. Correct answer: 'crude oil'.

3. INSTd as an IRT-Based Computer Adaptive Test

We implemented INSTd as a computer-adaptive test based on the 2PL model (van der Linde & Glas, 2010). The probability that an examinee with ability θ_i will answer item j correctly is given by the 2PL model:

$$P_j(\theta_i) = \frac{1}{1 + \exp\left[-k \cdot a_j(\theta_i - b_j)\right]}.$$

Here, k is a constant approximating the normal ogive model ($k = 1.7$ in INSTd), a_j is the item discrimination parameter, and b_j is the item difficulty parameter.

The average factor loading of the items was 0.521, and the IRT-based reliability coefficient was 0.768. These values indicate that the items measure a coherent latent trait and provide sufficiently precise ability estimates for diagnostic purposes.

These results confirm that INSTd meets essential psychometric criteria and can reliably assess definition-based comprehension across a wide range of examinee populations.

During the allotted five-minute testing period, examinees are instructed to answer as many questions as possible, as accurately as possible. The items presented to each examinee are selected based on their prior responses and accuracy, following a computer-adaptive testing algorithm explained in Section 3. As a result, each examinee completes a different number of items and encounters a unique sequence of questions.

4. Results

To investigate the relationship between INSTd and traditional measures of vocabulary knowledge, we analyzed its correlation with a vocabulary size survey conducted among 1,221 sixth-grade students in City A. In this survey, each examinee was presented with a list of 60 words and asked to indicate whether they knew each word. Responses were coded as binary values (1 = known, 0 = unknown or unsure), and the total number of known words was calculated as an estimate of vocabulary size.

The 60 words were carefully selected based on familiarity ratings from a large-scale database provided by NTT Communication Science Laboratories (Amano & Kondo, 1998). Words were chosen to span a wide range of familiarity levels, including highly familiar terms (e.g., “bank”), moderately familiar terms (e.g., “economy”), and low-familiarity terms (e.g., “recoinage”). This design ensured that the vocabulary size measure could differentiate examinees across a broad range of lexical knowledge.

The correlation between INSTd ability estimates and vocabulary size was found to be $r = 0.343$, statistically significant at $p < .001$. While moderate in strength, this correlation suggests that INSTd measures a construct that is related to—but distinct from—vocabulary breadth.

This result suggests that while hyperlink-based introduction of unfamiliar terms in digital learning platforms may support learners with extensive vocabularies, it may not adequately serve those with more limited lexical knowledge and underdeveloped definition-based comprehension skills as measured by INSTd.

5. Conclusion

This study introduced INSTd, a computer-adaptive test designed to assess learners’ ability to interpret and apply dictionary-style definitions—an ability that underpins many forms of learning in digital environments but has rarely been assessed directly.

Through rigorous psychometric development and large-scale validation, INSTd has been shown to capture a distinct dimension of lexical competence, one that complements vocabulary size. Its scalable design and diagnostic precision make it a promising tool for both educational assessment and personalizing digital reading experiences.

Acknowledgments

This research was supported by the Japan Society for the Promotion of Science (JSPS) under Grants-in-Aid for Scientific Research JP16H01819 and JP21H04416 and by Reading Skill Test, Inc. We also thank the Institute of Science for Education for providing access to INSTd data and screen captures used in this study.

References

Amano, S., & Kondo, T. (1998). Estimation of mental lexicon size with word familiarity database. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP)* (Vol. 5, pp. 2119–2122).

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 77–117). International Reading Association.

Arai, T., Todo, N., Ozaki, K., & Arai, N. H. (2025). *Assessing the ability to read and interpret mathematical definitions: Insights from INSTm*. In Proceedings of the 25th IEEE International Conference on Advanced Learning Technologies (ICALT 2025). IEEE.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27*(1), 101–118. https://doi.org/10.1177/0265532209340194.

Sanseido. (2024). *Reikai Gakushū Kokugo Jiten* (11th ed.) [例解学習国語辞典・第11版, in Japanese]. Sanseido.

Sternberg, R. J., Powell, J. S., & Kaye, D. B. (1987). Measuring the depth of vocabulary knowledge: Toward a bridge between vocabulary learning and reading comprehension. *Journal of Educational Psychology, 79*(4), 269–275.

van der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing*. Springer. https://doi.org/10.1007/978-0-387-85461-6

A Dual-Framework Approach to Evaluate Marker less AR Application: Insights from SUS and HARUS Questionnaires

Ajay Shankar Tiwari[a], Toluchuri SHALINI[b*], Utanko Mitra[c], & Kaushal Kumar BHAGAT[d*]

[a,b,c,d]Advanced Technology Development Centre, IIT Kharagpur
*kkntnu@hotmail.com

Abstract: This study evaluated a marker-less AR application for learning Engineering Drawing using SUS and HARUS to assess both general and AR-specific usability. Results showed high user satisfaction (SUS score: 88.6) and highlighted the application's effectiveness in engaging learners and supporting understanding of complex concepts. The findings emphasize the value of combining usability measures to guide the design of effective AR-based educational tools.

Keywords: Augmented Reality, Usability, Engineering Drawing, System Usability Scale, Hand-Held AR Usability Scale

1. Introduction

Augmented Reality (AR) is increasingly used in education to create immersive and interactive learning experiences (Azuma et al., 2001; Carmigniani et al., 2011; Lee, 2012). In visually intensive subjects like Engineering Drawing (ED), marker-less AR allows students to explore complex 3D models without physical markers, enhancing flexibility and engagement (Cheng et al., 2017; Chytas et al., 2020; Tiwari et al., 2024). Despite its potential, the usability of such tools remains underexplored, which can affect their educational impact. This study addresses this gap by evaluating a marker-less AR-based ED application using two instruments: the System Usability Scale (SUS) (Brooke, 1995) for general usability and the Hand-held AR Usability Scale (HARUS) (Santos et al., 2014) for AR-specific aspects like immersion and intuitiveness. Combining both tools offers a comprehensive assessment (Carrera et al., 2018; Derby & Chaparro, 2021; Law & Heintz, 2021). The study aims: (1) to assess general usability using SUS, (2) to evaluate AR-specific heuristics using HARUS, and (3) to compare both frameworks to derive insights for improving marker-less AR applications in education.

2. Methodology

The study involved 15 undergraduate students (9 males and 6 females) with prior knowledge of Engineering Drawing (ED), selected through random sampling to ensure diversity and reduce selection bias. Participants came from varied academic backgrounds but shared a basic understanding of ED concepts. They interacted with a previously developed marker-less AR application (Tiwari et al., 2024) (see Figure 1), which allowed them to visualize and manipulate 3D projections of engineering components in real-world settings without physical markers. Developed using Unity, the application featured a minimalistic, touch-based interface with intuitive gestures like pinch-to-zoom, swipe-to-rotate, and drag-to-move, making it accessible for first-time users. The marker-less functionality provided flexibility in model interaction from multiple angles. After a 10-minute session with the application, participants completed two usability questionnaires—SUS and HARUS—to assess both general and AR-specific usability factors. All responses were collected with assurances of confidentiality and voluntary participation.

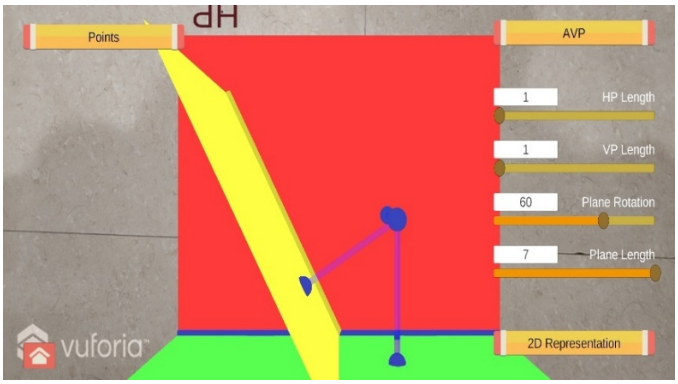


Figure 1. Projection of points.

3. Results

3.1 SUS Scores

We calculated SUS score using technique proposed by (Bangor et al., 2009). The overall SUS score for the AR-based application was calculated to be approximately 88.6. This score indicates that the application is perceived as highly usable and satisfactory by the participants (see Figure 2). The usability of the marker-less AR application was evaluated across two key dimensions: Manipulability and Comprehensibility, based on participant responses. The overall mean score for Manipulability was *4.34*, indicating that users generally found the application physically easy to use. Most participants agreed that operating the app required minimal effort, was comfortable for the hands and arms, and was simple to control. However, some noted slight discomfort in holding the device or potential grip issues. In terms of Comprehensibility, the application received an overall mean score of *4.10*, reflecting a generally positive cognitive experience. Users found the information clear, readable, and consistent, although a few noted minor concerns regarding screen clutter and display response time. Overall, both physical and cognitive usability aspects were rated favorably, suggesting that the application was user-friendly and accessible.

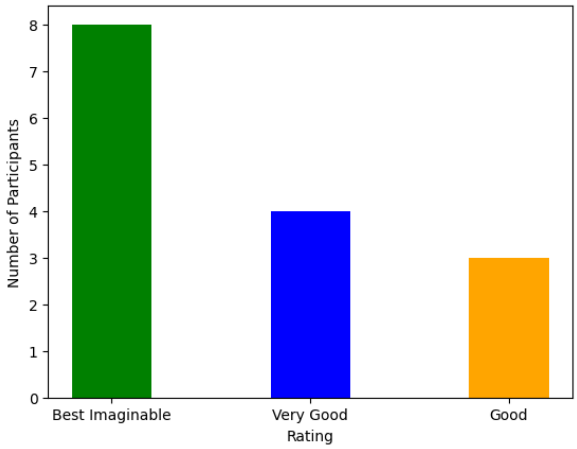


Figure 2. SUS Rating of AR Based application.

3.2 HARUS analysis

The evaluation of the marker-less AR application showed positive user perceptions in both manipulability (Mean = 4.34, *SD* = 1.48) and comprehensibility (Mean = 4.10, *SD* = 1.37). Participants generally found the application physically comfortable, easy to handle, and simple to operate, indicating minimal physical strain during interaction. Similarly, users perceived the

displayed information as readable, consistent, and responsive, suggesting manageable cognitive demands. However, slightly lower ratings on control precision, the appropriateness of displayed information, and system responsiveness highlight minor areas for improvement to further enhance the overall user experience.

4. Discussions and conclusions

This study investigated the usability of a marker-less AR application for learning Engineering Drawing (ED), guided by three research questions and evaluated using SUS and HARUS questionnaires. The high SUS score of 88.6 reflected excellent general usability, with participants describing the application as engaging and effective. HARUS results further highlighted the importance of AR-specific factors like immersion and interaction, while also revealing areas for improvement, such as interface sensitivity and the need for more advanced 3D models. Comparing SUS and HARUS underscored the value of using both tools for a comprehensive evaluation of usability and user experience. Despite promising results, the study's limitations—such as a small, homogenous sample, short interaction time, and reliance on self-reported data—suggest caution in generalizing findings. Future research should expand the participant pool, increase interaction duration, and integrate learning outcome assessments and adaptive features to better evaluate AR's long-term educational impact.

References

Azuma, R., Baillot, Y., Behringer, R., Feiner, S., Julier, S., & MacIntyre, B. (2001). Recent advances in augmented reality. *IEEE Computer Graphics and Applications, 21*(6), 34–47. https://doi.org/10.1109/38.963459

Cheng, J., Chen, K., & Chen, W. (2017). Comparison of marker-based AR and marker less AR: A case study on indoor decoration system. In *Lean and Computing in Construction Congress (LC3): Proceedings of the Joint Conference on Computing in Construction (JC3)* (pp. 483–490).

Brooke, J. (1995). SUS: A quick and dirty usability scale. *Usability Evaluation in Industry, 189*, 189–194.

Santos, M. E. C., Taketomi, T., Sandor, C., Polvi, J., Yamamoto, G., & Kato, H. (2014). A usability scale for handheld augmented reality. In *Proceedings of the 20th ACM Symposium on Virtual Reality Software and Technology* (pp. 167–176).

Carmigniani, J., Furht, B., Anisetti, M., Ceravolo, P., Damiani, E., & Ivkovic, M. (2011). Augmented reality technologies, systems and applications. *Multimedia Tools and Applications, 51*, 341–377. https://doi.org/10.1007/s11042-010-0660-6

Lee, K. (2012). Augmented reality in education and training. *TechTrends, 56*, 13–21. https://doi.org/10.1007/s11528-012-0559-3

Chytas, D., Johnson, E. O., Piagkou, M., Mazarakis, A., Babis, G. C., Chronopoulos, E., Nikolaou, V. S., Lazaridis, N., & Natsis, K. (2020). The role of augmented reality in anatomical education: An overview. *Annals of Anatomy - Anatomischer Anzeiger, 229*, 151463.

Tiwari, A. S., Bhagat, K. K., & Lampropoulos, G. (2024). Designing and evaluating an augmented reality system for an engineering drawing course. *Smart Learning Environments, 11*(1), 1.

Derby, J. L., & Chaparro, B. S. (2021). The challenges of evaluating the usability of augmented reality (AR). In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 65*, 994–998.

Carrera, C. C., Perez, J. L. S., & Cantero, J. d. l. T. (2018). Teaching with AR as a tool for relief visualization: Usability and motivation study. *International Research in Geographical and Environmental Education, 27*(1), 69–84.

Law, E. L.-C., & Heintz, M. (2021). Augmented reality applications for K-12 education: A systematic review from the usability and user experience perspective. *International Journal of Child-Computer Interaction, 30*, 100321.

Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies, 4*(3), 114–123.

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Multimodal Trait Scoring for Video Interviews Using Neural Models with Handcrafted Features and Trait-Attention

Taichi KITAJIMA* **& Masaki UTO***
The University of Electro-Communications, Japan
*{kitajima, uto}@ai.lab.uec.ac.jp

Abstract: Interview examinations are widely used in various educational assessments to evaluate students' interpersonal skills. Automated interview scoring methods that predict scores from video recordings of interviews using artificial intelligence technologies have recently attracted considerable attention. The primary limitations of traditional methods are twofold. First, they depend solely on either handcrafted or neural features. Second, although traditional methods are typically designed as trait-scoring models, they overlook inter-trait correlations. To address these limitations, this study proposes a trait-scoring model for interview examinations that predicts multiple trait scores by incorporating inter-trait correlations and combining handcrafted features with neural features derived from pre-trained language and computer vision models.

Keywords: Automated interview scoring, trait-scoring, large language models, computer vision, multimodal data analysis

1. Introduction

Interview examinations are widely used in various educational assessments to assess students' interpersonal skills, which are difficult to evaluate through objective tests. However, manual evaluation poses significant challenges, including dependency on rater characteristics and substantial time and cost requirements. To address these issues, automated scoring of interview examinations has been gaining increasing attention (Hemamou et al., 2023; Naim et al., 2015). Automated interview scoring aims to predict scores by using artificial intelligence technology to analyze multimodal data, including video, speech audio, and dialogue text recorded during interviews. Such systems can offer consistent and timely evaluations to students, thereby enhancing both the efficiency and fairness of evaluations, and holding significant potential for educational applications.

Traditional automated scoring methods for interview evaluations have relied primarily on handcrafted features (Naim et al., 2015). However, the handcrafted features might not fully capture the complex and valuable information inherent in multimodal data. In contrast, recent approaches have employed deep learning techniques to automatically extract features that are effective for automated scoring (Hemamou et al., 2023). These methods utilize either neural features or handcrafted features for each modality. However, neural and handcrafted features are often complementary (Ridley et al., 2021), suggesting that incorporating both types of features for each modality has the potential to improve interview scoring accuracy. Additionally, interview examinations typically involve trait scoring based on multiple evaluation criteria. However, conventional methods are designed to predict each trait score independently.

To address these challenges, the present study proposes a trait-scoring model for interview examinations that leverages inter-trait correlations and integrates hybrid input from both handcrafted and neural features across all modalities. The effectiveness of the proposed method is demonstrated through an evaluation using a real-world interview dataset.

2. Proposed Method

Letting A denote the audio data, V the video data, T the dialogue text data, and H the handcrafted features derived from $\{A, V, T\}$, our task is to construct the function $\boldsymbol{S} \leftarrow f_{\boldsymbol{\theta}}(A, V, T, H)$ that predicts the trait scores $\boldsymbol{S} = \{s_1, \ldots, s_K\}$, where K represents the number of traits, s_k is the score for the k-th trait, and $\boldsymbol{\theta}$ denotes the parameters. This study assumes that each interview follows a standardized format, in which all interviewees are asked the same questions in the same order. Accordingly, each set of modality data A, V, and T includes the interviewee's responses to each of the Q questions, denoted as A_q, V_q, and T_q, where each element corresponds to the modality data for the response to the q-th question.

Our proposed model is outlined in Figure 1. It first constructs neural features from the data $\{A_q, V_q, T_q \mid q \in \{1, \ldots, Q\}\}$ by extracting embedding vectors using pre-trained neural models. Specifically, Sentence-BERT extracts textual neural features E_q^T from dialogue text data T_q, while VideoLLaMA-2 extracts neural audio-visual features E_q^V from video data V_q and audio data A_q. Additionally, GhostFaceNets are used to extract facial neural features E_q^F from video data V_q, and ECAPA-TDNN (Emphasized Channel Attention, Propagation, and Aggregation in Time Delay Neural Networks) extracts neural speech features E_q^S from speech audio data A_q. For each question q, the neural features E_q^T, E_q^V, E_q^F, and E_q^S are each passed through a modality-specific fully connected layer. The resulting outputs are concatenated and further processed by two additional fully connected layers to form an embedding vector $\boldsymbol{z}_q$, where these layers share parameters across all questions. Subsequently, attention pooling is applied to the sequence of $\boldsymbol{z}_q$ to construct a neural feature for the k-th trait, defined as $\boldsymbol{h}_k = \sum_{q=1}^{Q} softmax\left(\boldsymbol{V}_k \cdot \tanh\left(\boldsymbol{W}_k \cdot z_q + \boldsymbol{b}_k\right)\right) \boldsymbol{z}_q$, where $\boldsymbol{W}_k$, $\boldsymbol{V}_k$, and $\boldsymbol{b}_k$ are trait-specific weights and bias parameters. Next, the handcrafted features H, derived from previous work (Naim et al., 2015), are concatenated with $\boldsymbol{h}_k$ to form a hybrid feature vector $\boldsymbol{r}_k$. Based on the hybrid feature $\boldsymbol{r}_k$, the trait-attention (Ridley et al., 2021) is applied to construct an inter-trait correlation-aware vector $\boldsymbol{x}_k$, defined as $\boldsymbol{x}_k = \sum_{k'=1}^{K} I(k' \neq k)[\exp(\boldsymbol{r}_k \cdot \boldsymbol{r}_{k'})/C]\, \boldsymbol{r}_{k'}$, where $I(\cdot)$ is the indicator function that returns 1 when its argument is true and 0 otherwise, and $C = \sum_{k'=1}^{K} I(k' \neq k) \exp(\boldsymbol{r}_k \cdot \boldsymbol{r}_{k'})$. Finally, the concatenated vector $\tilde{\boldsymbol{x}}_k = [\boldsymbol{x}_k, \boldsymbol{r}_k]$ is fed into a trait-specific two-layer fully connected network with sigmoid activation to produce the normalized scores for the k-th trait. The model is trained by minimizing the mean squared error loss through back-propagation after normalizing the gold-standard scores into the range $[0, 1]$. During inference, predicted scores can be linearly rescaled back to the original score scale.

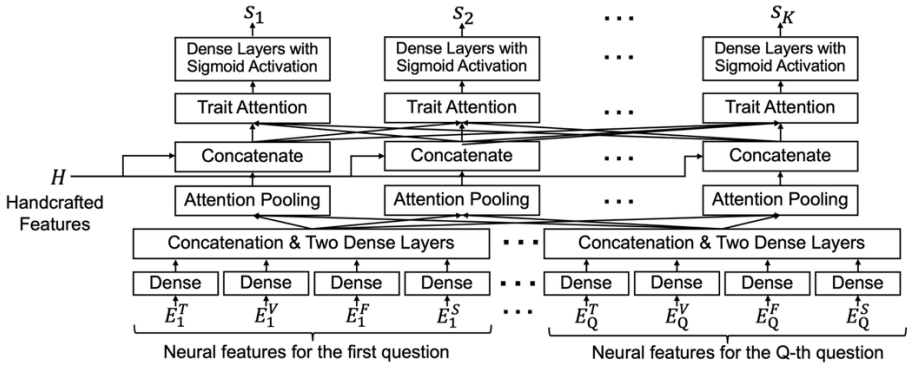


Figure 1. Architecture of the Proposed Model.

3. Evaluation Experiment

To evaluate the effectiveness of the proposed model, we conducted experiments using a real-world dataset comprising 138 interview videos of MIT students from a previous study (Naim et al., 2015). Each interview includes five standardized questions posed to all participants, with responses scored by nine human evaluators across the 16 traits, using a Likert scale ranging from 1 to 7. The average scores from nine raters were used as the gold standard.

Using this dataset, our experiment evaluated score-prediction accuracy, using 10-fold cross-validation, with the correlation coefficient as the evaluation metric. The proposed method was compared against 1) the support vector regression (SVR) and Lasso regression models, which use handcrafted features H as input, as proposed by Naim et al. (2015), 2) variants of the proposed model that exclude one of the neural features (E_q^T, E_q^V, E_q^F, or E_q^S) or the handcrafted features H from the input, and 3) the proposed model without the trait-attention mechanism. We also evaluated the accuracy of individual human raters' scoring as the average correlation between each rater's scores and the gold standard scores.

Table 1 presents the results, where $TrAtt$ denotes the trait-attention. Looking at the averaged performance, the proposed method achieves the highest accuracy. Additionally, a paired t-test between the proposed method and each of the others revealed that the proposed method significantly outperformed all others at the 5% significance level. These results indicate that integrating neural and handcrafted features across all modalities, together with accounting for inter-trait correlations, significantly improves scoring accuracy compared with the comparative models and even individual human raters.

Table 1. Experimental Results

Trait	Proposed model	Naim et al.		Proposed model without						Human rater
		SVR	Lasso	E_q^T	E_q^V	E_q^F	E_q^S	H	$TrAtt$	
1	**.702**	.650	.647	.671	.680	.676	.665	.562	.690	.616
2	**.710**	.652	.675	.643	.653	.671	.674	.540	.673	.617
3	**.739**	.669	.638	.647	.636	.694	.655	.465	.688	.707
4	**.792**	.790	.787	.776	.780	.789	.788	.556	.778	.661
5	.399	**.409**	.311	.406	.324	.323	.342	.352	.360	.600
6	.658	.656	**.710**	.612	.658	.673	.659	.540	.635	.642
7	.705	.731	**.755**	.661	.681	.726	.693	.551	.688	.661
8	**.528**	.473	.373	.491	.409	.502	.489	.404	.490	.355
9	.636	.595	**.739**	.627	.612	.589	.597	.447	.589	.628
10	.520	.501	.494	.537	**.541**	.484	.527	.199	.526	.389
11	**.513**	.477	.488	.469	.474	.487	.488	.332	.484	.444
12	.512	.458	.439	**.575**	.529	.436	.555	.229	.511	.513
13	**.631**	.587	.522	.567	.606	.595	.573	.469	**.631**	.525
14	**.594**	.588	.462	.549	.549	.543	.536	.439	.576	.510
15	.480	.399	.369	.480	.450	.352	**.494**	.300	.381	.494
16	**.636**	.618	.564	.624	.596	.591	.600	.551	.618	.591
Avg.	**.610**	.578	.561	.583	.574	.571	.583	.433	.582	.560

4. Conclusion

This study proposed a novel trait-scoring model for interview examinations and demonstrated its effectiveness. Future directions include extending the model to free-form interviews instead of standardized interview format, conducting comprehensive ablation studies and experiments with larger corpora, discussions of interpretability and endemic bias.

References

Hemamou, L., Guillon, A., Martin, J.-C., & Clavel, C. (2023). Multimodal hierarchical attention neural network: Looking for candidates behaviour which impact recruiter's decision. *IEEE Transactions on Affective Computing*, *14* (2), 969–985.

Naim, I., Tanveer, M. I., Gildea, D., & Hoque, M. E. (2015). Automated prediction and analysis of job interview performance: The role of what you say and how you say it. *Proceedings of 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 1–6.

Ridley, R., He, L., Dai, X., Huang, S., & Chen, J. (2021). Automated cross prompt scoring of essay traits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 13745–13753.

Learning Support System Visualizing Behavior of Student Program in Object Oriented Language Based on Teacher's Intent of Instruction

Koichi YAMASHITA[a*], Kenzo KOBAYASHI[b], Takuma KOKUBO[b], Satoru KOGURE[b], Yasuhiro NOGUCHI[b], Raiya YAMAMOTO[b], Tatsuhiro KONISHI[b] & Yukihiro ITOH[c]
[a]*Faculty of Business Administration, Tokoha University, Japan*
[b]*Faculty of Informatics, Shizuoka University, Japan*
[c]*Shizuoka University, Japan*
*yamasita@hm.tokoha-u.ac.jp

Abstract: This study describes a learning support system developed by extending a system that visualizes program behavior based on teachers' intent of instruction. The extended system supports functions to visualize concepts specific to object-oriented languages and the behavior of learners' program code. Our system was evaluated based on subjective evaluations by learners in practical classes where it was introduced and implemented, and test results from an evaluation experiment conducted with recruited participants. The evaluation results suggested that although the system's impact was not uniform across all learners, it proved highly effective for students with insufficient understanding of the programs.

Keywords: Programming education, program visualization system, educational authoring tool, classroom practice

1. Introduction

Thus far, we developed a program visualization (PV) system called the Teacher's Explaining Design Visualization Tool (TEDViT) for the C language and conducted several classroom practices by introducing it into actual classes (Kogure et al., 2014). A distinctive feature of TEDViT is that it allows teachers to freely customize PV according to their intent of instruction. The evaluation results from the classroom practice suggest that TEDViT effectively supports programming learning for novice learners. We also developed extensions to support the understanding of learning objectives specific to object-oriented languages to deal with the trends in languages used for novice programming education (Yamashita et al., 2022) and extensions to visualize the behaviors of learners' program codes to improve learner engagement (Kobayashi et al., 2024). These extensions were evaluated individually to obtain positive results regarding visualization validity and learning effects.

However, these studies were independent of each other. In this study, we developed an extended TEDViT that incorporates functions to visualize the behaviors of learners' program codes, supporting visualizations of concepts specific to object-oriented languages. This paper describes the extended TEDViT and its evaluation experiments. We evaluated student engagement through questionnaire surveys administered during practical classes that incorporated our system, while learning effectiveness was evaluated based on test results from an experiment conducted with recruited participants.

2. Proposed System

We developed a PV system by extending TEDViT, which supports the visualization of an object-oriented conceptual model (OOCM) (Kogure et al., 2019) for *modified content*. OOCM is a visualization model based on the Unified Modeling Language (UML) that can visualize the relationships among objects and classes, including the concepts of inheritance and polymorphism. *Modified content* is one of the levels of ownership of target programs in PV systems categorized by Sorva et al. (2013), including *given content*, where learners cannot modify the target program at all; *own cases*, where learners can only modify input and variable values; *modified content*, where learners can edit the target program; and *own content*, where learners can observe their program code as the target program. Sorva et al. (2013) pointed out that the closer the content ownership of the target program is to *own content*, the stronger the learner's engagement, and the better the learning effectiveness of the system.

Bringing content ownership closer to *one's own content* is an inherently difficult approach in a PV system that cannot reflect a teacher's intent of instruction. Therefore, no similar attempts have been made to date. The reason for limiting the correspondence to *modified content* was to simplify the process of determining the correspondence between the teacher's intent of instruction and the elements of the learner's program. Although this reduces the coding flexibility of the target program, we believe that this constraint will not be a major problem when introducing the system in programming education for novice learners.

TEDViT displays PV and interaction on a web browser, considering the ease of integration into various classroom environments. This implies that the coding environment for learners and the PV observation environment were provided separately. However, as Brusilovsky et al. (1994) pointed out, in the visualization of *modified content*, PVs are considered as feedback for exploratory learning and problem solving. Therefore, in this study, we developed our system as an extension plugin to Microsoft Visual Studio Code (VSCode). VSCode is a widely used text editor for coding, and its use was recommended for students participating in the classroom practice described in the next section.

3. Evaluation and Conclusion

We conducted experiments to evaluate the effectiveness of the system developed in this study. We set two hypotheses to be verified: our system can improve student engagement, and our system can improve learners' program understanding. For simplicity, we had learners work on exercises using our system in actual classes and conducted questionnaire surveys to investigate whether they demanded the use of our system for learning. To measure the learners' understanding of the programs, we conducted an evaluation experiment in which we recruited participants and used our system to measure their understanding based on tests.

To investigate whether learners demanded the use of our system, we conducted a five-point scale questionnaire survey after students participated in exercises using the system during actual classes. The exercises were conducted in an actual programming class for second-year university students majoring in computer science, where they observed program behavior using the proposed system. The average score of the responses collected from the 63 students was 3.65. Of the 63 students, two responded with 1 or 2, indicating a negative attitude toward using our system, whereas 32 responded with 4 or 5, indicating a positive attitude. These results suggest that more than half of the users were inclined toward continued use of the system, indicating at least some positive impact on engagement.

An experiment to evaluate learners' understanding of the program was conducted with 22 participants, ranging from first-year university students to first-year graduate students, who responded to our call for participants. After a brief explanation of the evaluation experiment, a pre-test was conducted. The participants were divided into two groups, A and B, with 11 participants in each, so that the average pretest scores in each group were approximately the same. Both groups performed the same exercises. However, to reduce the order effect, the order of the exercises was reversed. Group A worked on exercises using our system, took a middle test, and worked on exercises without using our system before taking a post-test. Group B worked on exercises without our system, took a middle test, and then worked on exercises with our system before taking a post-test.

Although the learning order differed between the two groups, the correct answer rates (CARs) showed a similar trend, and the results did not show a clear advantage in learning using our system. Therefore, based on the results of the pre-test, we further divided the two groups into upper, middle, and lower groups according to their scores and investigated the changes in the scores of each group. Here, grouping based on pretest scores was performed by dividing the subjects into the lower group with scores below the first quartile, the upper group with scores above the third quartile, and the middle group with scores between the first and third quartiles. Table 1 shows the changes in CARs for each group. The underlined values represent the CARs before and after the implementation of the exercises using our system.

Table 1. *Average correct answer rates for each group*

	Upper A	Middle A	Lower A	Upper B	Middle B	Lower B
Pre-test	.784	.580	.417	.834	.570	.384
Middle-test	.884	.860	.750	.900	.740	.634
Post-test	.987	.950	.884	.934	.820	.850

According to the values in Table 1, the CAR for Group A, which used our system between the pretest and middle test, increased by .334, while that of Group B, which did not use our system, increased by .250. Between the middle and post-tests, the CAR for Group B, which used our system, increased by .217, while that for Group A, which did not use our system, increased by .134. Because both groups showed an increase in the CARs when using our system, these results suggest that our system is effective in supporting learners with insufficient understanding of programming concepts. When combined with the results from the subjective evaluations based on questionnaire surveys, the results suggest that our system improves student engagement and program understanding to a certain extent. Although the same trend was not observed in the upper and middle groups of the pretest CARs, we conclude that the two experimental hypotheses were supported to a certain degree overall, even if not fully.

Acknowledgements

This study was supported by JSPS KAKENHI Grant Numbers JP19K12259, JP22K12290 and JP24K15219.

References

Brusilovsky, P., Kouchnirenko, A., Miller, P., & Tomek, I. (1994). Teaching Programming to Novices: A Review of Approaches and Tools. Proceedings of World Conference on Educational Multimedia and Hypermedia (ED-MEDIA94), 103-110.

Kobayashi, K., Kogure, S., Noguchi, Y., Yamamoto, R., Yamashita, K., Konishi, T., & Itoh, Y. (2024). Program Learning Support System with Visualization Reflecting Teacher's Intent for Learner's Code. Proceedings of the 32nd International Conference on Computers in Education, 417-419.

Kogure, S., Fujioka, R., Noguchi, Y., Yamashita, K., Konishi, T., & Itoh, Y. (2014). Code reading environment according to visualizing both variable's memory image and target world's status. Proceeding of the 22nd International Conference on Computers in Education, 343-348.

Kogure, S., Ogasawara, K., Yamashita, K., Noguchi, Y., Konishi, T., & Itoh, Y. (2019). Application of Programming Learning Support System to Object-Oriented Language. Proceedings of the 26th International Conference on Computers in Education, 348-350.

Sorva, J., Karavirta, V., & Malmi, L. (2013). A Review of Generic Program Visualization Systems for Introductory Programming Education. ACM Transactions on Computing Education (TOCE), 13(4), 15.

Yamashita, K., Suzuki, Y., Kogure, S., Noguchi, Y., Yamamoto, R., Konishi, T., & Itoh, Y. (2022). Learning Support System Visualizing Relationships Among Classes and Objects Based on Teacher's Intent of Instruction. Proceedings of the 30th International Conference on Computers in Education, 314-316.

Development of a Coding Framework for Cognitive Engagement Based on Online Asynchronous Discussion

JunHao WAN, Guang CHEN*
Faculty of Education, Beijing Normal University, China
*guang@bnu.edu.cn

Abstract: Online learning has broken the constraints of time and space, enhancing the quality and accessibility of higher education. However, real-time monitoring and effective assessment of learners' cognitive engagement within online learning environments present a major challenge. Due to variations across different learning contexts, general cognitive engagement coding frameworks increasingly reveal limitations in applicability, underscoring the need to develop context-specific frameworks. Therefore, this study adopts a grounded theory approach, analyzing 180 asynchronous discussion entries from six courses on the Chinese University MOOC. Through bottom-up coding, a three-level indicator system was established. Subsequently, the Delphi method was employed for validation, and the Analytic Hierarchy Process (AHP) was applied to determine indicator weights. The study also found that non-cognitive factors such as emotions and behaviors influence cognitive engagement, while the impact of interaction is somewhat diminished in this context.

Keywords: Online learning, asynchronous discussion, cognitive engagement, coding framework

1. Introduction

With the widespread adoption of asynchronous discussion in online learning and Massive Open Online Courses (MOOCs), the effective and scalable assessment of learners' cognitive engagement has become crucial (Peng et al., 2020). Cognitive engagement refers to a learner's willingness to make an effort to understand content and master skills (Huang et al., 2019). Asynchronous discussion forums serve as the primary space for interaction between learners and instructors (Almatrafi & Johri, 2019), and high-quality discussions have been shown to enhance learners' engagement and academic performance (Galikyan et al., 2021). However, existing cognitive engagement coding frameworks, such as the classic ICAP framework (Chi & Wyile, 2014), are often constructed in a top-down manner (Xu et al., 2023; Rivera et al., 2024). This approach leads to limitations such as coarse granularity (Liu, et al., 2022) and a lack of data-driven objectivity when applied to automated analysis, thereby restricting their practicality and accuracy. These frameworks demonstrate insufficient capability in distinguishing complex cognitive behaviors and struggle to adapt to large-scale, diverse online learning data.

To address this gap, this study aims to develop and validate a novel coding framework for cognitive engagement applicable to online asynchronous discussion contexts. Through a bottom-up grounded theory analysis of authentic MOOC discussion data (Strauss & Corbin, 1990), followed by validation via the Delphi method and weighting using the Analytic Hierarchy Process (AHP) to develop a more fine-grained and context-specific framework. This framework provides a more solid and objective benchmark for future automated cognitive engagement assessment systems, enabling more precise understanding of and support for learners' cognitive states.

2. Methods

This study analyzed 180 authentic asynchronous discussion posts from six diverse courses (covering humanities, social sciences, and STEM) on the Chinese University MOOC platform. A three-stage mixed-methods approach was employed. Firstly, grounded theory was applied, using three-level coding to generate the initial framework. This coding process was validated with substantial inter-coder reliability (Cohen's Kappa = 0.71) and continued until theoretical saturation. Next, two rounds of the Delphi method were conducted, inviting six experts in educational technology to anonymously revise and validate the framework until consensus was achieved. Then, based on expert judgments, the AHP was used to calculate the relative weights of all indicators, ensuring the framework possesses a robust internal structure.

3. The Coding Framework for Cognitive Engagement

The final output of this research is a multi-level coding framework for cognitive engagement. As shown in Table 1, the framework consists of 2 primary, 6 secondary, and 17 tertiary indicators, each with an assigned weight reflecting its relative importance. This fine-grained, multi-level structure enables a comprehensive and differentiated assessment of learners' cognitive engagement in asynchronous discussion, ranging from passive participation to deep knowledge construction.

Table 1. *Cognitive Engagement Coding Framework with Weights*

Primary indicators	Secondary indicators	Tertiary indicators	Description of the tertiary indicators
Multidimensional Orientation (0.326)	Emotional Express (0.130)	Grateful (0.033)	Express gratitude towards teachers and other learners
		Joyful (0.020)	Express joy that arises during the learning process.
		Hopeful (0.026)	Express hopes or aspirations regarding certain expectations.
		Moved (0.052)	A deep understanding and resonance with the statements or emotions of other learners.
	Information Consultation (0.196)	Transactional Consultation (0.078)	Questions regarding the course, system functionality, stability, compatibility, etc.
		Learning-related Consultation (0.117)	Questions regarding the accessibility and availability of learning materials.
Cognitive Orientation (0.674)	Passive (0.060)	Repetition (0.030)	The content of the comments is largely or entirely derived from others.
		Meaning-lessness (0.030)	The content of the comments is devoid of value.
	Active (0.129)	Declarative Statement (0.042)	Learners' direct responses or clear statements to questions or scenarios set by the instructor.
		Autonomous Expression (0.087)	Learners' ability to demonstrate independent thinking and autonomy during the learning process.
	Interactive (0.207)	Response (0.049)	Learners provide explanations or answers to questions posed by other learners.
		Reflection (0.060)	Learners adopt a critical stance toward the views of teachers or other learners.
		Support (0.029)	Learners understand or accept the viewpoints of other learners.
		Extension (0.070)	Building upon others' viewpoints or ideas to expand and elaborate, thereby enriching existing perspectives or proposing new insights or reflections.
	Constructive (0.278)	Transfer (0.071)	Learners apply existing knowledge or concepts to new contexts or fields.
		Application (0.094)	Learners apply newly acquired knowledge, theories, or concepts to real-world situations to support their viewpoints or explain phenomena.
		Output (0.113)	Learners propose original viewpoints, theories, or strategies in discussion.

4. Discussion and Conclusion

A key finding, emerging directly from our bottom-up, data-driven approach, is the significant role of “Multidimensional Orientation” (weight 0.326). Unlike traditional top-down models that often focus exclusively on purely cognitive indicators, our grounded framework reveals that non-cognitive factors, such as “Emotional Express” and “Information Consultation”, are integral components of engagement in this online asynchronous context. This finding challenges a narrowly cognitive view and underscores the necessity of a more holistic assessment model.

Notably, within the core “Cognitive Orientation” dimension, “Constructive” activities (0.278) outweigh “Interactive” activities (0.207). This finding offers a significant, context-specific amendment to influential frameworks like ICAP, which often place a heavy emphasis on interaction. It suggests that in the asynchronous, time-displaced environment of MOOC forums, fostering individual knowledge construction (e.g., application, original output) may be a more critical pedagogical lever for enhancing deep cognitive engagement than simply promoting more interaction.

Future research will focus on validating this framework across larger, more diverse datasets to address potential sample bias and enhance its generalizability across various disciplines and learning contexts. This will ultimately provide a robust foundation for developing large-scale, automated cognitive engagement assessment tools through integration with natural language processing (NLP) models.

References

Almatrafi, O., & Johri, A. (2019). Systematic review of discussion forums in massive open online courses (MOOCs). *IEEE Transactions on Learning Technologies*, 12(3), 413–428.

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4), 219–243.

Galikyan, I., Admiraal, W., & Kester, L. (2021). MOOC discussion forums: The interplay of the cognitive and the social. *Computers & Education*, 104133.

Huang, B., Hew, K. F., & Lo, C. K. (2019). Investigating the effects of gamification-enhanced flipped learning on undergraduate students’ behavioral and cognitive engagement. *Interactive Learning Environments*, *27*(8), 1106–1126.

Liu, S., Liu, S., Liu, Z., Peng, X., & Yang, Z. (2022). Automated detection of emotional and cognitive engagement in MOOC discussions to predict learning achievement. *Computers & Education*, 181, 104461.

Peng, X., Han, C., Ouyang, F., & Liu, Z. (2020). Topic tracking model for analyzing student-generated posts in SPOC discussion forums. *Smart Learning Environments*, 17(1).

Rivera, D. A., Frenay, M., & Paquot, M. (2024). The role of MOOC forum discussion tasks in learners' cognitive engagement. *Journal of Computer Assisted Learning*, 40(5), 2103–2120.

Strauss, A., & Corbin, J. (1990). Basics of qualitative research: Grounded theory procedures and techniques. *Sage Publications*.

Xu, Q., Wei, Y., Gao, J., Yao, H., & Liu, Q. (2023). ICAPD framework and simAM-YOLOv8n for student cognitive engagement detection in classroom. *IEEE Access,* 11, 136063–136076.

C4
Technology Enhanced Learning for Mobility of Learners and Learning Experiences (TEML)

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

A Cross-Platform Engagement Analysis for Supporting Knowledge Transformation Through Complex Learning Settings

Manabu ISHIHARA[a*], Izumi HORIKOSHI[b] & Hiroaki OGATA[c]
[a]*Graduate School of Informatics, Kyoto University, Japan*
[b]*Uchida Yoko Institute for Education Research, Uchida Yoko Co., Ltd., Japan*
[c]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
*ishihara.manabu.46z@st.kyoto-u.ac.jp

Abstract: To promote the transformation of conceptual knowledge into practical knowledge, educators would incorporate real-world situations as a learning setting. Yet, it is uncertain whether learners who actively acquired knowledge in the classroom will also be engaged in applying knowledge in the real world. Educators have scarce clues to understand learners' status and provide adequate support in such learning processes. Utilizing digital tools, especially cross-platform approaches, has the potential to address this issue by collecting and analyzing activity logs of various tools used in each learning setting. Nonetheless, prior studies rarely examined the continuity of engagement across complex learning settings involving the real world. This study aims to fill the gap by analyzing a dataset from a university course in Japan. We investigated the activity logs from a material reader for knowledge acquisition in the classroom and those from a mobile app dedicated to knowledge application in the real world. As a result, we found that engagement in the knowledge acquisition activity was not promising engagement in the knowledge application activity. We then identified four types of engagement across the complex learning settings. Notably, the types that were relatively less engaged in the knowledge application activity showed lower understanding levels, despite a certain engagement in the knowledge acquisition activity. The findings emphasize the importance of scaffolding during activities in the real world. Further, we discuss the potential of cross-platform analytics toward more effective knowledge transformation that leverages complex learning settings.

Keywords: Engagement, cross-platform analytics, learning analytics, knowledge transformation

1. Introduction

One of the educational goals is transforming conceptual knowledge into practical knowledge. In higher education, educators may include hands-on practices in addition to lectures in light of pedagogical strategies such as the experiential learning cycle (Kolb, 1984) and learning by doing (Gibbs, 1988). In particular, real-world situations can be a suitable learning setting for mitigating theory-practice gaps (Morley & Jamil, 2021). However, it is challenging to tailor a smooth transition from knowledge acquisition to applications with adequate understanding levels. In particular, understanding learners' engagement beyond the classroom settings is difficult, whereas educators have essential roles in maintaining learning ecologies across formal and informal circumstances (Peters & Romero, 2019).

Harnessing digital tools can enhance a wide range of learning experiences, including lectures in the classroom, online courses, and even field activities. Moreover, learning analytics have the potential to support learners by tracing their progress in complex learning settings (Moon et al., 2023). It leads to the assumption that analyzing learners' digital activities in various situations allows us to understand learners' progress of knowledge acquisition and application. Further, such technology-enhanced approaches could contribute to smooth

knowledge transformation through the classroom and real-world settings. Nonetheless, prior analyses have tended to focus on one-sided analysis (Mangaroska et al., 2021).

This study thus examines engagement across knowledge acquisition in the classroom and knowledge application in real-world settings with the following research questions:
RQ1: How are engagement in knowledge acquisition and understanding levels related to engagement in knowledge application?
RQ2: What kinds of engagement types emerge across complex learning settings?
RQ3: How do understanding levels differ among the engagement types?

2. Related Works

Real-world situations can be a suitable option to promote knowledge transformation with memorable events (Morley & Jamil, 2021). For the implementation, learning designs tend to be complex because of the mixture of preparatory and reflective activities in the classroom that amplify the effectiveness of field experiences (Lee et al., 2020).

Prior studies of ubiquitous learning have proposed mobile apps to support knowledge application in real-world settings by connecting curricular topics covered in the classroom. For instance, language learning in life space (Yang & Song, 2023), cultural heritage learning (Ruiz-Calleja et al., 2023), and science education at a museum (Hsu et al., 2016). However, the solutions relying on a single digital tool may not be scalable. Educators would employ different digital tools suitable for specific learning settings. Thus, there will be a need for other approaches that can continuously support across complex learning settings.

Several studies have proposed cross-platform analytics. Such approaches enable us to collect and analyze learner behaviors derived from different digital tools. A study of programming education gathered data from discrete tools for online lectures and coding practice to analyze the important features of programming performance (Mangaroska et al., 2021). Regarding the knowledge application in real-world situations, a seamless learning environment that accumulates data from modulated tools, including a mobile app, was introduced (Flanagan & Ogata, 2018). It can preserve activity logs from a material reader for classroom learning and also from a mobile app that stores the records of field-based activities. A study employed this environment for situated language learning. Through cross-platform analytics, the researchers demonstrated visual feedback depicting the connections between the vocabulary taught in class and actual instances found in life space (Mouri et al., 2018). Another study detected relevance between the learners' text outputs about real-world activities and the course content in a lecture series (Ishihara et al., 2024). As such, cross-platform analytics can open ways to understand continuous mobility of learning across various contexts and spaces.

Yet, none of these studies fully explored the associations of learners' engagement in complex learning settings. Real-world experiences cannot be replayed. Therefore, letting learners be less engaged may result in limited achievement. In the domain of technology-enhanced teaching and learning, studies advocate that activity logs in digital tools proximate engagement (Ahmadi et al., 2023; Motz et al., 2019). Such logs derived from multiple digital tools provide educators with clues to improve learning designs equipped with them (Horikoshi et al., 2025). Thus, it is valuable to investigate learners' distributed engagement through cross-platform analytics.

3. Method

3.1 Context of Data Collection

We collected a dataset from a university course in Japan. The subject of the course was human interface, and the main language was Japanese. This weekly course started in October 2024 and ended in January 2025. Five teachers took turns lecturing, and 39 students registered for the course. From the first to the fourth week, the students tackled a weekly

assignment that required them to find and analyze actual human interfaces in life space. Then, they wrote an analysis report about the interfaces they found by referring to the previous course content. In the subsequent class, they took five multiple-choice quizzes as an understanding check activity, of which scores ranged from zero to 10. After that, they discussed each one's analysis among peers then moved on to the next lecture to acquire new knowledge of course content (Figure 1 a).

The students' engagement in each activity and achievement can be framed in the revised Bloom's taxonomy (Krathwohl, 2002). It helps us categorize various learning activities into the six levels of learning achievement, from simple to complex: remember, understand, apply, analyze, evaluate, and create. Engagement in the lectures, which expected the acquisition of basic knowledge of human interface, corresponds to the achievement of remembering in the taxonomy. Scores of the quizzes represent the achievement of understanding. Engagement in real-world tasks that expected knowledge application corresponds to the achievement of applying. The quality of the analysis reports can indicate the achievement of analyzing. Engagement in peer discussions that expected mutual evaluations of each one's analysis report can be regarded as the achievement of evaluating.

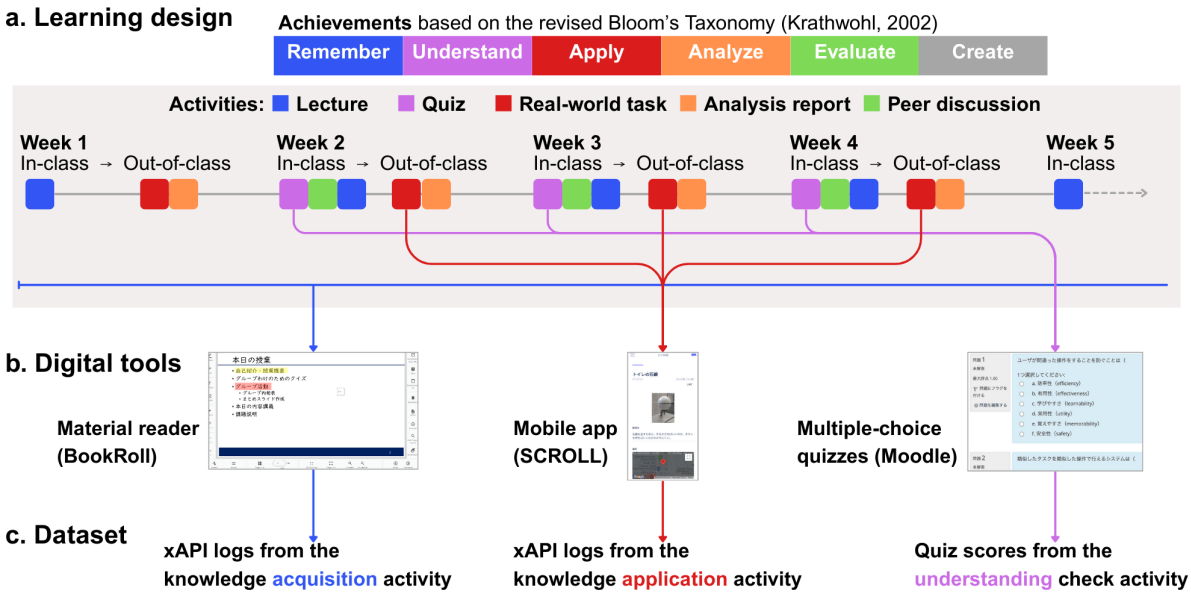


Figure 1. Overview of learning design, tools, and dataset.

3.2 Digital Tools

To investigate the students' engagement in complex learning settings, we employed a seamless learning platform (Flanagan & Ogata, 2018), which enables to implement technological assistance in a wide range of learning contexts. It can flexibly accommodate modular digital learning tools, for instance, a learning management system called Moodle, a material reader for classroom learning called BookRoll, and a mobile app for field-based activities called SCROLL (Figure 1 b). After logging in to Moodle, users can access such tools without additional login. During the course, Moodle functioned as a hub for accessing BookRoll, SCROLL, assignment submission pages, and quiz pages. Whenever the students use BookRoll or SCROLL, activity logs are accumulated in a database.

The material reader called BookRoll was used throughout the course period as a viewer of the lecture materials tailored by the course teachers. Operations on the material reader, such as browsing pages and leaving memos, represent the behaviors of knowledge acquisition from lectures. At the end of each class, the students were encouraged to leave memos and/or highlights on BookRoll. It expected students to make lasting impressions of some course content in the lecture prior to a real-world task.

The mobile app called SCROLL, accessible from smartphones and any other connected devices, was introduced for real-world tasks. The students used the app to record

photos of actual interfaces and notes about their situational findings. Such a saving behavior substantiated knowledge application because the students were supposed to relate certain course content when they recorded interfaces. Afterward, they reviewed their records on the app as a process of writing an analysis report. Additionally, they could browse peers' records on the app. During the out-of-class periods, there was no reminder nor encouragement for the use of the app.

The structure of log data from the tools above is aligned with the Experience Application Programming Interface (xAPI), a standardized format for e-learning activities. Depending on the operation, xAPI conveys context information about a specific operation, an actor of the operation, and an object of the operation. Such a predefined data structure eases interpreting user operations into learning behaviors. Standardized data is also beneficial for integrating activity logs from different tools, so that the data can be analyzed seamlessly, considering in-class and out-of-class contexts.

3.3 Dataset

We were allowed to access data from the first to the fifth week in the course. The activity logs of the digital tools and the average quiz scores were gathered from 26 students who granted the research use of their data. The frequency of the activity logs from the material reader was considered engagement in the knowledge acquisition activities. Average quiz scores represented the understanding levels of the acquired knowledge of previous lectures. The frequency of the activity logs from the mobile app was regarded as engagement in the knowledge application activities (Figure 1 c).

We categorized the activity logs per tool. On the material reader side, the operations related to browsing a page, searching in lecture material, and clicking a link in a page to access an external information were classified in the *View* category. The operations of outputting highlights, memos, the "favorite" emotions, and bookmarks on a lecture material page were considered the *Save* category. The category also included the operations of editing of memos.

On the mobile app side, the operation of creating a new record was regarded as the *Save* category. The operation of browsing one's own records was considered the *View Own* category. In addition, visiting peers' records was classified as the *View Other* category. Although there were other operations, such as adding "like" emotions and leaving comments on recorded items, these were not taken into account since the instructions did not expect such behaviors. A summary of the dataset is shown in Table 1.

Table 1. *Dataset (n=26)*

Data source	Category	Total	M	SD
Logs from material reader for knowledge acquisition	Save	224	8.62	10.32
	View	35,507	1,365.65	639.64
Logs from mobile app for knowledge application	Save	84	3.23	1.53
	View Own	152	5.85	5.65
	View Other	473	18.19	31.21
Quiz scores (0-10) for understanding check	Average score	26	8.10	1.33

M: Mean, SD: Standard deviation

3.4 Data Analysis

To answer RQ1 concerning the associations of engagement in the knowledge acquisition with other factors, we measured the correlations between the activity logs and the average quiz scores. we standardized each category's logs and then examined its normality of distribution using the Shapiro-Wilk test. Since the logs did not follow the normal distribution, we measured Spearman's rank correlation.

Regarding RQ2, which addresses the engagement types across the classroom and real-world settings, we performed an unsupervised clustering based on the activity logs of the knowledge acquisition and application activities. To identify the best clustering result, we conducted the following procedure: dimensional reduction for standardized activity logs, clustering, and silhouette analysis. For dimensional reduction, we used t-distributed Stochastic Neighbor Embedding (t-SNE), which is effective for non-linearly distributed data (Dhalmahapatra et al., 2019). We then conducted k-means clustering with an optimal number of clusters determined by the elbow method. The quality of the clustering results was assessed based on silhouette analysis. As shown in Figure 2, we identified the optimal clustering condition: dimensionality reduction using t-SNE with a perplexity of 2 and k-means clustering with the number of clusters set to 4. The four clusters were compared through box plots to interpret their characteristics of engagement.

For RQ 3, we tested whether there is any statistical difference in the understanding levels of course content among the clusters. As the sample size in each cluster was not sufficient for parametric measures, we performed the Kruskal-Wallis test, a nonparametric measure diagnosing statistical differences among two or more groups. For the post-hoc test that identifies specific pairs holding the differences, we used the Dunn-Bonferroni test.

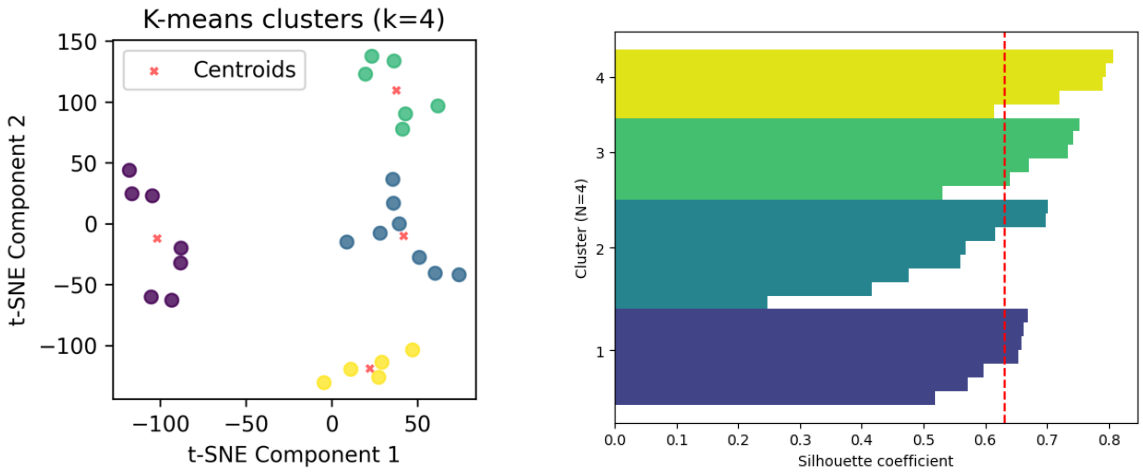


Figure 2. Optimal k-means clustering and its silhouette coefficients

4. Results

4.1 Correlations of Engagement for RQ1

The results of Spearman's rank correlation indicated several statistically significant positive correlations among engagement in the knowledge acquisition activity, average quiz scores as understanding levels of acquired knowledge, and engagement in the knowledge application activity (Table 2). For the association between the knowledge acquisition and application activities, there was a moderate level of statistically significant positive correlation: *View* of the material reader in the knowledge acquisition activity and *View Own* of the mobile app in the knowledge application activity ($\rho(24) = .47$, $p = .015$).

Regarding the associations with understanding levels, the average quiz scores showed moderate levels of statistically significant positive correlations with all categories under the knowledge application activity: *Save* ($\rho(24) = .46$, $p = .02$), *View Own* ($\rho(24) = .61$, $p = .001$), and *View Other* ($\rho(24) = .54$, $p = .004$). In contrast, there was no significant correlations between the categories under the knowledge acquisition activity. The rest of the statistically significant positive correlations were found within the knowledge application activity: *Save* and *View Own* ($\rho(24) = .42$, $p = .035$) and *View Own* and *View Other* ($\rho(24) = .64$, $p < .001$).

Table 2. *Spearman's Rank Correlations Among Categories (n=26)*

Data source	Category	1	2	3	4	5	6
Logs from material reader for knowledge acquisition	1. Save	-					
	2. View		-				
Logs from mobile app for knowledge application	3. Save			-			
	4. View Own		.47*	.42*	-		
	5. View Other				.64**	-	
Quiz scores for understanding check	6. Average score			.46*	.61**	.54**	-

*p < .05, **p < .01

4.2 Types of Engagement for RQ2

The clustering addressed RQ2 that focuses on engagement types across the classroom and real-world learning settings. The results identified four clusters. The box plots depicted each cluster's characteristics (Figure 3). The students in Cluster 1 showed remarkable active usage of mobile app in the knowledge application activity, whereas their *Save* behaviors on the material reader in the knowledge acquisition activity was under average. For those in Cluster 2, all their engagement was under average. Cluster 3 holds students who notably showed the *Save* behaviors on the material reader in the knowledge acquisition activity, whereas other behaviors were near average. Students in Cluster 4 showed active *View* behaviors on the material reader in the knowledge acquisition activity, while their knowledge application activity using the mobile app was under average. The results indicate that the students' engagement differed depending on the learning setting and the operational behavior.

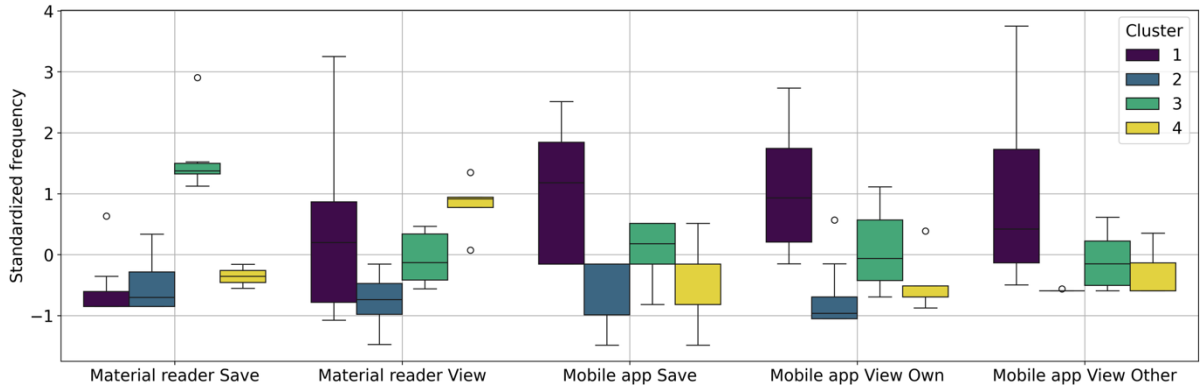


Figure 3. Standardized frequency of activity logs per cluster.

4.3 Difference in Understanding Level Among Clusters for RQ3

For RQ3, the statistical tests revealed the difference among the clusters regarding average quiz scores. The descriptive statistics of average quiz scores per cluster are shown in Table 3. A Kruskal-Wallis test indicated a statistically significant difference among the four clusters ($H(3) = 8.99$, $p = .03$). Then the post hoc Dunn-Bonferroni test identified one significant statistical difference between Clusters 1 and 2 (adjusted $p = .04$), as shown in Figure 4.

The descriptive statistics indicate that Cluster 1, highly engaged in the knowledge application activity, marked the highest average quiz scores. Cluster 3, near average overall but actively saved during the knowledge acquisition activity, followed Cluster 1. Cluster 4, less engaged in most activities except frequent viewing during the knowledge acquisition activity, resulted in third place. Cluster 2, relatively low engagement in all activities, marked the lowest scores.

Table 3. *Descriptive Statistics of Average Quiz Scores per Cluster*

Cluster	Number of students	M	SD	Median
Cluster 1 (Application-engaged)	7	9.14	0.47	9.33
Cluster 2 (Less-engaged-overall)	8	7.17	1.51	6.83
Cluster 3 (Acquisition-save-engaged)	6	8.56	0.89	8.83
Cluster 4 (Acquisition-view-engaged)	5	7.60	1.23	7.33

M: Mean, SD: Standard deviation

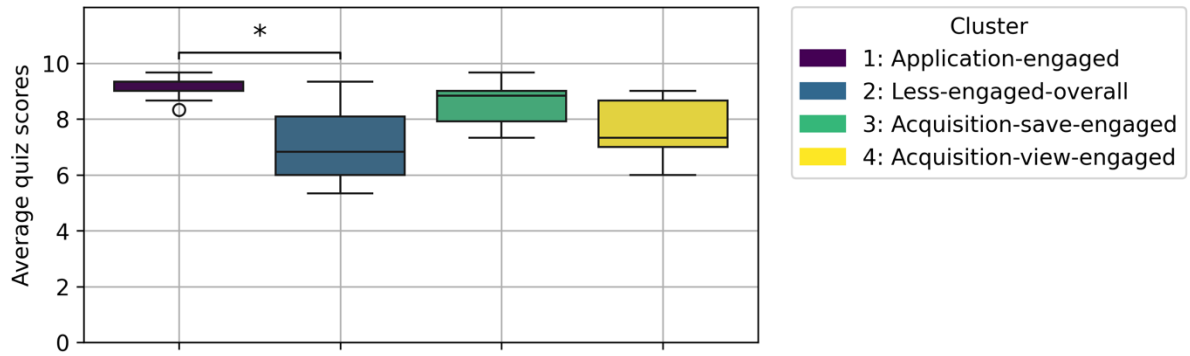


Figure 4. Average quiz scores per cluster.

5. Discussion

5.1 Key Findings

The correlation analysis for RQ1 suggests that engagement in the knowledge acquisition activity in the classroom setting was not promising engagement in the knowledge application activity in real-world setting. We observed only the category of *View* under the knowledge acquisition activity showed a significant positive correlation with the category of *View Own* under the knowledge application activity. Engagement in the knowledge acquisition did not show any correlation with the performance of quizzes in the classroom setting. In contrast, the students engaged in the knowledge application activity tended to build better understandings, even though their engagement in the knowledge acquisition activity was not always high. These findings suggest that engagement in the knowledge acquisition activity has limited associations with higher-order achievement. Rather, the results emphasize the importance of engagement in knowledge application in different situations, which reinforces understanding of knowledge with multiple aspects (Spiro et al. 2013).

Regarding RQ2 and RQ3, we identified four engagement types across the classroom and real-world learning settings. Cluster 1, highly engaged in the knowledge application activity, achieved the best average quiz scores. Although their engagement in the category of *Save* in the knowledge acquisition activity was below average, it might be not essential for developing profound understanding. Their average quiz scores tended to be better than Claster 3 that showed the highest engagement in the category of *Save* in the knowledge acquisition activity. Clusters 2 and 4 showed relatively lower engagement among the clusters. Also, their understanding levels were lower than Clusters 1 and 3. In particular, the difference in the scores between Clusters 1 and 2 was statistically significant. Although no cluster fell into the at-risk level of understanding, the students in Clusters 2 and 4 might need an enhancement for more active knowledge applications, which could lead to deeper understandings of course content.

Notably, the variance of the average quiz scores tended to be narrower when engagement in the knowledge application activity was high. The standard deviation of Cluster 1 was 0.47. That of Cluster 3 was 0.89. Clusters 4 and 2 followed them with their scores of 1.23 and 1.51, respectively. The tendency implies that it was quite promising that the students engaged in the knowledge application activity achieved higher understanding levels.

5.2 Implications

Our cross-platform engagement analysis revealed that engagement can vary depending on the context (knowledge acquisition in the classroom or application in the real world) and the behavior (*View* or *Save*). That means, high engagement in a learning context does not promise continuous engagement in another context. Additionally, the results indicated that the students in the clusters with lower engagement in the knowledge application activity showed relatively lower understanding levels. These findings advocate the need for learning analytics that support wider learning experiences (Ferguson et al., 2019).

Cross-platform analytics could satisfy the need. One potential is to ease the difficulty of evaluating learner achievement while keeping contextual and spatial mobility. In this study, for instance, the students' engagement and understanding levels can be mapped to the revised Bloom's taxonomy by cluster. As shown in Figure 5, the students in Cluster 1 can be assessed as a well-accomplished group in terms of active knowledge application with high understanding levels. In contrast, the students in Cluster 2 may be subject to scaffolding due to lower engagement and understanding levels.

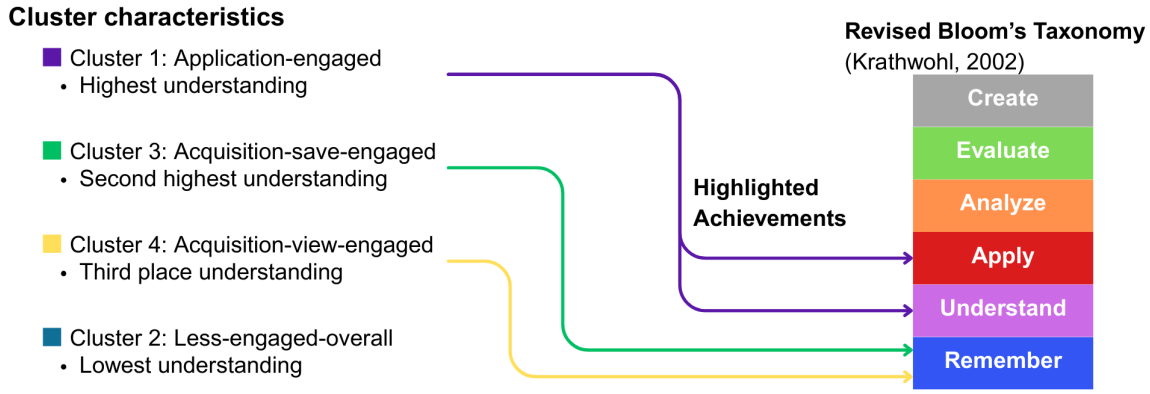


Figure 5. Mapping of highlighted achievements of clusters to the revised Bloom's taxonomy.

Given such inconsistent learner engagement, ensuring learners' continuous engagement is vital. This entails the need for further research. In this study, the material reader BookRoll and the mobile app SCROLL enabled the students to interact with learning resources in the classroom and also their findings in the real world. However, there was a lack of timely feedback. Feedback based on cross-platform analytics could enhance contextual and temporal reflection. For educators, visualizing the transition of learner engagement in tool usage will be helpful in terms of reviewing and improving activity designs (Horikoshi et al., 2025). For learners, feedback based on one's own engagement trajectories can be used for self-monitoring. Frequent self-monitoring is a key to establishing self-regulated learning and higher learning performances (Yang & Song, 2023). In light of such potential, future studies of cross-platform analytics should consider feedback systems to sustain the continuity of engagement across heterogeneous learning settings.

5.3 Limitations

Despite the findings and the potential of cross-platform analytics, this study holds limitations. Our analysis focused on viewing and saving activity logs as the basic engagement underlying

the two digital tools. It might overlook other essential logs related to engagement. In addition, the dataset was collected from a limited part of a course. Other activities, such as analysis report assignments and peer discussions, were not considered. Analyzing the content of such outcomes may offer clues to infer the students' higher-order achievement, i.e., analyzing, evaluating, and creating in the revised Bloom's taxonomy.

Also, our analysis did not consider the temporal and social perspectives. The data was collected from a weekly course, and there were interactions among peers. Further analysis should address the students' temporal changes of engagement and the effects of social interactions.

Another main limitation is the lack of generalizability of the findings. We performed the non-parametric statistical tests because of the small sample size of which data was not normally distributed. Also, the degree of each student's prior knowledge and motivation of the course content may influence the behaviors. Such factors should be controlled in future analyses. For more rigorous discussions, additional case studies using larger datasets are needed.

6. Conclusion

This study demonstrated cross-platform analytics to address the difficulty of understanding learners' engagement across knowledge acquisition in the classroom and application in the real world. The results revealed that engagement appeared inconsistently. Moreover, the student groups that had less engagement in the knowledge application activity showed lower understanding levels. The trackability of cross-platform analytics can support educators and learners in leveraging contextual, spatial, and temporal mobility. Despite current limitations, further studies could exploit its potential for a more effective knowledge transformation through complex learning settings.

Acknowledgements

This work was partly supported by CSTI SIP Program Grant Number JPJ012347 and JSPS KAKENHI Grant Number 23H00505.

References

Ahmadi, G., Mohammadi, A., Asadzandi, S., Shah, M., & Mojtahedzadeh, R. (2023). What are the indicators of Student Engagement in Learning Management Systems? A Systematized Review of the Literature. *The International Review of Research in Open and Distributed Learning*, *24*(1), 117–136. https://doi.org/10.19173/irrodl.v24i1.6453

Dhalmahapatra, K., Shingade, R., Mahajan, H., Verma, A., & Maiti, J. (2019). Decision support system for safety improvement: An approach using multiple correspondence analysis, t-SNE algorithm and K-means clustering. *Computers & Industrial Engineering*, *128*, 277–289. https://doi.org/10.1016/j.cie.2018.12.044

Ferguson, R., Clow, D., Griffiths, D., & Brasher, A. (2019). Moving Forward with Learning Analytics: Expert Views. *Journal of Learning Analytics*, *6*(3), 43–59. https://doi.org/10.18608/jla.2019.63.8

Flanagan, B., & Ogata, H. (2018). Learning analytics platform in higher education in Japan. *Knowledge Management & E-Learning*, *10*(4), 469–484. https://www.kmel-journal.org/ojs/index.php/online-publication/article/view/40/41

Gibbs, G. (1988). *Learning by Doing: A Guide to Teaching and Learning Methods*. Oxford Polytechnic.

Horikoshi, I., Toyokawa, Y., Nakamura, K., Majumdar, R., & Ogata, H. (2025). Learning stream plot: classroom activity visualization with daily learning log data. *Smart Learning Environments*, *12*(1), 1–30. https://doi.org/10.1186/s40561-025-00380-7

Hsu, T.-Y., Kuo, F.-R., Liang, H.-Y., & Lee, M.-F. (2016). A curriculum-based virtual and physical mobile learning model for elementary schools in museums. *The Electronic Library, 34*(6), 997–1012. https://doi.org/10.1108/EL-08-2015-0146

Ishihara, M., Horikoshi, I., & Ogata, H. (2024, November). Linking Real-World Experiences with Course Contents: A Text Mining Approach Toward Effective “There and Back Again.” In A. Kashihara, B. Jiang, M. M. Rodrigo, & J. O. Sugay (Eds.), *Proceedings of 32nd International Conference on Computers in Education Conference Volume I* (pp. 450–459). Asia-Pacific Society for Computers in Education. https://doi.org/10.58459/icce.2024.4881

Kolb, D. A. (1984). *Experiential Learning: Experience As the Source of Learning and Development* (Vol. 1). Prentice-Hall.

Krathwohl, D. R. (2002). A revision of bloom’s taxonomy: An overview. *Theory into Practice*, *41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Lee, H., Stern, M. J., & Powell, R. B. (2020). Do pre-visit preparation and post-visit activities improve student outcomes on field trips? *Environmental Education Research*, *26*(7), 989–1007. https://doi.org/10.1080/13504622.2020.1765991

Mangaroska, K., Vesin, B., Kostakos, V., Brusilovsky, P., & Giannakos, M. N. (2021). Architecting analytics across multiple E-learning systems to enhance learning design. *IEEE Transactions on Learning Technologies*, *14*(2), 173–188. https://doi.org/10.1109/tlt.2021.3072159

Moon, J., Lee, D., Choi, G. W., Seo, J., Do, J., & Lim, T. (2023). Learning analytics in seamless learning environments: a systematic review. *Interactive Learning Environments*, *32*(7) 3208–3225. https://doi.org/10.1080/10494820.2023.2170422

Morley, D. A., & Jamil, M. G. (2021). Introduction: Real world learning—recalibrating the higher education response towards application to lifelong learning and diverse career paths. In *Applied Pedagogies for Higher Education* (pp. 1–17). Springer International Publishing. https://doi.org/10.1007/978-3-030-46951-1_1

Motz, B., Quick, J., Schroeder, N., Zook, J., & Gunkel, M. (2019, March). The validity and utility of activity logs as a measure of student engagement. In *Proceedings of the 19th Learning Analytics and Knowledge Conference* (pp. 300–309). Association for Computing Machinery. https://doi.org/10.1145/3303772.3303789

Mouri, K., Uosaki, N., & Ogata, H. (2018). Learning Analytics for Supporting Seamless Language Learning using E-book with Ubiquitous Learning System. *Journal of Educational Technology & Society*, *21*(2), 150–163. https://www.jstor.org/stable/26388390

Peters, M., & Romero, M. (2019). Lifelong learning ecologies in online higher education: Students’ engagement in the continuum between formal and informal learning. *British Journal of Educational Technology*, *50*(4), 1729–1743. https://doi.org/10.1111/bjet.12803

Ruiz-Calleja, A., Bote-Lorenzo, M. L., Asensio-Pérez, J. I., Villagrá-Sobrino, S. L., Alonso-Prieto, V., Gómez-Sánchez, E., García-Zarza, P., Serrano-Iglesias, S., & Vega-Gorgojo, G. (2023). Orchestrating ubiquitous learning situations about Cultural Heritage with Casual Learn mobile application. *International Journal of Human-Computer Studies*, *170*, 102959. https://doi.org/10.1016/j.ijhcs.2022.102959

Spiro, R. J., Vispoel, W. P., Schmitz, J. G., Samarapungavan, A., & Boerger, A. E. (2013). Knowledge acquisition for application: Cognitive flexibility and transfer in complex content domains. In *Executive Control Processes in Reading* (1st Edition, pp. 41–74). Routledge. https://doi.org/10.4324/9780203063552-7

Yang, Y., & Song, Y. (2023). Understanding primary students’ self-regulated vocabulary learning behaviours on a mobile app via learning analytics and their associated outcomes: a case study. *Journal of Computers in Education*, *10*(3), 469–498. https://doi.org/10.1007/s40692-022-00251-x

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

A Study of Strategies to Enhance Student Engagement in Middle School Classrooms with Instructional Feedback in a Smart Paper - Pen Environment

Nan Dang[a*],Jingyi Zhang[a],Xinyu Sun[a],Xiaoyan Wang[a], Huilong Li[a]
Minsheng Fan[b]&Xiaodong Wei[a]
[a]*Institute of Educational Technology*
Northwest Normal University
Lanzhou, China
[b]*Key Laboratory of Education Digitalization of Gansu Province*
Lanzhou, China
*Minsheng Fan@corresponing.author

Abstract: In the context of educational digitization, middle school classrooms face dual challenges: nsufficient student engagement and the simplification, standardization, and rigid sourcing of traditional instructional feedback. This study, based on an analysis of feedback issues in middle school classrooms and a survey of students' feedback needs, combines the technical advantages of Smart Paper-Pen Systems in terms of scientific teaching evaluation and diversified classroom interaction. Following principles such as interactivity and continuity, it constructs a theoretical model for enhancing student engagement through instructional feedback in a Smart Paper-Pen Environment, and explores a strategic framework for improving student engagement by optimizing instructional feedback. This study provides a theoretical framework and practical pathways for integrating instructional feedback with student engagement in smart classrooms, aiding the deep integration of educational technology into classrooms, and holds both theoretical innovation and practical guidance significance..

Keywords: Smart Paper-pen System; instructional feedback; student engagement; middle school classroom

1. Introduction

With the advancement of education digitalization strategy, smart classroom has become an important way to improve the quality of education. In recent years, the education field has gradually recognized the importance of instructional feedback in the process of student learning. Instructional feedback is not only the teacher's evaluation of students' learning results, but also an important means to promote students' in-depth thinking and adjustment of learning strategies. In the middle school classroom, the problem of insufficient student engagement needs to be solved, and instructional feedback, as a key link in classroom teaching, has a potential impact on improving student engagement.

In the traditional classroom, teachers often provide feedback through verbal evaluation, homework correction, etc., but these methods have certain limitations, such as feedback is not timely enough and lack of personalization. The emergence of the smart paper and pencil system, on the other hand, provides new possibilities for instructional feedback. As an emerging educational technology tool, Smart Paper-pen System integrates the advantages of traditional paper-pen writing and modern information technology. It can record students' writing process, answer questions and other learning behavior data in real time, and transform these data into visual feedback information to help teachers more accurately understand the

students' learning status, so as to provide more targeted instructional feedback. At the same time, the Smart Paper-pen System also supports a variety of interactive features, such as classroom voting, real-time reviews, etc., which can enhance the interest of the classroom and the sense of participation of students.

This study is dedicated to exploring how to optimize instructional feedback by using the Smart Paper-pen System, which in turn enhances student engagement in middle school classrooms.

2. Definition of relevant concepts

2.1 Smart Paper-pen System

Smart Paper-pen System is an innovative educational tool system that integrates advanced information technology and traditional writing scenarios, which breaks the single mode of traditional paper-pen writing, and builds up a digital and intelligent learning and teaching ecology. The Smart Paper-pen System consists of dot-matrix digital pen, digital paper and cloud service. The system is able to collect students' learning data in real time, providing teachers with accurate instructional feedback basis, while supporting a variety of interactive methods to enhance the interest of the classroom and students' participation.

2.2 Instructional Feedback

Instructional feedback refers to the process by which teachers convey information about students' learning progress during the teaching process to promote their academic development. Its primary functions include providing correct answers, analyzing errors, offering problem-solving strategies, and guiding learning methods and strategies. Effective instructional feedback helps students understand their learning status, adjust their learning strategies, and thereby enhance learning outcomes. At its core, instructional feedback is a dynamic, two-way communication process. Teachers obtain information about students' learning through observing their learning behaviors and homework performance, and then provide corresponding feedback. Students, in turn, reflect on their learning processes based on teachers' feedback, identify issues, and implement improvement measures, thereby enhancing their learning autonomy and self-regulation abilities.

2.3 Student Engagement

Student engagement, also known as student involvement, is defined by Astin as the physical and psychological energy students invest in the learning process (Astin A W, 1984). Finn argues that student engagement encompasses both behavioral and emotional aspects. Behaviorally, it manifests in whether students frequently participate in classroom and school activities; emotionally, it refers to whether students feel they “belong” to the school environment and value outcomes related to the school (Finn, 1993). In 2001, Kuh formally defined student engagement, emphasizing that it involves the time, effort, and energy students invest in educational activities (Kuh, 2001). Chinese education scholar Professor Kong Qiping synthesized the consensus from domestic and international research, pointing out that although there are differences in terminology, scholars agree that student participation transcends overt behavior and encompasses deeper activities such as psychological, cognitive, and emotional processes. Ultimately, Professor Kong Qiping concluded that student participation involves activities across three dimensions: behavioral, cognitive, and emotional (Kong Qiping et al., 2003).

3. The Basis for Constructing Strategies

Traditional instructional feedback, due to its overly foundational nature, group-based feedback, and fixed sources, makes it difficult for students to achieve knowledge transfer,

often neglecting individual learning needs and leading to reduced learning engagement (Li Maoshen et al., 2022). The lack of peer feedback and self-reflection limits the development of students' self-reflection and metacognitive abilities (Wang Yun et al., 2021).

Therefore, this study conducted a needs survey on 88 first-year middle school students at a certain middle school in Guangzhou using the Instructional feedback Scale (Mo Muzhen et al., 2012). The data showed that in terms of feedback content, students desired timely feedback and targeted analysis, while also expecting positive recognition. Students of different academic levels had varying feedback needs: struggling students focused on filling gaps in foundational knowledge, average students focused on guidance in problem-solving approaches, and high-achieving students sought cross-disciplinary expansion and improvement. In terms of emotional needs, most students indicated that teachers' praise can enhance their willingness to speak up, and they prefer private feedback on mistakes rather than public criticism.

Based on the above results, the smart paper-pen system precisely meets students' feedback needs through four functions: scientific teaching evaluation, diversified classroom interaction, instant classroom feedback, and visualization of learning thinking. Specifically, it visualizes students' mastery of knowledge points through learning situation data; stimulates learning motivation through diversified interactive modules such as quick response and voting; provides timely feedback through real-time tracking of answer progress and grading; and displays the logic of problem solving through screen sharing and handwriting playback.

The principles of instructional feedback are based on Bandura's ternary interaction decision theory, the Hurlock effect, and Vygotsky's zone of proximal development theory, and are constructed from four dimensions: interactivity, continuity, positivity, and flexibility. The Smart Paper-Pen System collects and analyzes student performance data to drive teachers to dynamically adjust their teaching strategies; A comprehensive feedback system covering pre-class, in-class, and post-class phases is established to avoid fragmented teaching interventions; by reducing the frequency of negative feedback and stimulating students' intrinsic motivation (John Hattie et al., 2015); implementing dynamic tiered feedback—providing step-by-step guidance for students with weaker foundations and delivering cross-disciplinary extension tasks for advanced learners—ultimately maximizing the effectiveness of precise teaching.

Based on the above foundation, this study constructed strategies for improving student engagement in middle school classrooms through instructional feedback in a Smart Paper-Pen Environment.

4. Theoretical Models for Smart Paper-Pen Environment

The model of instructional feedback to enhance student engagement in middle school classroom under the Smart Paper-Pen Environment is shown in Figure 1, which refers to the “Emotional Design Supported Instructional Agent Feedback Model” proposed by Liu Qingtang and other scholars (Liu Qingtang et al., 2024), and is based on the theory of feedback teaching, the constructivist learning theory, the theory of emotional response, and the theory of ternary interaction decision-making.

Feedback-based teaching theory emphasizes the irreplaceable role of instructional feedback in learning motivation, cognitive development, and behavioral change during the teaching process. Smart paper-pen systems, which integrate information collection and interaction, cover all stages of teaching, including lesson preparation, instruction, interaction, and assessment. By efficiently integrating learning data, these systems provide a scientific basis for instructional feedback, which can be categorized into learning progress feedback and emotional feedback.

According to constructivist learning theory, learning is a process in which learners actively construct knowledge. Based on instructional feedback, learners are able to generate cognitive conflict and reflection, and behavioral, cognitive, and emotional involvement occurs so as to actively construct their personal knowledge system.

Emotional response theory asserts that teachers' language and emotions can directly influence students' emotional states and engagement levels (Sean M et al., 2012). Therefore,

instructional feedback should present positive emotional feedback. The reprocessed and integrated student learning data from the smart paper-and-pen system can serve as the basis for teachers' subsequent feedback, while also incorporating teachers' intuitive judgments of students' classroom participation.

Ternary interaction decision theory believes that there is mutual interaction and influence between the environment, the individual, and behavior. In teaching scenarios, teachers, students, teaching behaviors, and the environment engage in multiple interactions. In the whole process, the smart paper-pen, as the teaching environment and medium, interacts and dynamically generates with the teacher, the students, and the behaviors of the teacher and students.

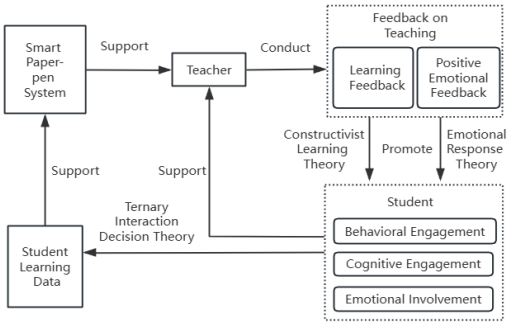


Figure.1 A model for improving student engagement in middle school classrooms with instructional feedback in a Smart Paper-pen Environment

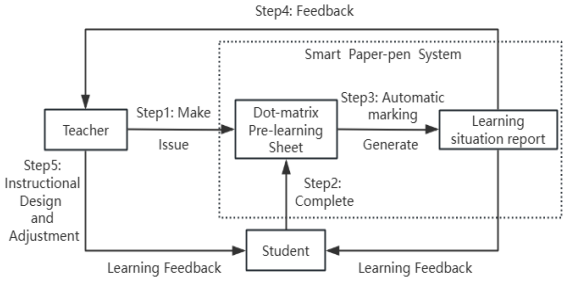


Figure.2 Flowchart of the guided learning strategy for the dot-matrix pre-learning sheet

5. Strategies for Improving Student Engagement in Middle School Classrooms

Based on the theoretical models for Smart Paper-Pen Environment, this paper proposes five strategies with instructional feedback to enhance student engagement in middle school classrooms.

5.1 Dot Matrix Pre-learning Sheet Guided Learning Strategy

Dot Matrix Pre-Learning Sheet is a pre-learning sheet presented in the form of dot matrix paper in Smart Paper and Pen. This guided learning strategy is primarily used before class, as shown in Figure 2. Teachers must design the pre-learning worksheet in advance and print it as a dot matrix paper version. The dot matrix pre-learning worksheet leverages the smart pen and paper system's automatic grading, generation of learning progress reports, and analysis functions to help teachers quickly and scientifically assess students' starting points. By understanding students' prior knowledge levels, teachers can design and adjust their teaching plans accordingly, making the teaching more targeted. Additionally, after completing the pre-learning worksheet and receiving feedback, students can gain a clearer self-awareness, focus more intently on the lesson, and enhance their behavioral and emotional engagement.

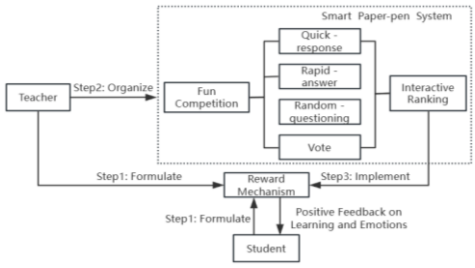


Figure.3 Fun Competitive Strategy Flowchart

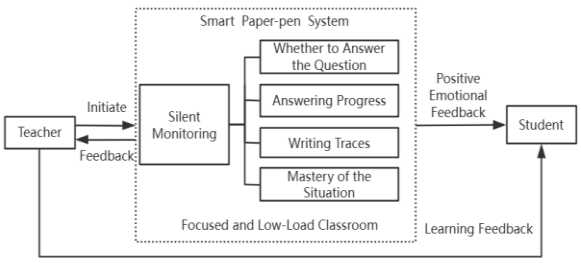


Figure.4 Silent Monitoring Strategy Flowchart

5.2 Fun Competitive Motivational Strategies

The fun competition motivation strategy is mainly used in the classroom. As shown in Figure 3, the essence of this strategy lies in utilizing the interactive leaderboard feature of the Smart Paper-Pen System, ensuring that every student interaction triggers a reward mechanism to enhance learning motivation. The quick response, fast response, and random question modules all include a reward star feature. The quick response feature displays the names of other students who attempted to answer when one student successfully answers, allowing teachers to award stars to these students to alleviate the frustration of those who failed to answer. Conducting tests through voting can help assess students' mastery of knowledge. Teachers distribute stars to students based on their test performance, and the stars earned by students are aggregated into an interactive leaderboard. Skinner's reinforcement theory suggests that intermittent reinforcement (as opposed to continuous reinforcement) can lead to higher response rates and lower extinction rates. Therefore, teachers can implement rewards based on the interactive leaderboard on a weekly or monthly basis. The frequency and magnitude of rewards should be mutually corresponding to give students a sense of “leveling up” participation. It is worth noting that competition and incentive mechanisms jointly developed by students and teachers are more effective.

5.3 Silent Monitoring Strategies

The silent monitoring strategy is mainly used in the middle of the class. As shown in Figure 4, this strategy uses the monitoring and voting functions of the smart paper-and-pen system to monitor students' learning progress in real time. The system allows teachers to view students' answers in real time from the podium, such as their progress in answering questions. Additionally, the handwriting replay function helps teachers quickly identify students' errors and areas of weakness. The voting function includes both named and anonymous voting, anonymous voting encourages students to express their true thoughts, and the system's real-time data statistics enable teachers to accurately understand the learning situation.

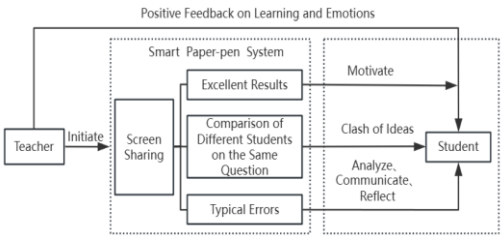


Figure.5 Achievement Display Strategy Flowchart

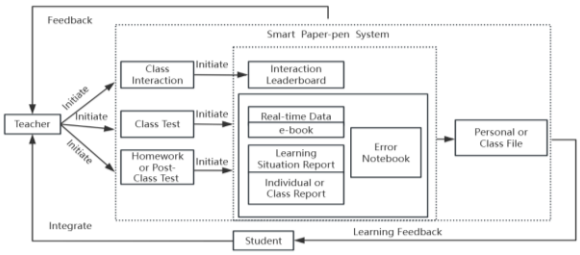


Figure.6 Comprehensive Evaluation and Feedback Strategy Flowchar

5.4 Achievement Display Strategy

Achievement display strategy refers to the use of the same-screen display function of the Smart Paper-Pen System to carry out feedback, which is mainly used in the middle of the lesson. As shown in Figure 5, this strategy primarily encompasses three key components: firstly, showcasing students' outstanding achievements, such as assignments with high accuracy rates, neatly organized exams, and innovative projects; Secondly, collectively analyzing and correcting errors. For simple errors, encourage students to point them out and correct them among themselves; for complex errors, let students discuss and encourage them to use the system to replay the writing process and explain their problem-solving approach. Thirdly, display and compare different solution methods on the same screen. Through this approach, students can clearly see the similarities and differences between different solutions and provide feedback to each other as equals, fostering divergent and exploratory thinking.

5.5 The comprehensive evaluation and feedback strategy

The comprehensive evaluation and feedback strategy can be flexibly applied before, during and after class. As shown in Figure 6, the comprehensive evaluation and feedback strategy collects students' multi-dimensional learning data through the Smart Paper-pen System, and eventually converges to form digital learning files, realizing the effective combination of process evaluation and summative evaluation. Together with the diagnostic evaluation implemented with the help of dot-matrix pre-learning sheet Guided learning strategy, comprehensive evaluation and feedback can be realized.

6. Summary of the study

This study demonstrates the unique advantages of smart paper-pen system in instructional feedback through the deep integration of technology and education. It not only provides practical strategies for improving student engagement in middle school classrooms, but also offers frontline teachers an operational model for technology-integrated teaching, laying the theoretical foundation for future empirical studies or classroom trials.

Acknowledgements

This work was supported by the [Humanities and Social Sciences Youth Foundation of Ministry of Education] under Grant [24YJC880091]; [National Natural Science Foundation of China] under Grant [62467007]. This work was supported by Key Laboratory of Education Digitalization of Gansu Province.

References

Astin A W.Student involvement: A developmental theory for higher education[J].Journal of College Student Development, 1984, 40(5):518-529.

Jeremy.School engagement and students at risk[C]//1993.

D.Kuh.Assessing What Really Matters to Student Learning: Inside the National Survey of Student Engagement[J].Change,2001,Vol.33(3): 10-17,66.

Kong Qi-Ping. Student Participation in the Process of Mathematics Teaching [M]. Shanghai: East China Normal University Press, 2003.

Li Maoshen, Zhu Leilei. The dilemma of instructional feedback in primary and secondary school classrooms and the path to improvement: Based on the perspective of Hattie's “visible learning”[J]. Educational Development Research,2022, Vol.42(20): 77-84.

John Hattie, Jin Yinglian, Hong Chao, Pei Xinning. Visible Learning (Teacher's Edition) - Maximizing Learning [M]. Beijing: Education Science Press, 2015:19.

WANG Yun,LI Zhixia,BAI Qingyu,et al. A study of peer feedback for deep learning in blended learning [J]. Modern Educational Technology, 2021, 31(05): 67-74.

Liu Qingtang,Yin, Xinghan,Zhang, Guangqiong,et al. A study on the construction and application of emotional design-supported instructional agent feedback model in online collaborative conversations[J]. Research on Electronic Education,2024,45(01):84-91,99.

Sean M. Horan;Matthew M. Martin;Keith Weber;.Understanding Emotional Response Theory:The Role of Instructor Power and Justice Messages. [J]. Communication Quarterly,2012,Vol.60(2): 210-233.

Mo Muzhen, Huang Yinghua, Yao Jingjing, Scott George PARIS, Gordon STANLEY, Jim TOGNOLINI. Analysis of the Psychological Properties of the “Instructional feedback Scale” for Primary School Students Using the Rasch Measurement Model [J]. Psychological Exploration, 2012, 32 (05): 387-396.

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Predicting Students' Online Learning—Character Strengths, Online Self-Regulated Learning, and Learning Satisfaction

Chin Lee KER[a*], Gaoxia ZHU[a]
[a]*National Institute of Education, Nanyang Technological University, Singapore*
** chinlee.ker@nie.edu.sg*
gaoxia.zhu@nie.edu.sg

Abstract: Online learning provides numerous benefits to students. However, it remains unclear how psychological and learning-related factors affect a student's performance in online learning environments. In this paper, we examined various psychological and learning-related factors (e.g., character strengths, online self-regulated learning, self-efficacy, learning satisfaction) that may predict learning performance in online environments. Results indicated that after controlling gender, disciplinary background and proportion of online or hybrid courses, self-efficacy in online learning was a significant positive predictor of learning performance. A possible explanation is that self-efficacy pushes students to persist in achieving their academic goals. In contrast, humanity (a character strength) and time management (a subconstruct of online self-regulated learning) were negatively associated with learning performance. Other predictors were not statistically significant. The results suggest that when one focuses more on nurturing relations with others, exhibiting humanity strengths, they may have less energy for their academic work; when one has good time management for online modules, they may neglect their face-to-face modules and do not perform well overall. Future research should investigate the underlying mechanisms of how humanity and time management impact students' learning performance in online learning environments.

Keywords: Online learning, character strengths, self-regulated learning, self-efficacy

1. Introduction

During the COVID-19 pandemic, modules were forced to shift online, and even in the post-pandemic era, online and blended learning is still widely implemented (Jamilah & Fahyuni, 2022). Online learning refers to experiencing education via technological tools and having connection to the internet (Moore et al., 2011). It also encompasses mobile learning, whereby students can use their mobile phones to access digital learning management systems like Google Classrooms or communication platforms like Teams to learn online. Online learning offers advantages like accessibility, flexibility, and opportunities for learners to collaborate, access resources, and study conveniently (Waschull, 2001). Despite these benefits, some students may have poor academic performance or poor course persistence when learning online, making it critical to study predictors that may influence how they perform to help improve the quality of online courses (Xu & Jaggers, 2013).

In online settings, many factors can influence how well a student performs academically. For example, character strengths like being persistent or having love for learning are associated with student enjoyment and achievement (Wagner et al., 2020). As students mostly learn independently in an online setting, self-regulation was crucial for them to succeed (Wijekumar et al., 2006). Satisfaction with online courses is essential as it is associated with achievement of course learning outcomes (Vikas & Mathur, 2022). Students

also need to have self-efficacy in handling and planning online learning behaviors (Zimmerman, 2008).

According to Peterson and Seligman (2004), character strengths are identifiable pathways to demonstrate virtues, which are valuable characteristics that are universal and classified into— wisdom (cognitive traits); courage (emotional traits); humanity (relational traits); justice (strong community life); temperance (protective traits); and transcendence (associations to the universe and purpose). Character strengths relate to educational outcomes— love of learning (wisdom) and perseverance (courage) are associated with achievement; prudence (temperance) is strongly related to overall school achievement; and social intelligence (humanity) is associated with self-rated group achievement (Wagner et al., 2020).

Online self-regulated learning (OSRL) includes setting goals, environment structuring, adopting task strategies, managing time, seeking help, and evaluating oneself (Barnard et al., 2008). There is a positive relationship between OSRL and Grade Point Average (GPA) (Barnard et al., 2008). In an online environment, there are multiple formats of complex digital content (Zimmerman, 2008), making self-regulation critical for students to coordinate and integrate content into workable cognitive representations (Azevedo et al., 2004).

Satisfaction with online learning (SOL), as well as online learning perception, is associated with achievement of course learning outcomes (Vikas & Mathur, 2022). Self-efficacy in online learning refers to how much students believe that they can utilize technology or online tools competently, and complete online learning work effectively; increase in self-efficacy in online learning contributes to students' better grades and academic achievement (Won et al., 2024). It could be possible that self-efficacy in online learning benefits online learning engagement (Kuo et al., 2021), which may translate to good performance. Academic satisfaction, the degree one enjoys their study experience (Lent et al., 2005), predicts achievement (Kumar & Dileep, 2006).

The studies mentioned above have shown that various factors influence learning. However, there are limited studies on how these factors predict students' learning performance in online settings, especially during the post-pandemic period. There is also a lack of research on predictors such as character strengths, OSRL, SOL or self-efficacy. To tackle these research gaps, this paper aims to study the different predictors of learning performance in online learning environments. This study is one of the first exploring the relationship between character strengths, OSRL, SOL and performance in higher education. Specifically, this study is guided by the research question: When gender, disciplinary background and proportion of online learning modules are controlled, how do different factors (character strengths, OSRL, SOL, self-efficacy in online learning, academic satisfaction) influence individual learning performance?

2. Methods

2.1 Participants and Instruments

829 responses were collected from a university in Singapore. After removing the duplicated and incomplete responses, there were 695 students. After filtering outliers and careless responses, 542 students remained in our final sample.

Participants' mean age was 25.32 years, with more females (n = 359). 72.14% were undergraduates, 25.65% were postgraduates, and 2.21% declared their program as "Others", as they were in special programs like "Postgraduate Diploma in Education". Participants were from— business (n = 147), engineering/ computer science (n = 126), science (n = 55), education (n = 108), humanities/ language (n = 44), medicine (n = 41), design/ communication (n = 23), or social sciences (n = 62). One could belong to multiple faculties. Most participants (n = 237) had less than 20% courses in online/hybrid format; 137 had 30% to 40% and 78 had

50%. Few (n = 48) had 60% to 80% and even fewer (n = 42) had more than 80% of courses in online/hybrid format.

Data was collected through an online self-reported survey between October 2023 and March 2024. Based on the Value in Action (VIA) Classification system (Peterson & Seligman, 2004), the 24-item Character Strengths-Semantic Differential Scale was used (CS-SDS, Chan et al., 2007). Participants rated descriptions on a 7-point continuum (e.g., Conforming-Creative). In our study, internal consistency is acceptable– α (wisdom)= 0.60; α (courage)= 0.60; α (humanity)= 0.59; α (temperance)= 0.55; α (transcendence)= 0.60. As the internal consistency for justice was low (α = 0.31), justice-related items were removed, after which the overall internal consistency is good (α = 0.86).

For OSRL, the Online Self-regulated Learning Questionnaire (OSLQ, Barnard et al., 2008) was used. The 24-item questionnaire measures— goal setting (α= 0.79), environment structuring (α= 0.77), task strategies (α= 0.67), time management (α= 0.71), help seeking (α= 0.67), and self-evaluation (α= 0.79). Overall internal consistency is good (α= 0.89).

SOL was measured using a 4-item scale developed by Lin (2005) (α= 0.90). Self-efficacy in online learning was measured using the Motivated Strategies for Learning Questionnaire (MSLQ, Pintrich et al., 1993). Using a 7-point Likert scale, students rated 8 statements (α= 0.93). Academic satisfaction was measured using a 7-item measure by Lent et al. (2005), using a 5-point Likert scale (α= 0.89). Participants were also asked to report their proportion of modules that were in online or hybrid format and their GPA category, used as a measure of online learning performance (out of 5)— < 2.0, 2.0 to 3.0, 3.0 to 4.0, or > 4.0.

2.2 Data Analysis

Firstly, 91 outliers were detected using “mahalanobis” function in the R stats package (R Core Team, 2025), with a *p*-value of 0.001 as a cutoff. Next, the careless R package (Yentes & Wilhelm, 2023) was applied to identify careless responses. We calculated the longstring index, then removed data points that exceeded 1.5 times the interquartile range above the third quartile of average longstring index, using “identify outliers” function in rstatix package (Kassambara, 2023). 63 careless responses were identified, with 1 case flagged as both outlier and careless. In total, 153 responses were removed, leaving us a final sample of 542. Lastly, as GPA category was ordinal, multiple ordinal logistic regression was done via “clm” function in the ordinal package (Christensen, 2023). Controlling gender, faculty, proportion of online/hybrid courses, these predictors were entered: subconstructs of character strengths and OSRL, SOL, self-efficacy in online learning, and academic satisfaction.

3. Results

Table 1 shows the results of the ordinal logistic regression. The model was statistically significant, $\chi^2(24) = 124.73$, $p < .001$, indicating that the predictors significantly distinguished outcome groups. The model explained between 21% (Cox & Snell R^2) and 25% (Nagelkerke R^2) of variance, with McFadden’s $R^2 = 0.13$. For control variables, being female is associated with 39% lower odds of being in a higher GPA category. Business (2.43 times), humanities/language (3.06 times), and social science students (3.37 times) are more likely to be in a higher GPA category than their respective counterparts. Humanity (β = -0.28, Odds Ratio = 0.76, $p = .02 < 0.05$) and time management (β = -0.56, Odds Ratio = 0.57, $p < .001$) were negatively associated with GPA category. In contrast, self-efficacy in online learning (β = 0.79, Odds Ratio = 2.20, $p < .001$) was a significant positive predictor of GPA category. Other predictors were not statistically significant.

Table 1. *Ordinal logistic regression to predict students’ individual learning performance*

Predictor	Estimate	Standard Error	Z	*p*	Odds

Gender (Female)	-0.50	0.21	-2.40	0.02*	0.61
Business	0.89	0.35	2.54	0.01*	2.43
Engineering/ Computer Science	0.21	0.35	0.58	0.56	1.23
Science	0.28	0.40	0.70	0.48	1.32
Education	-0.15	0.37	-0.42	0.68	0.86
Humanities/Language	1.12	0.41	2.73	0.01**	3.06
Medicine	-0.67	0.47	-1.42	0.15	0.51
Design/ Communication	0.76	0.55	1.38	0.17	2.13
Social Science	1.21	0.41	2.99	0.00**	3.37
Proportion of online/hybrid courses	0.04	0.08	0.45	0.65	1.04
Wisdom	0.01	0.17	0.04	0.97	1.01
Courage	0.09	0.18	0.49	0.62	1.09
Humanity	-0.28	0.12	-2.28	0.02*	0.76
Temperance	0.16	0.14	1.19	0.24	1.18
Transcendence	-0.27	0.17	-1.55	0.12	0.76
Goal Setting	0.27	0.19	1.42	0.15	1.31
Environment Structuring	-0.10	0.17	-0.58	0.56	0.91
Task Strategies	-0.18	0.15	-1.14	0.25	0.84
Time Management	-0.56	0.15	-3.81	<.001***	0.57
Help Seeking	0.07	0.17	0.44	0.66	1.08
Self-Evaluation	0.19	0.16	1.13	0.26	1.20
SOL	-0.03	0.10	-0.26	0.80	0.97
Self-Efficacy in Online Learning	0.79	0.12	6.66	<.001***	2.20
Academic Satisfaction	-0.09	0.21	-0.45	0.65	0.91

Notes: *$p < .05$; **$p < .010$; ***$p < .001$

4. Discussion and Conclusion

This study examines how various factors, such as character strengths or learning-related factors, influence how well a student performs in online settings. The result shows that each unit of increase in humanity is associated with 24% lower odds of being in a higher GPA category. We posit that an overemphasis on relational traits might detract participants from academic focus. Our finding is consistent with Azizi et al. (2019), who found a significant negative association between use of social networks and academic performance. A potential explanation is that if students spend too much time on social networking, they might compromise their academic performance in online learning, where they have more flexibility dividing their time between learning and socialization. Another explanation is if giving kindness is overdone and not reciprocal, it may be negatively associated with institutional identity and stress reduction, and the person giving kindness may not persist academically (Hosoda & Estrada, 2024).

Surprisingly, each unit increase in online time management is associated with 43% lower odds of a higher GPA. This finding contradicts with West and Sadoski (2011), who found that time management predicts academic performance, as those with good time management were less likely to delay content revision. This inconsistency may be related to how time management is defined and measured. For West and Sadoski (2011), good time management

refers to managing time well for medical school/ practical courses. In our study, good time management refers to fixing a schedule to study for online/ hybrid courses, and allocating extra studying time for them. Furthermore, as online learners could be flexible and arrange time in an unstructured way (Zheng et al., 2018), they may have compensated for the lack of in-person interaction by allocating more time for online/ hybrid modules. This may backfire if they neglect face-to-face modules, which still comprise the majority of modules.

Higher self-efficacy is associated with 2.20 times the odds of achieving higher GPA categories. This could be related to learners' confidence and belief that they will succeed academically. Our findings echo Won and colleagues (2024), who found that academic self-efficacy in online learning predicted students' grades. The belief that one can complete an academic task successfully increases achievement likelihood (Won et al., 2024). With higher self-efficacy, students tend to have better online learning engagement (Kuo et al., 2021) that may transform into academic achievement.

There are some limitations: (1) our participants were from a university in Singapore, making our findings not generalizable to a wider population; (2) GPA was used as a measure of learning performance, which involves academic outcomes for all modules, instead of only online modules. Future studies can explore how to isolate the outcomes of online and hybrid learning from overall outcomes to better understand how various factors influence online learning performance, to inform the design of online learning strategies.

In conclusion, this paper contributes to the literature regarding online learning and informs educators to pay attention to the specific attributes of humanity, time management, and self-efficacy of students when delivering online learning classes. To improve online learning, educators could: (1) remind students to not be excessively relational or over focus on social networking; (2) remind students to have balanced time management across all modules, instead of allocating too much time resources for online/ hybrid modules only; and (3) help students build self-efficacy, for example, by creating scaffolded tasks, enabling them to have opportunities to feel academically self-efficacious (Won et al., 2024).

Acknowledgements

We would like to thank the participants who made this study possible. This study is supported by the Ministry of Education (MOE) Academic Research Fund (AcRF, Reference No. RG 115/22).

References

Azevedo, R., Cromley, J. G., & Seibert, D. (2004). Does adaptive scaffolding facilitate students' ability to regulate their learning with hypermedia? *Contemporary Educational Psychology, 29*(3), 344–370. https://doi.org/10.1016/j.cedpsych.2003.09.002

Azizi, S. M., Soroush, A., & Khatony, A. (2019). The relationship between social networking addiction and academic performance in Iranian students of medical sciences: A cross-sectional study. *BMC Psychology, 7*(1), 28–28. https://doi.org/10.1186/s40359-019-0305-0

Barnard, L., Paton, V., & Lan, W. (2008). Online self-regulatory learning behaviors as a mediator in the relationship between online course perceptions with achievement. *International Review of Research in Open and Distance Learning, 9*(2), 1–11. https://doi.org/10.19173/irrodl.v9i2.516

Chan, F., Chou, C. C., Keegan, J., & Ditchman, N. (2007). *Development and validation of the character strengths—semantic differential scale (CS-SDS).* Department of Rehabilitation Psychology and Special Education, University of Wisconsin-Madison.

Christensen, R. H. B. (2023). *ordinal— Regression models for ordinal data*. R package version 2023.12-4.1. https://CRAN.R-project.org/package=ordinal.

Hosoda, K. K., & Estrada, M. (2024). The influence of kindness on academics' identity, well-being and stress. *PloS One, 19*(10), e0312269-. https://doi.org/10.1371/journal.pone.0312269

Jamilah, J., & Fahyuni, E. F. (2022). The future of online learning in the Post-COVID-19 era. *KnE Social Sciences*, 497-505. https://doi.org/10.18502/kss.v7i10.11251

Kassambara, A. (2023). *rstatix: Pipe-friendly framework for basic statistical tests*. R package version 0.7.2. https://CRAN.R-project.org/package=rstatix

Kumar, S. P. K., & Dileep, P. (2006). *Academic life satisfaction scale and its effectiveness in predicting academic success.* (ERIC Document Reproduction Service No. ED491869). Retrieved from ERIC database.

Kuo, T. M., Tsai, C., & Wang, J. (2021). Linking web-based learning self-efficacy and learning engagement in MOOCs: The role of online academic hardiness. *The Internet and Higher Education, 51*, 100819. https://doi-org.libproxy.nie.edu.sg/10.1016/j.iheduc.2021.100819

Lent, R. W., Singley, D., Sheu, H.-B., Gainor, K. A., Brenner, B. R., Treistman, D., & Ades, L. (2005). Social cognitive predictors of domain and life satisfaction: Exploring the theoretical precursors of subjective well-being. *Journal of Counseling Psychology, 52*(3), 429–442. https://doi.org/10.1037/0022-0167.52.3.429

Lin, Y. M. (2005). *Understanding students' technology appropriation and learning perceptions in online learning environments.* University of Missouri-Columbia.

Moore, J. L., Dickson-Deane, C., & Galyen, K. (2011). E-Learning, online learning, and distance learning environments: Are they the same? *The Internet and Higher Education, 14*(2), 129–135. https://doi-org.libproxy.nie.edu.sg/10.1016/j.iheduc.2010.10.001

Peterson, C., & Seligman, M. E. (2004). *Character strengths and virtues: A handbook and classification.* Oxford university press.

Pintrich, P. R., Smith, D. A. F., Garcia, T., & Mckeachie, W. J. (1993). Reliability and predictive validity of the motivated strategies for learning questionnaire (MSLQ). *Educational and Psychological Measurement, 53*(3), 801–813.

R Core Team. (2025). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.r-project.org/

Vikas, S., & Mathur, A. (2022). An empirical study of student perception towards pedagogy, teaching style and effectiveness of online classes. *Education and Information Technologies, 27*(1), 589–610. https://doi.org/10.1007/s10639-021-10793-9.

Wagner, L., Holenstein, M., Wepf, H., & Ruch, W. (2020). Character strengths are related to students' achievement, flow experiences, and enjoyment in teacher-centered learning, individual, and group work beyond cognitive ability. *Frontiers in Psychology, 11*, 1324–1324. https://doi.org/10.3389/fpsyg.2020.01324

Waschull, S. B. (2001). The online delivery of psychology courses: Attrition, performance, and evaluation. *Teaching of Psychology*, *28*(2), 143–147. https://doi.org/10.1207/S15328023TOP2802_15

West, C., & Sadoski, M. (2011). Do study strategies predict academic performance in medical school? *Medical Education, 45*(7), 696–703. https://doi.org/10.1111/j.1365-2923.2011.03929.x

Wijekumar, K., Ferguson, L., & Wagoner, D. (2006). Problem with assessment validity and reliability in web-based distance learning environments and solutions. *Journal of Educational Multimedia & Hypermedia*, https://www.proquest.com/docview/205853455?pqorigsite=gscholar&fromopenview=true

Won, S., Kapil, M. E., Drake, B. J., & Paular, R. A. (2024). Investigating the role of academic, social, and emotional self-efficacy in online learning. *The Journal of Experimental Education, 92*(3), 485–501. https://doi.org/10.1080/00220973.2023.2183375

Xu, D., & Jaggars, S. S. (2013). The impact of online learning on students' course outcomes: Evidence from a large community and technical college system. *Economics of Education Review*, *37*, 46–57. https://doi.org/10.1016/j.econedurev.2013.08.001

Yentes R, Wilhelm F (2023). *careless: Procedures for computing indices of careless responding.* R package version 1.2.2.

Zheng, C., Liang, J.-C., Li, M., & Tsai, C.-C. (2018). The relationship between English language learners' motivation and online self-regulation: A structural equation modelling approach. *System, 76*, 144–157. https://doi-org.remotexs.ntu.edu.sg/10.1016/j.system.2018.05.003

Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal, 45*(1), 166–183. https://doi.org/10.3102/0002831207312909

Meta AI for Leadership Training: A Pilot Study on Mobile Social Media–Based Blended and Self-Regulated Learning

Riska SAPUTRA[a*], Zainul ARIFIN[b], Tsaqufal JALILIY [c], & Husni MUBAROK[d]
[a]*Graduate Institute of Digital Learning and Education, National Taiwan University of Science and Technology, Taiwan*
[b]*Department of Da'wah Management, Syarifuddin Islamic University, Indonesia*
[c]*Faculty of Language Art & Culture, Yogyakarta State University, Indonesia*
[d]*Educational Technology, Faculty of Education, Surabaya State University, Indonesia*
*riskysaputra17@gmail.com

Abstract: This exploratory pilot study investigates the integration of Meta AI into WhatsApp as a mobile-based tool for leadership education compared with traditional book-based methods. Eighteen university students were divided into two groups: one engaged in Blended Learning with Meta AI and Self-Regulated Learning (BMAI-SRL), while the other relied on Blended Learning with SRL but without Meta AI (B-SRL). Data were collected through pre-/post-tests on leadership knowledge, questionnaires measuring critical thinking, self-efficacy, and SRL, as well as open-ended reflections. Both groups demonstrated significant within-group improvements, but no significant differences were found between groups. Thematic analysis revealed that BMAI supported efficiency, personalization, and flexibility but also posed challenges of reliability and distraction, whereas books provided cultural grounding, deeper understanding, and more consistent engagement. Overall, the findings suggest that a blended approach combining the adaptability of AI with the cultural depth of traditional methods may enhance leadership learning, though larger and longer-term studies are needed to validate these findings.

Keywords: Meta AI, Mobile-Assisted Learning, Leadership Training, Self-Efficacy, Critical Thinking, Self-Regulated Learning

1. Introduction

The rapid growth of digital technologies has transformed higher education worldwide, particularly through mobile platforms and artificial intelligence (AI). Generative AI increasingly shapes how students learn, engage, and regulate their studies (Fullan et al., 2024). Mobile-Assisted Learning (MAL) leverages portable devices and ubiquitous connectivity to support continuous access, while AI-powered tools provide adaptive feedback and personalized support (Moya & Camacho, 2024). These developments present opportunities for leadership education, which requires critical thinking, ethical reasoning, decision-making, and communication skills (Effendi et al., 2020). In Indonesia, leadership is deeply rooted in cultural traditions, emphasising collective responsibility, ethical judgment, and spiritual values (Andriyanti et al., 2024). Traditionally cultivated through books, religion, and community, these competencies now face tension as younger generations turn to digital tools and social media.

Self-Regulated Learning (SRL) provides a bridge by promoting autonomy, goal setting, monitoring, and reflection (Zimmerman, 2000). The integration of Meta AI into WhatsApp—a platform with over 86 million active users in Indonesia (Kumar, 2025)—offers potential to embed SRL strategies within interactive digital scaffolding, making leadership training more engaging and learner-centred. Yet, limited empirical research has compared AI-enabled MAL with traditional book-based methods in leadership education. To address this gap, the present study pilots Meta AI integration into WhatsApp and contrasts it with conventional learning, examining their effects on leadership knowledge, critical thinking, self-efficacy, and SRL, while also capturing students' qualitative experiences.

2. Literature Review

2.1 Leadership Training in Higher Education

Leadership training in higher education often combines cognitive, affective, and social learning. It is not only about acquiring theoretical knowledge but also about cultivating character, empathy, and decision-making skills that reflect community and cultural contexts (Effendi et al., 2020). In Indonesia, culturally responsive leadership training emphasizes local wisdom, which integrates religious teachings, collective values, and respect for authority (Andriyanti et al., 2024). Book-based learning remains widely used because it embeds cultural narratives and provides structured content that encourages reflection. At the same time, leadership education faces challenges such as limited engagement, lack of personalization, and difficulty in connecting abstract principles with real-world practice. Emerging digital approaches, particularly those using AI and mobile platforms, offer new ways to engage students. However, scholars caution that cultural grounding must remain central to avoid decontextualized or shallow learning experiences (Fullan et al., 2024).

2.2 Mobile-Assisted Learning and Self-Regulated Learning

MAL provides ubiquitous, interactive, and flexible learning opportunities (Hwang et al., 2024). By leveraging smartphones and social media platforms, learners can access educational content anytime and anywhere. When integrated with AI, MAL can offer dynamic adaptation of content and immediate feedback, supporting all three stages of SRL—forethought, performance, and reflection (Weng et al., 2024). For example, AI can recommend tasks, adjust content difficulty, and provide real-time assessments tailored to learners' progress. These affordances align well with leadership training, where reflection and adaptability are key. However, MAL and AI also introduce challenges: learners may struggle with distractions, misinformation, and variable digital literacy. Moreover, overreliance on AI may reduce deep reflection unless carefully scaffolded. Taken together, the literature suggests that while MAL and AI hold promise for enhancing SRL and leadership competencies, their effectiveness depends on careful integration with cultural values and structured learning designs.

3. Method

As shown in Figure 1, Eighteen undergraduate students from a leadership course at a private university in East Java, Indonesia, participated in a three-week study (December 2024–January 2025), randomly assigned to an experimental group using Blended Learning with Meta AI via WhatsApp and SRL strategies (BMAI-SRL) or a control group using book-based blended learning with SRL but without Meta AI (B-SRL). Both groups attended an initial webinar on leadership styles and local wisdom, followed by self-study using their assigned methods.

As illustrated in Figure 2, students set weekly goals, created study plans, and evaluated their progress. The BMAI-SRL group interacted with a chatbot to request definitions, analyze cases, and explore leadership scenarios, supplemented by textbooks and online articles, while the B-SRL group accessed curated readings without AI support. While Figure 3, show students' interaction with Meta AI via WhatsApp.

Quantitative data included pre- and post-tests (10 items, scored 0–100) and post-intervention questionnaires using 5-point Likert scales on critical thinking (Chai et al., 2015; $\alpha = 0.71$), self-efficacy (Pintrich et al., 1991; $\alpha = 0.79$), and SRL (Barnard et al., 2009; $\alpha = 0.90$). Open-ended questions captured qualitative reflections. Data analysis employed Wilcoxon Signed-Rank tests for within-group changes, Mann–Whitney U tests for between-group comparisons, and gain score analysis for knowledge improvement. Thematic analysis followed GenAI-assisted coding procedures (Turobov et al., 2024), code generation (Step 1), code clustering (Step 2), and developing AI-specific codes (Step 3) and a frequency analysis was conducted to determine how often each code was mentioned across both groups.

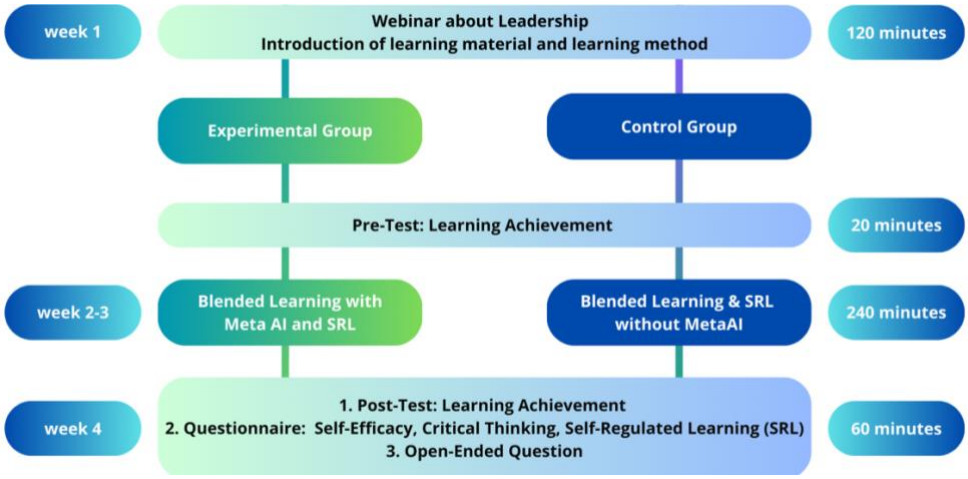


Figure 1. Experiment Process

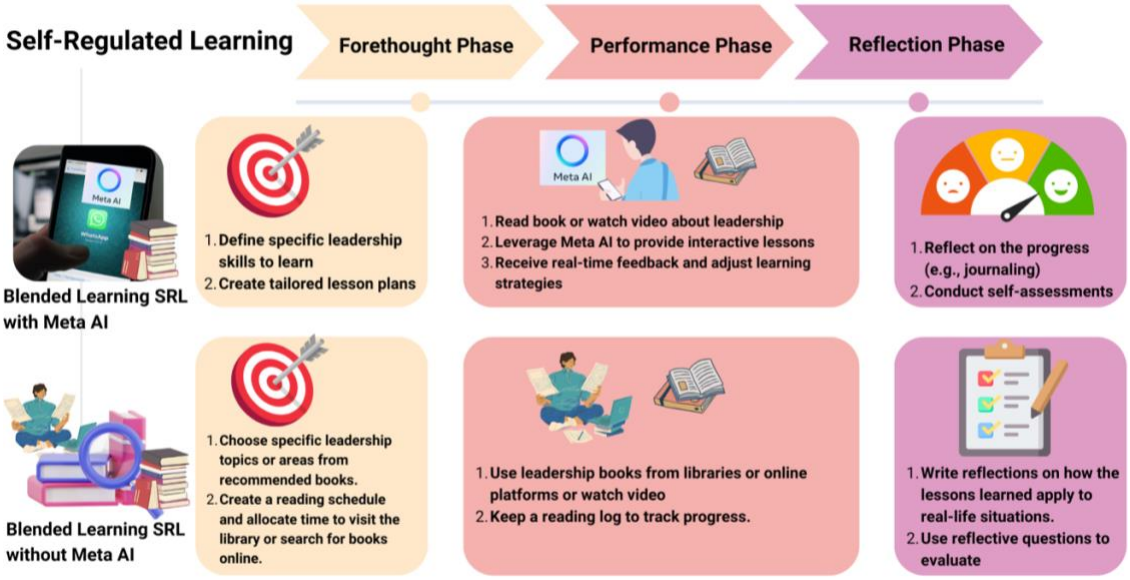
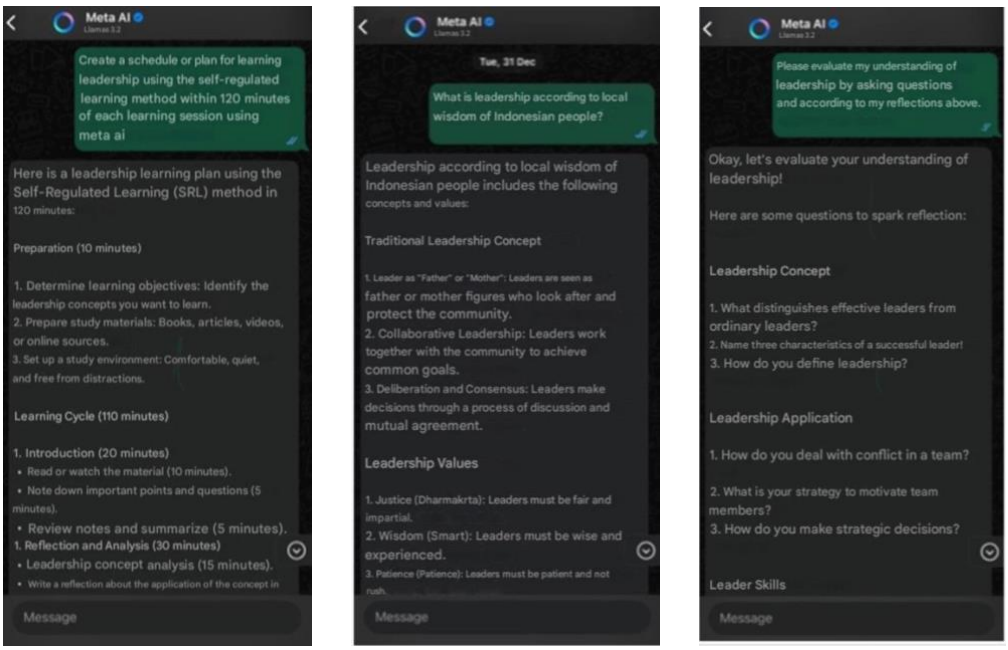


Figure 2. Learning Activity Process



(a) Setting study plan (b) Interactive learning (c) Evaluation

Figure 3. Screenshot of learning with Meta AI

4. Results

4.1 Learning Achievement

Table 1. *Wilcoxon Test for learning achievement*

Group	Variance	N	Mean	SD	Z	α
Blended Learning Meta AI-SRL (BMAI-SRL)	Pre Test	9	10.0	8.6	2.687	0.007**
	Post Test	9	74.4	12.3		
Blended Learning SRL without Meta AI (B-SRL)	Pre Test	9	10.0	7.07	2.692	0.007**
	Post Test	9	73.3	10		

**p-value <0.01

Table 2. *Mann-Whitney U Test for gain score of learning achievement*

Variance	N	Mean Rank	Sum of Ranks	U	Z	α
BMAI-SRL	9	9.83	88.50	37.500	-0.277	0.782
B-SRL	9	9.17	82.50			

p-value > 0.05

Table 1 shows significant improvements in both the BMAI-SRL and B-SRL groups from pre-test to post-test. As seen in Table 2, the gain scores between groups were not significantly different suggesting both methods were equally effective in this pilot study.

4.2 Critical-Thinking, Self-Efficacy, and Self-Regulated Learning

Table 3. *Mann-Whitney U Test for Critical Thinking*

Variance	N	Mean Rank	Sum of Ranks	U	Z	α
BMAI-SRL	9	9.83	88.50	37.500	-0.268	0.789
B-SRL	9	9.17	82.50			

p-value > 0.05

Table 3 indicates no significant difference in critical thinking. Both groups improved, though the B-SRL group showed slightly more consistent outcomes.

Table 4. *Mann-Whitney U Test for Self-Efficacy*

Variance	N	Mean Rank	Sum of Ranks	U	Z	α
BMAI-SRL	9	7.17	64.50	19.500	-1.886	0.059
B-SRL	9	11.83	106.50			

p-value > 0.05

As shown in Table 4, the B-SRL group reported higher confidence, but the difference was not significant. This suggests both methods supported self-efficacy, with the book group showing greater stability.

Table 5. *Mann-Whitney U Test for Self-Regulated Learning (SRL)*

Item	Variance	N	Mean Rank	Sum of Ranks	U	Z	α
Goal Setting	BMAI-SRL	9	9.44	85.00	40.000	-0.046	0.963
	B-SRL	9	9.56	86.00			
Environment	BMAI-SRL	9	10.22	92.00	34.000	-0.587	0.557
	B-SRL	9	8.78	79.00			
Task Strategy	BMAI-SRL	9	9.28	83.50	38.500	-0.189	0.850
	B-SRL	9	9.72	87.50			
Time Management	BMAI-SRL	9	9.00	81.00	36.000	-0.407	0.684
	B-SRL	9	10.00	90.00			
Help Seeking	BMAI-SRL	9	8.56	77.00	32.000	-0.804	0.422
	B-SRL	9	10.44	94.00			
Self-Evaluation	BMAI-SRL	9	8.61	77.50	32.500	-0.755	0.450
	B-SRL	9	10.39	93.50			

p-value > 0.05

Table 5 shows no significant differences across SRL dimensions. Both groups engaged in SRL strategies, with the B-SRL group slightly stronger in help seeking and self-evaluation, while BMAI-SRL showed greater variability.

Table 6. *Thematic Analysis*

Theme	Code	The number of times mentioned	
		BMAI-SRL	B-SRL
Depth of Understanding and Value Integration	New knowledge, understanding values, simple, sacred, relevant, wide view but sometimes shallow	19	15
Emphasis on Interaction and Shared Experience	Discussion, teamwork, teacher's encouragement, Master Solid	9	10
Learning Challenges	Difficulty understanding, time management, applying theory, laziness, unstable network, AI validity, distractions,	14	12
Desire for Practical Application and Real-World Relevance	Practical approach, tech/media, case studies, direct practice, less boring, desire for practice, real-world visits, human interaction.	11	9
Relevance and Motivation	Relevant (pesantren), team, self-leadership, self-will, culturally meaningful, personally motivating	16	14
Application and Confidence	Apply in teaching, moral/ethics, care/support, build solid team, applying learning, responsibility, growing confidence	13	13

To capture qualitative perspectives, students were asked the following open-ended reflection question. These open-ended questions aim to capture participants' impressions, challenges, understanding, motivation, and the overall impact of learning leadership based on local wisdom on their cultural relevance, engagement, and leadership development. Thematic analysis revealed that both groups shared similar experiences. Depth of understanding and value integration appeared most often (BMAI-SRL = 19, B-SRL = 15), followed by relevance and motivation (16, 14). Common themes also included learning challenges (14, 12) and the desire for practical application (11, 9). Both groups equally highlighted application and confidence (13, 13), indicating that learners valued putting knowledge into practice and building self-assurance regardless of medium.

5. Discussion

This exploratory pilot study found that both the Meta AI–supported blended learning (BMAI-SRL) and traditional book-based blended learning (B-SRL) significantly improved leadership knowledge, critical thinking, self-efficacy, and self-regulated learning, with no statistically significant differences between groups—suggesting that combining SRL with either digital or traditional modalities supports short-term learning gains in leadership education (Zimmerman, 2000; Weng et al., 2024; Hwang et al., 2024). The BMAI-SRL group valued efficiency, personalization, and flexibility, though they also faced challenges related to AI reliability and distraction, aligning with concerns about the need for robust scaffolding in AI-mediated learning environments (Moya & Camacho, 2024; Fullan et al., 2024). Conversely, the B-SRL condition offered deeper cultural grounding and more consistent engagement, in line with literature emphasizing the importance of embedding local wisdom and cultural narratives in leadership development (Andriyanti et al., 2024; Effendi et al., 2020). However, the study's generalizability is constrained by its small sample size, short intervention duration, and reliance on self-reported measures, while the imbalance in social and instructional support across groups introduces a design confound that needs attention. Future research should address these limitations by recruiting larger and more diverse samples, implementing longer-term and longitudinal designs, balancing collaborative and AI-guided learning structures, and incorporating objective performance measures alongside qualitative reflections. Despite these caveats, the results underscore the promise of a blended instructional model that harmonizes AI-driven adaptability with culturally embedded reflective materials in nurturing context-sensitive, adaptive leadership learning.

6. Conclusion

This study concludes that both BMAI-SRL and B-SRL, when supported by blended learning and self-regulated learning, significantly enhanced leadership knowledge with no major

differences in critical thinking, self-efficacy, or SRL outcomes. While BMAI-SRL offered flexibility and personalization but faced issues of reliability and distraction, whereas B-SRL provided cultural grounding and consistency engagement. Given the study's small sample size, short duration, and reliance on self-reported data, future research should adopt larger, more diverse samples, longer-term designs, and objective performance measures to validate and extend these findings. Overall, the integration of AI-driven adaptability with culturally embedded reflective learning holds promise for developing context-sensitive leadership education.

7. Acknowledgment

We acknowledge the support of the National Science and Technology Council (NSTC), Taiwan, and our partner universities in Indonesia for their contributions to this study.

References

AlAfnan, M. A. (2024). Taxonomy of educational objectives: Teaching, learning, and assessing in the information and Artificial Intelligence era. *Journal of Curriculum and Teaching, 13*(4), 173. https://doi.org/10.5430/jct.v13n4p173.

Andriyanti, M., Suwandayani, B. I., & Amelia, D. J. (2024). The power of integrity of local wisdom in basic education: Sustainable development. *Journal of Learning Improvement and Lesson Study, 4*(2), 10–23. https://jlils.ppj.unp.ac.id/index.php/jlils/article/view/110

Barnard, L., Lan, W. Y., To, Y. M., Paton, V. O., & Lai, S.-L. (2009). Measuring self-regulation in online and blended learning environments. *The Internet and Higher Education, 12*(1), 1–6. https://doi.org/10.1016/j.iheduc.2008.10.005

Breitwieser, J., Nobbe, L., Biedermann, D., & Brod, G. (2023). Boosting self-regulated learning with mobile interventions: Planning and prompting help children maintain a regular study routine. *Computers & Education,* (104879), 104879. https://doi.org/10.1016/j.compedu.2023.104879

Burke, P. F., Schuck, S., Burden, K., & Kearney, M. (2025). Mediating learning with mobile devices through pedagogical innovation: Teachers' perceptions of K-12 students' learning experiences. *Computers & Education, 227*(105226), 105226. https://doi.org/10.1016/j.compedu.2024.105226

Chai, C. S., Deng, F., Tsai, P.-S., Koh, J. H. L., & Tsai, C.-C. (2015). Assessing multidimensional students' perceptions of twenty-first-century learning practices. *Asia Pacific Education Review, 16*(3), 389-398. https://doi.org/10.1007/s12564-015-9379-4

Effendi, Y. R., Bafadal, I., Degeng, I. N. S., & Arifin, I. (2020). The principal's transformational leadership approach based on local wisdom in strengthening the character of students. *MOJEM: Malaysian Online Journal of Educational Management*, *8*(4), 24-42. https://mojem.um.edu.my/article/view/26379

Fullan, M., Azorín, C., Harris, A., & Jones, M. (2024). Artificial intelligence and school leadership: challenges, opportunities and implications. *School Leadership and Management, 44*(4), 339–346. https://doi.org/10.1080/13632434.2023.2246856.

Hwang, G.-J., Rahimi, M., & Fathi, J. (2024). Enhancing EFL learners' speaking skills, foreign language enjoyment, and language-specific grit utilising the affordances of a MALL app: A microgenetic perspective. *Computers & Education, 214*(105015), 105015. https://doi.org/10.1016/j.compedu.2024.105015

Kumar, N. (2025). *WhatsApp Users Statistics (2025): How Many People Use It?*. DemandSage. https://www.demandsage.com/whatsapp-statistics/

Moya, S., & Camacho, M. (2024). Leveraging AI-powered mobile learning: A pedagogically informed framework. *Computers and Education: Artificial Intelligence, 7*(100276), 100276. https://doi.org/10.1016/j.caeai.2024.100276

Pintrich, P.R., Smith, D.A.F., Garcia, T., & McKeachie, W.J. (1991). A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ). MI: National Center for Research to Improve Postsecondary Teaching and Learning. (ERIC Document Reproduction Service No. ED 338122). 1991. https://files.eric.ed.gov/fulltext/ED338122.pdf

Weng, X., Xia, Q., Ahmad, Z., & Chiu, T. K. F. (2024). Personality traits for self-regulated learning with generative artificial intelligence: The case of ChatGPT. *Computers and Education: Artificial Intelligence,* 7(100315), 100315. https://doi.org/10.1016/j.caeai.2024.100315.

Turobov, A., Coyle, D., & Harding, V. (2024). *Using ChatGPT for thematic analysis.* https://doi.org/10.48550/arXiv.2405.08828

C5
Educational Gamification and Game-based Learning (EGG)

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Impact of Gamification on Engagement and Learning in Video-Based Platforms

Pasan PEIRIS[a]*, Matthias GALSTER[b], Antonija MITROVIC[a], Sanna MALINEN[a], Ehsan BOJNORDI[a] & Jay HOLLAND[a]
[a]*University of Canterbury, New Zealand*
[b]*University of Bamberg, Germany*
*pitigalagepasan.peiris@pg.canterbury.ac.nz

Abstract: Video-based learning provides flexible and self-paced learning. However, passive video consumption results in shallow learning. To address this, reflective activities such as commenting on videos and reviewing peers' comments are used to increase engagement. Incorporating gamification can further boost engagement. This study investigates the impact of gamification, specifically badges, on student engagement and learning in video-based learning. We conducted a comparative study between a gamified and a non-gamified version of the same platform. Findings show that badges significantly increased learner engagement. Student feedback indicated a positive reception of badges. Also, students who used the gamified platform wrote higher quality reflective comments on videos, leading to increased learning.

Keywords: Gamification, Badges, Video-Based Learning, Engagement, Learning

1. Introduction

Video-based learning (VBL) offers flexibility and allows learners to control the pace of learning (Sablić et al., 2021). However, passive consumption of videos often results in low knowledge retention (Chatti et al., 2016). Various strategies can increase learner engagement, including annotations (Chatti et al., 2016), quizzes (Koedinger et al., 2015), and interactive note taking (Mitrovic et al., 2016). Learners who take quality notes by capturing and reflecting on key video concepts achieve better knowledge gains than those who do not (Taskin et al. 2019).

Another way to increase engagement is gamification (Lee & Kim, 2025). Gamification leverages principles of game design and integrates game-like elements, such as points, badges, and leaderboards into non-game contexts (Deterding et al., 2011). Gamification motivates university students to achieve certain goals and behaviors, resulting in increased knowledge and performance (Tahir et al., 2022; O'Donovan et al., 2013).

In this paper we use gamification in VBL for presentation skills, one of the most demanded soft skills in industry (Galster et al., 2023). Presentation skills are the "Ability to exchange (send and receive) information in different form orally (e.g., in formal presentations or informal conversations)" (Galster et al., 2023). We address three research questions (RQs):

- **RQ1:** To what extent does gamification influence learner engagement in VBL?
- **RQ2:** To what extent does gamification affect learning outcomes in VBL?
- **RQ3:** What are learners' attitudes towards gamification in VBL?

We make the following contributions:

- We describe how gamification influences engagement based on interaction measures (e.g., platform usage, activity completion rates). We found that gamification enhances engagement in reflective activities and generally increases the use of the VBL platform.
- We explore how gamification influences learning outcomes in a VBL platform. We found that gamification influences learning by motivating students to write more high-quality comments when reflecting on video content.
- Regarding learners' perceptions towards gamification, gamification is well received by students. Also, self-assessed perceptions about the impact of gamification on learning are consistent with the actual learning effectiveness of gamification.

2. Background

Video-based learning: Videos present knowledge in a consistent and attractive manner (Yousef et al., 2014). By incorporating both audio and visual elements, videos offer a structured and engaging way to present training content for better understanding and knowledge retention (Yousef et al., 2014). VBL also removes physical and time barriers, allowing training in a setting and at a time that best suits learners (Sablić et al., 2021). Video-based training is effective and valued by students (Giannakos et al., 2016).

Despite its popularity, VBL provides limited human interaction and limited interaction with video content (Chatti et al., 2016). To avoid passive learning, active video watching (AVW) utilizes reflective activities such as annotations, writing comments, and rating comments (Correia & Chambel, 1999; Lau et al., 2016). Galster et al. (2018) conclude that AVW is promising for teaching presentation skills. Since then, there have been multiple developments in AVW to increase engagement and learning. For example, Mohammadhassan et al. (2022) enhanced AVW-Space (an AVW platform) with nudges, leading to a notable increase in conceptual knowledge. These improvements highlight that increasing engagement in learning activities increases learning in AVW platforms.

Gamification: Gamification is the use of game design elements in non-game contexts (Deterding et al., 2011). Example game elements include points, badges, leaderboards, challenges, and levels (Klock et al., 2020). Gamification aims to increase the intrinsic and extrinsic motivation of the user and get them involved in tasks by combining playful activities (Buckley & Doyle, 2014). Intrinsic motivation is the satisfaction from undertaking a task for its implicit satisfaction (for example, employees performing their tasks out of enjoyment), while extrinsic motivation is the satisfaction from achieving explicit rewards (for example, certificates) (Manzano-León et al., 2021; Fischer et al., 2019).

Gamification is beneficial for different educational levels, from primary to tertiary (Manzano-León et al., 2021). For example, Tahir et al. (2022) integrated gamification into an Intelligent Tutoring System (ITS) for a database course. Challenges, goals, and badges motivated students to perform more complex tasks. O'Donovan et al. (2013) gamified a course on computer game development using storyline, points, progress bars, badges, and leaderboards in an online learning tool. Student engagement and performance improved as a result of gamification, compared to the previous non-gamified course.

3. Research Method

3.1 Study Design

We utilized purposive sampling (Baltes & Ralph, 2022) to recruit participants from students enrolled in ENGR101, a mandatory first-year course for all engineering students at the University of Canterbury, New Zealand in the years of 2023 and 2024. The study was conducted in the context of AVW-Space (Lau et al., 2016), a VBL platform which facilitates engagement through three learning activities: watching instructional videos, commenting to reflect on and interacting with the content, and reviewing peers' comments to encourage further reflection. The study received approval from the institution's human ethics board.

The study was conducted in two phases: In the first phase, students completed a pre-survey (Survey 1) to provide demographic data and answers to conceptual knowledge questions on presentation skills (see Section 3.3). Then, the students watched and commented on the eight videos. At the end of this phase, the instructor selected relevant student comments and anonymously published them in AVW-Space. In the second phase, the students reviewed their peers' comments. Finally, the students completed a post-survey (Survey 2), which included the same conceptual knowledge questions. In 2024 we added a treatment to AVW-Space to gamify the learning through badges to further enhance the engagement. Consequently, in 2024 we also surveyed students on their perception towards

badges and the gamified AVW-Space. Students in both years received 1% of course credit for undertaking the exercise.

3.2 Game Elements and Implementation

We selected badges as the game element. Prior work showed that badges have the potential to support learning (Denny et al., 2018; Hakulinen et al., 2015; Tahir et al., 2022). Other elements, such as points and leaderboards which foster competitiveness (Amo et al., 2020) were discarded due to their potential negative impact on certain user demographics, such as extroverts (Codish & Ravid, 2014) and their drawbacks in educational contexts (Toda et al., 2017). Also, competition and social comparison have detrimental effects on motivation, behavior and peer relationships (Verheijen et al., 2019; Hanus & Fox, 2015). To mitigate these risks, we implemented badges using task-based awarding conditions, promoting individual progress without fostering competition. We designed 10 badges reflecting the three main learning activities in AVW-Space. We implemented three badges for watching videos, five badges for commenting on videos, one badge for reviewing peers' comments, and one badge to reflect the achievement of all other badges, see Figure 1. Table 1 presents the awarding conditions.

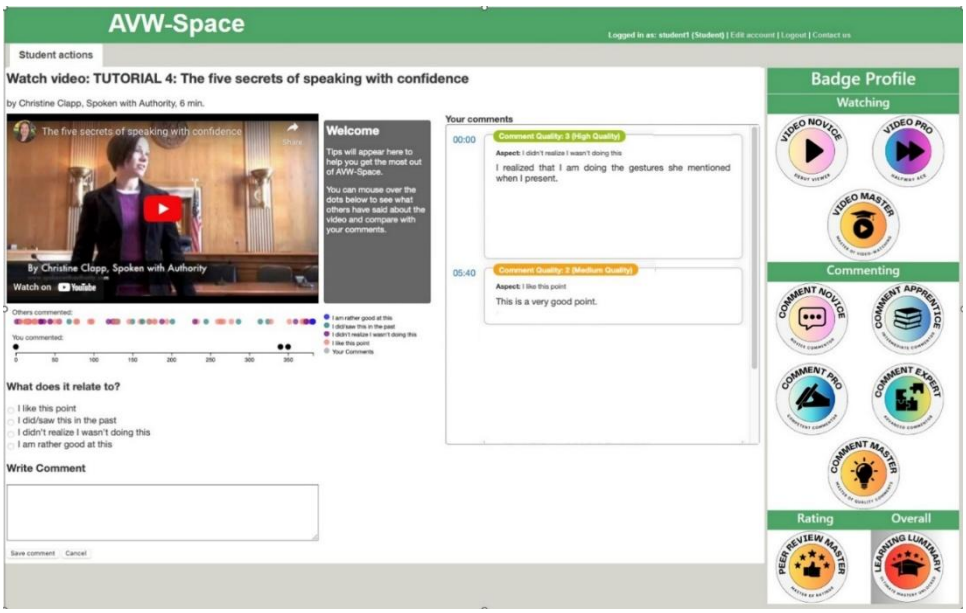


Figure 1. Gamified AVW-Space

3.3 Data Analysis

To measure engagement **(RQ1),** we analyzed AVW-Space logs: data on the number of distinct days users logged in, the number of user sessions, the number of videos watched, the number of comments posted, the number of nudges (personalized prompts) received, the number of high-quality comments written, and the number of peer comments reviewed.

Since assessing learning **(RQ2)** of students during live presentations was impractical due to the large cohort, we used an alternative approach. We asked students to complete timed pre- and post-surveys, listing terms related to delivery, structure, and visual aids. We automatically scored their responses using ontology developed in prior research (Dimitrova & Mitrovic, 2022) and used the resulting Conceptual Knowledge scores (CK1 and CK2) as a proxy for learning. We utilized descriptive statistics and statistical tests to compare the gamified (2024 group) and non-gamified (2023) groups. Perception surveys **(RQ3)** included closed and open-ended questions. For closed questions, we used descriptive statistics and quantitative methods. For open-ended questions, we analyzed responses using content analysis (Seidel, 1998). We started by extracting key phrases relevant to each question. Then, we identified recurring themes and organized them into distinct categories (Seidel, 1998).

Table 1. *Badges and Awarding Conditions*

Badge	Task-based awarding conditions
Video Novice	Watch your first video
Video Pro	Watch 4 videos
Video Master	Watch all 8 videos
Comment Novice	Write your first high-quality comment
Comment Apprentice	Write two high-quality comments for the first video
Comment Pro	Write two high-quality comments for the next video
Comment Expert	Write three high-quality comments each for the next 3 videos
Comment Master	Write five high-quality comments each for the next 3 videos
Peer Review Master	Rate 15 peer comments
Learning Luminary	Accumulate all other badges

Furthermore, we used structural equation modelling (SEM) to examine the impact of gamification on learning outcomes **(RQ2)** in the gamified group (Figure 2). We defined four hypotheses to test as part of answering RQ2. Prior research indicates that badges moderate the effect of prior knowledge on time-on-task (Tahir et al., 2022). From that we infer our first hypothesis (H1) that students with higher prior knowledge are more likely to earn more badges. Our second hypothesis (H2) suggests that the number of badges earned has a positive effect on the number of high-quality comments written by students. This is because badges are designed to prompt users to write more high-quality comments. Our third hypothesis (H3) suggests that high quality comments influence the post-test knowledge. Previous research has established that high-quality comments are positively correlated with post-test knowledge (CK2) (Mohammadhassan et al., 2022). Our fourth hypothesis (H4) suggests that pre-existing knowledge positively influences the posttest knowledge. Studies in video-based training have shown a positive relationship between CK1 and CK2 (Mitrovic et al., 2019). Based on these established correlations and our hypotheses, we propose that high-quality comments mediate the relationship between badges and CK2.

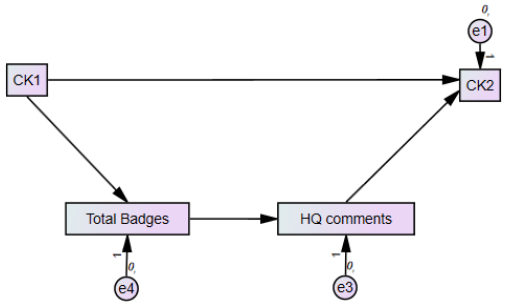


Figure 2. Hypothesized path analytic model

4. Results

We analyzed the data from students who completed Survey 1 and watched at least one video. The non-gamified group included 645 students (Year 2023), while the gamified group had 223 students (Year 2024).

4.1 Engagement (RQ1)

Engagement data were not normally distributed (Shapiro-Wilk test), so we employed the Mann-Whitney U test to compare the gamified and non-gamified groups. We found significant

differences in all engagement metrics except the number of videos watched (Table 2). This may be because students were told to watch all eight videos. Overall, the analysis indicates that the gamified group engaged significantly more than the non-gamified group.

Table 2. *Measures for Engagement (* indicates significant difference)*

	Gamified [2024] (N=223)	Non-gamified [2023] (N=645)	Significance
	Mean (sd), Median	Mean (sd), Median	
Distinct Days*	4.46 (3.63), 3	3.09 (2.24), 3	U=57864, p<.001
User sessions*	5.74 (5.11), 4	3.83 (3.25), 3	U=55713.5, p<.001
Videos watched	7.27 (1.95), 8	7.35 (1.86), 8	U=72705, p=.67
Comments posted*	25.15 (15.21), 31	12.39 (9.73), 9	U=38933.5, p<.001
Nudges received*	32.33 (16.45), 34	23.08 (13.40), 21	U=45326, p<.001
High quality comments written*	19.20 (12.59), 28	6.70 (6.44), 6	U=38152, p<.001
Comments reviewed*	19.09 (20.93), 15	16.54 (21.15), 10	U=59498, p<.001

4.2 Learning (RQ2)

We compared the learning outcomes of students who completed both Survey 1 and Survey 2 (Table 3). 149 students from the gamified group and 493 students from the non-gamified group completed both surveys. Data were not normally distributed, so we used the Mann-Whitney U test. Results showed no significant difference in learning gains (CK2-CK1) between the two groups. However, there was a significant difference between the two groups on the CK1 scores, indicating that students did not start with similar levels of prior knowledge. This suggests that although both groups improved similarly, the initial disparity in prior knowledge may have influenced how each group engaged. Therefore, interpreting learning gains without considering baseline differences could be misleading. As a result, we cannot conclusively explain the impact of gamification using CK1 and CK2 when comparing the two groups directly. However, a Wilcoxon Signed-Rank Test conducted separately for the gamified group ($Z=2.73$, $p=.006$) and the non-gamified group ($Z=2.608$, $p=.009$) indicated significant within-group improvement from pre-test (CK1) to post-test (CK2). This shows that both groups improved their post-test scores compared to their pre-test scores.

Table 3. *Learning (* indicates significant difference)*

	Gamified (N=149)	Non-gamified (N=493)	Significance
	Mean (sd), Median	Mean (sd), Median	
CK1*	11.60 (5.14), 11	14.04 (5.82), 14	U=45955.5, p=<.001
CK2*	13.06 (6.44), 13	14.14 (6.75), 14	U=41832, p=.001
CK2 – CK1	1.46 (6.43), 1	0.70 (6.58), 1	U=34377, p=.24

We evaluated the path diagram using IBM SPSS Amos version 28.0.0, analyzing data from 149 participants with complete data (Figure 3). Based on the inputs and number of variables, the model includes 12 parameters. A sample size of at least 10 participants per parameter is recommended (Raykov & Widaman, 1995), making our dataset appropriate. All variables in the model are observed. The chi-square test ($\chi^2 = 2.915$, df = 2, p = .23) indicates that the model's predictions are not significantly different from the observed data. The model fit indices further support this: the Comparative Fit Index (CFI) is .998 and the Root Mean Square Error of Approximation (RMSEA) is .027. These values meet accepted thresholds (CFI > .90, RMSEA < .06) for a good model fit (Hu & Bentler, 1999; Teo et al, 2013).

The model indicated that higher CK1 scores directly cause higher CK2 scores (beta = .45, p<.001), confirming Hypothesis H4. Therefore, the effect of the number of high-quality comments on CK2 is adjusted for and above and beyond this influence (.1, p = 006) which confirms Hypothesis H3. The number of badges affects the number of high-quality comments (.96, p<.001). Therefore, hypothesis H2 is confirmed. However, pre-exiting knowledge (CK1) does not affect the total number of badges (.04, p = 382). Therefore, hypothesis H1 is rejected.

This suggests that gamification via badges enhances post-test knowledge (CK2) by encouraging students to produce high-quality comments. Interestingly, the number of badges earned appears to be independent of pre-existing knowledge (CK1) and instead exerts its own direct influence on learning outcomes. One explanation is that even students with strong prior knowledge may not necessarily know how to write high-quality comments. Additionally, those who already feel knowledgeable might be less motivated to engage fully.

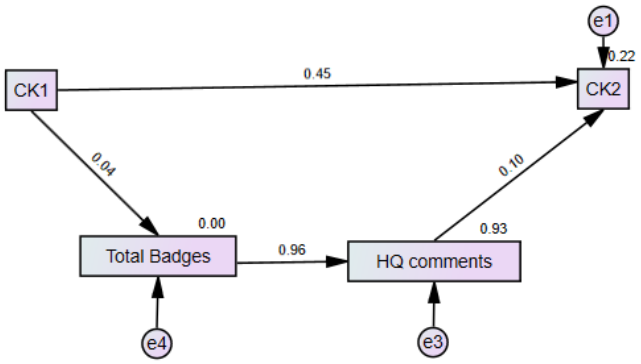


Figure 3. Path Diagram for the Gamified Group

4.3 Student Perception on Gamification (RQ3)

Figure 4 illustrates the participant perceptions on the badges implemented in the AVW-Space. A majority of respondents (43%) liked the badges and 2% indicated they did not see any badges. In the following, we excluded those who did not see any badges. Over half liked the visual appearance of the badges (53%) and felt that the badge names and descriptions accurately represented levels of achievement (64%). Most participants found the process of earning badges to be neither too easy nor too difficult (53%) and reported checking their badges frequently (57%). Additionally, many felt that badges motivated them (54%), increased their engagement (51%), and expressed overall satisfaction with the badge system (51%). Notably, the perception of increased engagement was consistent with the quantitative measures, which showed higher levels of engagement for the gamified group.

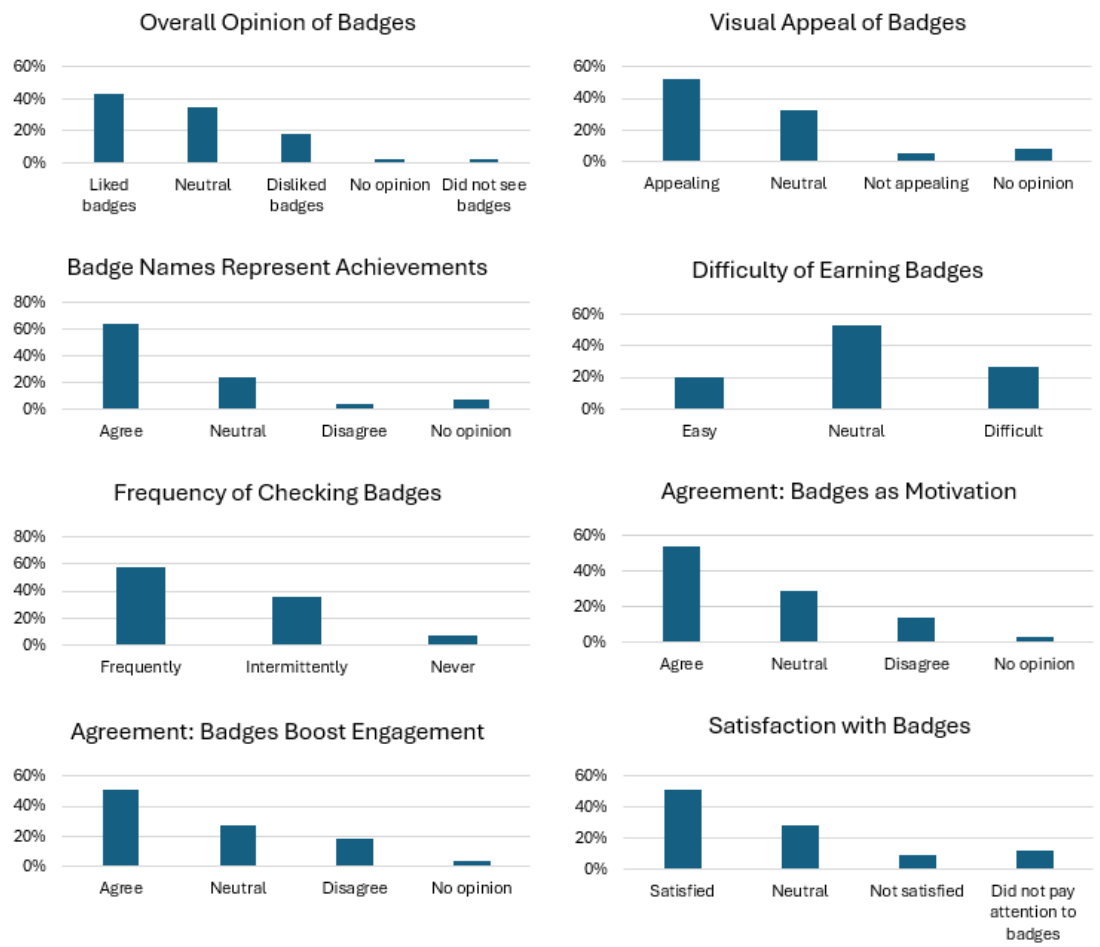


Figure 4. Perception of Badges

In addition to the above, we also asked open questions to participants who used the gamified platform to describe their experience with the badges in terms of their impact on motivation and engagement with the learning activities. The majority of the responses (n=80) reported a positive experience. Many participants (n=39) noted that the badges increased their motivation and engagement (which confirms the responses to the closed question discussed in the previous paragraph), often stating that they felt more encouraged to actively participate in the activities. For example, one participant wrote "*The badges motivated me to complete each task they gave, so in turn I believe they were effective*". Others went further to describe specific reasons for their positive experience. A frequently mentioned benefit was the sense of clear goals and direction provided by the badges (n=17). Participants appreciated that the badges helped break down larger tasks into smaller, more manageable steps. One participant commented "*The badges broke the overall large task into smaller tasks making it more manageable and easier to complete*". Six participants highlighted the usefulness of badges in making their progress visible. As one put it, "*The badges were a good way to tick off parts of the training, to see your progress*". In contrast, 27 responses were neutral. These participants typically indicated that, although they understood the purpose of the badges, they sometimes struggled to determine how to write high-quality comments. This may suggest the need for further interventions, such as providing clearer explanations for what constitutes a high-quality comment. Lastly, 27 reported a negative experience. The main issues were that earning the badges required too much effort and that the process felt tedious. This suggests that, for these participants, the badges did not support intrinsic motivation.

Furthermore, we asked participants how earning badges influenced their learning experience on the platform. The majority of participants (n=64) reported a positive impact, which was consistent with the results of the SEM analysis, indicating that badges had a significant influence on post-test knowledge. The key reasons for this were consistent with

those mentioned earlier, such as increased motivation to learn, clear goals, and visible progress. In addition, nine participants indicated that the badges helped them improve their learning and reflection. One participant noted, "*It made me think more deeply to get high quality comments*". However, ten participants reported a negative impact. Of these, five described the experience as frustrating, with one stating, "*It made the learning experience feel like a chore*". Another three participants mentioned that the focus on earning badges led to superficial learning and distraction, as one explained, "*It made me less focused on the task and more focused on getting the checkpoints* […]".

5. Discussion and Conclusion

We implemented badges associated with task-based awarding conditions to investigate their impact on students' perception, engagement, and learning within video-based learning. Our findings indicate that gamification significantly enhances engagement across various measures. In particular, gamification was effective in promoting constructive learning behaviors, such as writing high-quality comments and rating peers' comments. These findings align with the findings of previous research (Manzano-León et al., 2021; Tahir et al., 2022).

However, due to significant differences in pre-existing knowledge (CK1) between the gamified and non-gamified groups, we were unable to directly compare learning outcomes across groups. Despite this limitation, structural equation modelling within the gamified group revealed the influence of gamification on learning-related variables (e.g., high-quality comments, CK2). The analysis demonstrated that badges exert an indirect effect on post-test knowledge (CK2), mediated by the high-quality comments. These results suggest that badges contribute meaningfully to learning outcomes. These results align with previous research which state gamification contributes to the performance of students (O'Donovan et al., 2013). Future research could further explore different mediated pathways of gamification via different attributes associated with learning. Furthermore, we observed that pre-existing knowledge about presentation skills did not significantly influence the number of badges a student received. For future research, it would be valuable to explore other variables that may affect a student's level of immersion in gamification. Additionally, researchers could investigate whether a latent variable such as the ability to write constructive comments may be contributing to this level of immersion (note that in both groups, over 87% of participants who completed both surveys reported English as their first language, i.e., writing constructive comments was most likely not due to a lack of language skills).

In addition to performance measures, we also gathered participants' perceptions of the badges. The feedback was largely positive, particularly regarding aspects such as visual design, award criteria, motivation, engagement, learning experience and overall satisfaction. Our research empirically demonstrated that self-assessed perception data were also consistent with the observed quantitative improvements in engagement and learning, reinforcing the conclusion that badges are both a well-received and an effective intervention.

We acknowledge several threats to validity. In terms of construct validity, we used conceptual knowledge scores (CK1 and CK2) as a proxy for learning. While this approach provided an objective, scalable measure, it is an indirect indicator that may not fully capture actual competency in presentational skills that include performative dimensions (e.g., body language). Regarding internal validity and confounding factors, the students may have had different schedules impacting the time they spent on AVW-Space in different years. More critically, the two groups showed significant differences in pre-learning scores, limiting our ability to attribute learning gains specifically to badges. In terms of external validity, the study was conducted in a video-based learning environment (AVW-Space) for presentation skills and employed badges as the gamification intervention, which may limit the generalizability of findings to other platforms with different game elements or for training different skills. In terms of conclusion validity, comment quality as an engagement measure was determined using a machine learning model trained on prior data. However, superficial, keyword-rich comments may have been misclassified as high quality. There is also a risk that students gamed the system by deliberately including relevant terms without fully understanding them, potentially

weakening the validity of conclusions about badges promoting genuine cognitive engagement. Additionally, some students may not have been motivated to meaningfully complete the pre- and post-tests, which could compromise the accuracy of the CK1 and CK2 measures and threaten the validity of conclusions drawn from these variables. In addition, the reliability of the study's findings may be impacted by sample-related factors. Specifically, the 2024 cohort (gamified group) had a notably smaller size (n = 223) compared to the 2023 (non-gamified) group (n = 645), due to partial participation of eligible students. This imbalance could affect reliability, as the smaller sample size in the gamified group may have reduced the statistical power.

Overall, our findings offer empirical support for the use of badges as a gamification intervention to enhance engagement and learning VBL. They also indicate that badges are well received by students. Our findings offer practical guidance for researchers and designers of gamified learning, highlighting the value of integrating badges with clearly defined goals and pathways to encourage desired behaviors in educational settings.

References

Amo, L., Liao, R., Kishore, R., & Rao, H. R. (2020). Effects of structural and trait competitiveness stimulated by points and leaderboards on user engagement and performance growth: A natural experiment with gamification in an informal learning environment. European Journal of Information Systems, 29(6), 704–730.

Baltes, S., & Ralph, P. (2022). Sampling in software engineering research: A critical review and guidelines. Empirical Software Engineering, 27(4), 94.

Buckley, P., & Doyle, E. (2014). Gamification and student motivation. Interactive Learning Environments, 24(6), 1162–1175.

Chatti, M. A., Marinov, M., Sabov, O., Laksono, R., Sofyan, Z., Fahmy Yousef, A. M., & Schroeder, U. (2016). Video annotation and analytics in CourseMapper. Smart Learning Environments, 3(1), 10.

Codish, D., & Ravid, G. (2014, February). Personality based gamification-Educational gamification for extroverts and introverts. In Proceedings of the 9th CHAIS Conference for the Study of Innovation and Learning Technologies: Learning in the Technological Era (Vol. 1, pp. 36-44).

Correia, N., & Chambel, T. (1999). Active video watching using annotation. Proceedings of the Seventh ACM International Conference on Multimedia (Part 2), 151–154.

Denny, P., McDonald, F., Empson, R., Kelly, P., & Petersen, A. (2018). Empirical support for a causal relationship between gamification and learning outcomes. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1–13). Association for Computing Machinery.

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From Game Design Elements to Gamefulness: Defining Gamification. Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, 9-15.

Dimitrova, V., & Mitrovic, A. (2022). Choice Architecture for Nudges to Support Constructive Learning in Active Video Watching. Intl. Journal of Artificial Intelligence in Education, 32(4), 892–930.

Fischer, C., Malycha, C. P., & Schafmann, E. (2019). The Influence of Intrinsic Motivation and Synergistic Extrinsic Motivators on Creativity and Innovation. Frontiers in Psychology, 10.

Galster, M., Mitrovic, A., & Gordon, M. (2018). Toward Enhancing the Training of Software Engineering Students and Professionals Using Active Video Watching. 2018 IEEE/ACM 40th Intl. Conference on Software Engineering: Software Engineering Education and Training (ICSE-SEET), 5–8.

Galster, M., Mitrovic, A., Malinen, S., Holland, J., & Peiris, P. (2023). Soft skills required from software professionals in New Zealand. Information and Software Technology, 160, 107232.

Giannakos, M. N. (2013). Exploring the video-based learning research: A review of the literature. British Journal of Educational Technology, 44(6)

Giannakos, M. N., Krogstie, J., & Aalberg, T. (2016). Video-based learning ecosystem to support active learning: Application to an introductory computer science course. Smart Learning Environments, 3(1), 11.

Hakulinen, L., Auvinen, T., & Korhonen, A. (2015). The effect of achievement badges on students' behavior: An empirical study in a university-level computer science course. International Journal of Emerging Technologies in Learning, 10(1).

Hanus, M. D., & Fox, J. (2015). Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance. Computers & education, 80, 152-161.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modelling: A Multidisciplinary Journal, 6(1), 1–55.
Klock, A.C.T., Gasparini, I., Pimenta, M. S., & Hamari, J. (2020). Tailored gamification: A review of literature, International Journal of Human-Computer Studies,144, 1071-5819.
Koedinger, K., McLaughlin, E., Kim, J., Jia, J., & Bier, N. (2015). Learning is Not a Spectator Sport: Doing is Better than Watching for Learning from a MOOC. 111–120.
Lau, L., Mitrovic, A., Weerasinghe, A., & Dimitrova, V. (2016). Usability of an active video watching system for soft skills training. In Proc. 1st Int. Workshop on Intelligent Mentoring Systems, ITS.
Lee, J., & Kim, D. (2025). Levelling up learning in higher education: Gamification of in-video components to supercharge student learning and achievement in video-based learning. British Journal of Educational Technology.
Manzano-León, A., Camacho-Lazarraga, P., Guerrero, M. A., Guerrero-Puerta, L., Aguilar-Parra, J. M., Trigueros, R., & Alias, A. (2021). Between Level Up and Game Over: A Systematic Literature Review of Gamification in Education. Sustainability, 13(4), 2247
Michel, N., Cater, J. J., & Varela, O. (2009). Active versus passive teaching styles: An empirical study of student learning outcomes. Human Resource Development Quarterly, 20(4), 397–418.
Mitrovic, A., Dimitrova, V., Lau, L., Weerasinghe, A., & Mathews, M. (2017). Supporting Constructive Video-Based Learning: Requirements Elicitation from Exploratory Studies. Lecture Notes in Computer Science, 224–237.
Mitrovic, A., Dimitrova, V., Weerasinghe, A., & Lau, L. (2016). Reflective Experiential Learning: Using Active Video Watching for Soft Skills Training. Proceedings of the 24th International Conference on Computers in Education, 192–201.
Mitrovic, A., Gordon, M., Piotrkowicz, A., & Dimitrova, V. (2019). Investigating the Effect of Adding Nudges to Increase Engagement in Active Video Watching. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, & R. Luckin (Eds.), Artificial Intelligence in Education (pp. 320–332). Springer International Publishing.
Mohammadhassan, N., Mitrovic, A., & Neshatian, K. (2022). Investigating the effect of nudges for improving comment quality in active video watching. Computers & Education, 176, 104340.
O'Donovan, S., Gain, J., & Marais, P. (2013). A case study in the gamification of a university-level games development course. Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference, 242–251.
Raykov, T., & Widaman, K. F. (1995). Issues in applied structural equation modeling research. Structural Equation Modeling, 2(4), 289–310.
Robson, K., Plangger, K., Kietzmann, J., McCarthy, I. P., & Pitt, L. (2015). Is it all a game? Understanding the principles of gamification. Business Horizons, 58(4), 411–420.
Sablić, M., Mirosavljević, A., & Škugor, A. (2021). Video-Based Learning (VBL)—Past, Present and Future: An Overview of the Research Published from 2008 to 2019. Technology, Knowledge and Learning, 26(4), 1061–1077.
Seidel, J. V. (1998). Qualitative data analysis. In The Ethnograph v5.0: A user's guide (Appendix E).
Tahir, F., Mitrovic, A., & Sotardi, V. (2022). Investigating the causal relationships between badges and learning outcomes in SQL-Tutor. Research and Practice in Technol. Enhanced Learning, 17(1), 7.
Taskin, Y., Hecking, T., Hoppe, H. U., Dimitrova, V., & Mitrovic, A. (2019). Characterizing Comment Types and Levels of Engagement in Video-Based Learning as a Basis for Adaptive Nudging. In M. Scheffel, J. Broisin, V. Pammer-Schindler, A. Ioannou, & J. Schneider (Eds.), Transforming Learning with Meaningful Technologies (pp. 362–376). Springer International Publishing.
Teo, T., Tsai, L. T., & Yang, C. C. (2013). Applying structural equation modeling (SEM) in educational research: An introduction. In Application of structural equation modeling in educational research and practice (pp. 1–21). Brill Sense.
Toda, A. M., Valle, P. H., & Isotani, S. (2017). The dark side of gamification: An overview of negative effects of gamification in education. In Researcher links workshop: higher education for all (pp. 143-156).
Verheijen, G. P., Stoltz, S. E., van den Berg, Y. H., & Cillessen, A. H. (2019). The influence of competitive and cooperative video games on behavior during play and friendship quality in adolescence. Computers in Human Behavior, 91, 297-304.
Yousef, A. M. F., Chatti, M. A., & Schroeder, U. (2014). The State of Video-Based Learning: A Review and Future Perspectives. Advances in Life Sciences, 6, 122–135.

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Impact of Digital Learning Game Narratives on Affective Experiences and Learning

J. Elizabeth RICHEY[a*], Daniel POLLOCK[b], Mahboobeh MEHRVARZ[b], Aditi HAIMAN[b], Hayden STEC[b], Leah TEFFERA[b], Bruce M. MCLAREN[b]
[a]*Department of Psychology, University of Pittsburgh, USA*
[b]*Human-Computer Interaction Institute, Carnegie Mellon University, USA*
*jer177@pitt.edu

Abstract: Digital learning games can support learning by improving students' affective experiences; however, these experiences may depend on the game narrative, and there is some evidence that narrative preference may vary by gender. We developed two different narratives of the same digital learning game focused on decimal number operations and compared them to a non-game control with the same math content. Although one game was designed to appeal to more masculine preferences based on a student preference survey, results indicated that students' learning, engagement, and interest in the games did not vary by narratives or by gender. Students reported greater engagement with the games compared to the non-game, and girls learned more than boys across all three learning conditions. However, the more casual game narrative reduced students' test anxiety and increased their test self-efficacy compared to the non-game, while the battle-oriented narrative did not. Given that girls generally reported lower levels of test self-efficacy and higher test anxiety, these results suggest that digital learning games–particularly those with a more casual game narrative–may be especially effective for girls who experience test anxiety in math.

Keywords: Digital learning games, gender, engagement, self-efficacy, anxiety

1. Introduction

As technology-supported instruction becomes more integrated into students' lives, it is important to understand how different design choices affect students' learning and affective experiences. Digital learning games are one area in which design choices may be especially important and under-studied. Digital learning games can provide instruction and practice with important concepts in an environment that may keep students playing, and thus learning, with more engagement and less anxiety than non-game instruction (Anastasiadis et al., 2018). However, there needs to be more systematic research assessing the effects of different game design choices on affective experiences and learning outcomes, and particularly how learners' identities and preferences might interact with game features to promote learning. In this paper, we present an experimental study designed to assess the interaction of game narrative and gender identity on students' affective experiences and learning outcomes with *Decimal Point*, a math digital learning game focused on decimal number operations, and *Ocean Adventure,* a new narrative version of the decimal number game.

There is a great deal of evidence that digital learning games can promote learning; however, researchers often rely on pre/post learning gains or poorly controlled comparisons to support claims that digital learning games promote better learning and positive affective experiences (Clark et al., 2016; Mayer, 2019; Hussein et al., 2022; McLaren & Nguyen, 2023). Consequently, observed differences in affective and learning outcomes might reflect effects of the game itself or other differences between conditions, such as the use of instructional technology (in a game vs. non-technology-based comparison) or the amount of practice and instruction (in a game vs. "business as usual" comparison). Fewer studies have adopted an experimentally rigorous approach comparing games to matched non-game controls (Hussein et al., 2022) or systematically varying features to assess impact (Mayer, 2019).

Decimal Point has been used to test game features in an experimental classroom setting for the past decade (McLaren, 2024). Controlled studies with *Decimal Point* have indicated that engagement is a significant factor in explaining why games support better learning than similar non-game learning systems. Richey et al. (2021) found that gaming the system—a measure of behavioral disengagement—mediated learning differences between *Decimal Point* and a non-game control with the same instructional content. In other words, the game reduced students' disengaged learning behaviors, which in turn led to greater learning. *Decimal Point* is a useful platform for studying learning and engagement with digital games, and it serves as an opportune context for investigating how different game design choices—in this case, narrative—influence students' affective experiences and learning.

Given engagement's role in learning from digital games (Richey et al., 2021), it is critical to explore how games can be as engaging as possible for a diverse audience of students. Students' game preferences vary widely across individuals (Nguyen et al., 2023; Moradi et al., 2024), and some evidence suggests these preferences vary systematically in relation to aspects of students' identities (Chen et al., 2010; Kim et al., 2021), including students' gender identity (Hawkins et al., 2019; Khan et al, 2017; Yeo et al., 2022). Students of all genders play and enjoy digital games; girls and boys report intending to play digital games at similar frequencies, but their preferred styles of games can vary by gender (Aleksić et al., 2017; Hamari & Keronen, 2017; Romrell, 2014). Similarly, research focused on digital learning games has found that they can promote learning, engagement, and self-efficacy among students of all genders (Sitzmann, 2011; Vogel et al., 2006). Some studies have found differences based on gender, with girls demonstrating greater learning or enjoyment (Khan et al., 2017; Tsai, 2017), while other studies have found no gender differences (Yen et al., 2011; Haruna et al., 2023).

In the case of *Decimal Point*, we have found consistent affective and learning benefits for girls compared to boys (Hou et al., 2020; Nguyen et al., 2022). We propose two possible explanations for this result. First, it is possible that girls find *Decimal Point*, which has an amusement park narrative, to be more engaging than boys do ("the engagement hypothesis"). Prior research has revealed some gender differences in game preferences, with boys tending to prioritize action, challenge, and visual appeal and girls tending to prioritize social interactions and puzzle-style features (Arroyo et al., 2013; Dele-Ajayi, 2018). Nguyen et al. (2023) found that boys and students with more masculine-typed traits reported preferring action, sports, and racing games, and girls and students with more feminine-typed traits said they preferred casual, music, and party games. The casual gameplay and friendly, collaboration-focused narrative of *Decimal Point* aligns more with the reported preferences of girls and students with feminine-typed traits. If girls are more engaged than boys with *Decimal Point* because they find the game design and narrative style more appealing, then changes to *Decimal Point* aimed at increasing its appeal for boys would minimize or eliminate the affective and learning advantages for girls.

Alternatively, it may be that learning math in the playful environment of a digital learning game minimizes the salient math cues of the content and, in turn, reduces math or test anxiety that might otherwise be triggered by math content ("the math anxiety hypothesis). Anxious thoughts consume working memory resources, increasing extraneous cognitive load and leaving fewer cognitive resources available for the mental work of problem solving and learning (e.g., managing the inherent complexity of solving math problems, storing new information in long-term memory, building connections between prior knowledge and new information; Beilock et al., 2010; Maloney et al., 2013; Sweller, 1994). If *Decimal Point*'s playful environment triggers less anxiety than traditional math instruction, it would free up more working memory space for learning about mathematical concepts. Girls tend to experience greater math and test anxiety in math domains and lower self-efficacy (Rocha & Dondio, 2021; Zhang et al., 2024), so minimizing math anxiety might be especially beneficial for girls. If this were the case, we would expect math digital learning games to benefit girls more than boys regardless of any gender differences in narrative appeal, and changes to make *Decimal Point* more appealing to boys should not change the affective and learning benefits for girls.

The current study aims to test these two hypotheses for why *Decimal Point* has led to better learning outcomes for girls—and to explore how game narrative might interact with

gender more generally in digital learning games—by investigating the impact of game narrative and gender on affective and learning outcomes. We created a novel version of the game, *Ocean Adventure*, with a narrative designed to appeal more to boys. We based the design on a game preference survey administered to 333 students (Nguyen et al., 2023); in *Ocean Adventure*, students play the role of a special agent battling pirates and criminals. Results showed *Decimal Point* was equally appealing to boys and girls, but the narrative description for *Ocean Adventure* was significantly more appealing to boys. By comparing two narrative versions of the game, *Decimal Point* and *Ocean Adventure*, we tested the following research questions:

RQ1: Do game narrative and gender interact to predict learning outcomes?

RQ2: Do game narrative and gender interact to predict affective experiences?

The engagement hypothesis predicts that girls benefit more from *Decimal Point* than boys because they find the game more engaging than boys do; therefore, narrative changes that appeal more to boys should reduce the affective and learning advantages for girls in *Ocean Adventure*. Conversely, the math anxiety hypothesis predicts that girls benefit more from *Decimal Point* than boys because they are more affected by math anxiety and low self-efficacy. It predicts that encountering math concepts in a playful game narrative reduces anxiety and potentially boosts self-efficacy compared to a non-game control. According to this hypothesis, the learning advantage for girls should not be changed by a narrative version that is more appealing to boys so long as the game still effectively reduces math saliency and thus anxiety. If boys experience greater affective and learning benefits from *Ocean Advent*ure than girls, it will support the engagement hypothesis. If girls' affective and learning benefits are similar across games, it will support the math anxiety hypothesis, particularly if anxiety is lower or self-efficacy is higher across narratives.

2. Method

Four hundred and eighty-one 5th and 6th grade students from six middle and elementary schools in a northeastern United States metropolitan area participated in the study as part of their normal math instruction. Twenty-one students' data could not be used due to a technical error in the intervention, and an additional 128 students were dropped from analysis for failing to complete one or more test or the instructional materials. The remaining 332 students included 180 who identified as girls and 152 who identified as boys via an open-ended prompt. No students in the final sample reported identifying as non-binary, another gender, or declined to report their gender. The final sample included students who were assigned to play the original digital learning game *Decimal Point* (48 girls, 39 boys), the redesigned digital learning game *Ocean Adventure*, which had narrative that was rated as more appealing among boys (54 girls, 40 boys), or a digital non-game control (78 girls, 73 boys) that included the same instructional materials but without the game narrative or playful design. Students were randomly assigned to a game or non-game condition at the class level to avoid issues with students playing the game and non-game side-by-side; students were randomly assigned to one of the two game conditions at the individual level. Teachers and researchers present in the classroom avoided providing any direct content help to students; instead, they clarified questions about game mechanics, offered encouragement to struggling students, and pointed students to the built-in hints for content questions.

Decimal Point is a digital learning game teaching decimal number operations by counteracting common misconceptions in students' understanding of decimal numbers. Tasks within the game require students to place decimal numbers on a number line or to perform operations on decimal numbers (add, order, sequence, and compare). The game presents students with a sci-fi, light-hearted amusement park narrative: Aliens have come to earth, and they want to learn about decimal numbers while having fun. The player is tasked with guiding the aliens through an amusement park, where each uniquely themed section has decimal content embedded in instructional mini-games. The aliens encourage students with light-hearted positive dialogue, such as "Give this a try! You'll have fun!" Students complete games

in different sections of the amusement park, including "Catch the Ghost" in the Haunted House area of the park and "Escape the Aliens" in the Space Adventure area.

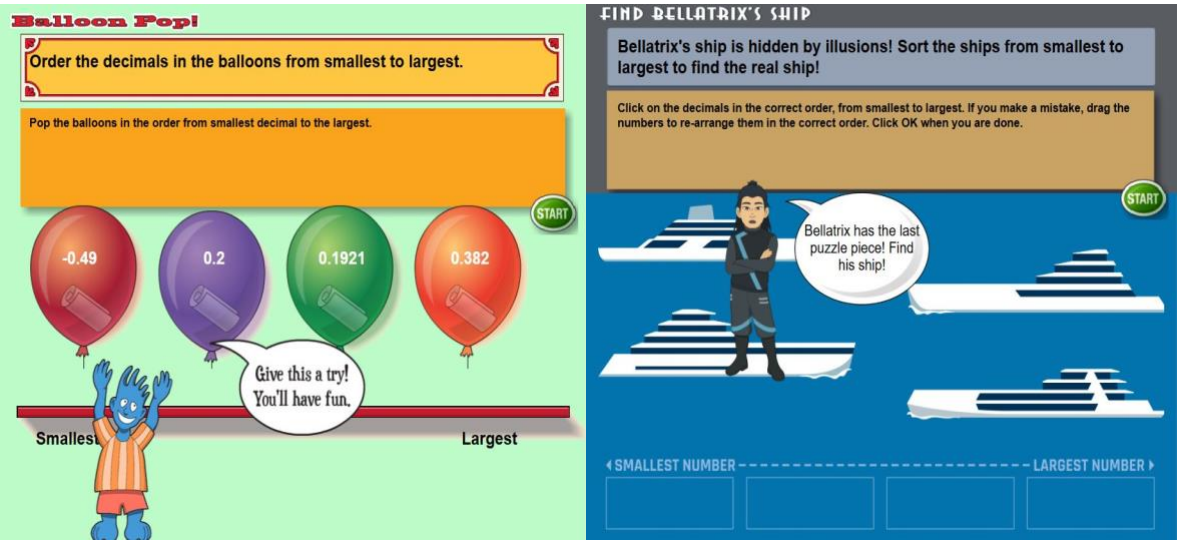


Figure 1. A mini-game within *Decimal Point* (left), with an amusement park narrative, and a mini-game within *Ocean Adventure (*right), with a battle narrative.

Ocean Adventure consists of the same instructional content, but it was redesigned with a new narrative to appeal more to boys based on data collected on students' narrative preferences (Nguyen et al., 2023). Specifically, Nguyen et al. (2023) asked students about the styles of games they preferred; they were also given four game titles with brief narrative summaries and asked to rank their preferences. Boys expressed a preference for action games and overwhelmingly preferred the game titled "War at Sea" with a narrative description that formed the basis of *Ocean Adventure.* Prior research has shown *Decimal Point* typically benefits girls more than boys, and we wanted to experimentally test whether these benefits might be attributed to gender differences in the narrative appeal of the game. In *Ocean Adventure,* students play a special agent who sets out to sea to defeat and capture pirates and criminals. Students receive mission-focused advice in an adventurous, serious tone, such as, "Bellatric has the last puzzle piece! Find his ship!" The game is organized around different pirate hideouts instead of theme park areas, and students complete battles such as "Mirror Maze" in the area focused on pirate villain Bellatrix, or "Break into the Lab" in the area focused on the Doom Island hideout. The non-game control contains the same instructional content as both games but has no game-like narrative. It is a basic, black-and-white digital tutor interface.

Three isomorphic tests were used to assess learning prior to completing the instructional materials (pretest), immediately after completing them (posttest), and approximately one week after completing them (delayed posttest). The tests have been validated in prior studies (Nguyen et al., 2022) and consisted of similar mathematical content to the games, with assessment for near, medium, and far transfer. After completing the learning materials but prior to completing the test, students responded to surveys targeting their interest and engagement in the learning materials and their anxiety and self-efficacy toward the test. We focus on four surveys in this paper: a three-item survey of affective engagement (e.g., "I felt frustrated or annoyed"; Ben-Eliyahu et al., 2018); a three-item survey of situational interest (e.g., "The game/tutor was exciting"; Linnenbrink-Garcia et al., 2010); a four-item survey about test anxiety (e.g., "If I did poorly on this activity, people would look down on me"; Spencer et al., 1999); and a five-item survey about test self-efficacy (e.g., "I could handle this activity"; Spencer et al., 1999). Surveys used a Likert-type scale of 1 (strongly disagree) to 5 (strongly agree).

3. Results

To assess whether students learned from the instructional materials, and whether learning differed based on gender or game condition, we conducted repeated-measures Analyses of Variance (ANOVAs) with time (pretest to posttest, pretest to delayed posttest) as a within-subjects factor and game condition and gender as between-subjects factors. This repeated-measures approach accounted for differences between groups at pretest. Table 1 reports means and standard deviations on all tests. Results indicated a significant effect of time, $F(1, 326) = 196.68$, $p < .001$, $\eta^2_p = .38$, with student performance significantly improving from pretest to posttest. There was also a significant interaction between time and gender, $F(1, 326) = 5.69$, $p = .018$, $\eta^2_p = .017$, with girls improving more than boys from pretest to posttest. There was no interaction of time and game condition, $F(2, 326) = 1.20$, $p = .30$, $\eta^2_p = .007$; and no time, game condition, and gender interaction, $F(2, 326) = 0.06$, $p = .94$, $\eta^2_p < .001$.

Results for the pretest to delayed posttest analysis followed the same pattern, with a significant increase in scores from pretest to delayed posttest, $F(1, 326) = 257.68$, $p < .001$, $\eta^2_p = .44$, and a significant interaction between time and gender, $F(1, 326) = 5.30$, $p = .022$, $\eta^2_p = .016$, with girls improving more than boys from pretest to delayed posttest. There was no interaction of time and game condition, $F(2, 326) = 2.42$, $p = .091$, $\eta^2_p = .015$; and no time, game condition, and gender interaction, $F(2, 326) = 0.82$, $p = .44$, $\eta^2_p = .0005$. These results indicate that students' scores improved from pretest to posttest and from pretest to delayed posttest, and that girls benefitted more than boys across all conditions. There was no evidence that students' learning differed by condition or that girls and boys benefitted differently based on condition.

We conducted a series of 2 (gender) x 3 (game condition) ANOVAs to assess effects on interest, affective engagement, test anxiety, and test self-efficacy. Table 1 reports means by gender and condition. For interest, there was a main effect of condition, $F(5, 315) = 9.758$, $p < .001$, $\eta^2_p = .058$, but no main effect of gender, $F(5, 315) = 0.24$, $p = .63$, $\eta^2_p = .001$, and no significant interaction, $F(5, 315) = 0.19$, $p = .83$, $\eta^2_p = .001$. A posthoc analysis with Bonferroni corrections revealed significantly lower levels of interest only for the non-game compared to *Decimal Point*, $p = .003$, and to *Ocean Adventure*, $p < .001$. This analysis included 321 students (11 missing responses).

For affective engagement, there were main effects of condition, $F(5, 325) = 11.29$, $p < .001$, $\eta^2_p = .065$, and gender, $F(5, 325) = 10.23$, $p = .002$, $\eta^2_p = .031$; girls reported lower engagement than boys. There was no significant interaction, $F(5, 325) = 0.53$, $p = .59$, $\eta^2_p = .003$. A posthoc analysis with Bonferroni corrections revealed significantly lower levels of engagement only for the non-game compared to *Decimal Point*, $p = .002$, and compared to *Ocean Adventure*, $p < .001$. This analysis included 331 students (1 missing).

For test anxiety, there were main effects of condition, $F(5, 314) = 4.68$, $p = .010$, $\eta^2_p = .029$, and gender, $F(5, 314) = 8.60$, $p = .004$, $\eta^2_p = .027$; girls reported higher test anxiety than boys. There was no significant interaction, $F(5, 314) = 0.59$, $p = .55$, $\eta^2_p = .004$. A posthoc analysis with Bonferroni corrections showed significantly higher levels of test anxiety only for the non-game compared to *Decimal Point*, $p = .036$. This analysis included 320 students (12 missing responses).

For test self-efficacy, there were main effects of condition, $F(5, 314) = 4.46$, $p = .012$, $\eta^2_p = .028$, and gender, $F(5, 314) = 5.72$, $p = .017$, $\eta^2_p = .018$, with girls reporting lower self-efficacy than boys. There was no significant interaction, $F(5, 314) = 0.16$, $p = .85$, $\eta^2_p = .001$. A posthoc analysis with Bonferroni corrections showed significantly lower test self-efficacy only for the non-game compared to *Decimal Point*, $p = .028$. This analysis included 320 students (12 missing responses).

Overall, girls reported lower affective engagement and test self-efficacy and higher test anxiety across all conditions. The games, regardless of narrative, tended to increase interest and affective engagement, while *Decimal Point* specifically reduced test anxiety and increased test self-efficacy. Boys and girls were not affected differently by the different game narratives or by the non-game control.

Table 1. *Means and Standard Deviations on Tests and Surveys by Game Condition and Gender*

Cond.	*N*	Pretest	Posttest	Delayed Posttest	Interest	Affective Engmnt.	Test Anxiety	Test self-efficacy
Dec. Point	87	25.1 (10.8)	29.5 (9.6)	30.2 (9.6)	3.1 (1.1)	3.3 (1.0)	2.4 (1.0)	3.7 (0.8)
Girls	48	23.0 (9.7)	28.0 (8.0)	29.5 (8.1)	3.1 (1.0)	3.2 (1.0)	2.5 (1.0)	3.16 (0.9)
Boys	39	27.6 (11.6)	31.2 (11.1)	31.1 (11.2)	3.0 (1.2)	3.6 (1.0)	2.2 (0.9)	3.9 (0.7)
Ocean Adv.	94	22.8 (10.5)	28.7 (9.4)	30.0 (9.2)	3.1 (1.2)	3.4 (0.9)	2.4 (0.9)	3.7 (0.8)
Girls	54	21.2 (9.6)	28.0 (8.6)	28.6 (8.9)	3.1 (1.2)	3.2 (0.9)	2.6 (1.0)	3.6 (0.8)
Boys	40	24.8 (11.4)	29.7 (10.4)	31.8 (9.4)	3.1 (1.2)	3.5 (1.0)	2.1 (0.8)	3.8 (0.9)
Non-game	151	21.8 (9.9)	27.1 (10.1)	28.5 (10.2)	2.6 (1.0)	2.9 (0.9)	2.7 (1.1)	3.5 (0.8)
Girls	78	20.2 (9.6)	26.4 (10.6)	27.8 (10.9)	2.6 (0.9)	2.8 (0.8)	2.8 (1.1)	3.4 (0.8)
Boys	73	23.6 (9.9)	27.9 (9.5)	29.2 (9.3)	2.6 (1.0)	3.0 (1.1)	2.6 (1.1)	3.5 (0.8)

These results indicate that girls reported lower levels of affective engagement and test self-efficacy and higher levels of test anxiety than boys across all conditions. They also reveal that the games, regardless of narrative, tended to increase interest and affective engagement, while *Decimal Point* specifically reduced test anxiety and increased test self-efficacy. There were no interactions between gender and condition, meaning boys and girls were not affected differently by the different game narratives or by the non-game control. Given that the game conditions affected these self-reported measures but did not affect learning outcomes, we examined correlations between the self-reported measures, pre-to-posttest gains, and pre-to-delayed posttest gains. Self-reported measures were generally *not* significantly correlated with learning gains, with the exception of a significant, small correlation between interest and pre-to-posttest gains ($r = .153$, $p = .006$). Other than interest, the remaining self-reported measures all showed significant, medium correlations with each other.

Table 1. *Correlations among test gains and self-reported measures. (*) p < .05.*

	2.	3.	4.	5.	6.
1. Post gain	.718*	.089	.153*	-.054	.023
2. Delayed gain		.063	.083	.010	-.002
3. Affective engagement			.420*	-.477*	.339*
4. Interest				-.029	-.054
5. Test anxiety					-.629*
6. Test self-efficacy					

4. Discussion and Conclusion

In this paper, we tested whether the narrative of a digital learning game would affect learners' affective experiences and learning outcomes. We also investigated whether different narratives would affect boys' and girls' learning and affective experiences differently. In doing so, we tested competing hypotheses that girls' learning advantages from *Decimal Point* might be based on their enjoyment of the game narrative or that they might generally benefit from learning with digital games because they reduce math saliency and, consequently, math and test anxiety. Building on prior research that has shown a consistent learning advantage for girls with *Decimal Point*, we hypothesized that girls would show greater engagement, interest, and learning with *Decimal Point.* If their learning gains were driven primarily by greater engagement compared to boys with the narrative of *Decimal Point,* we hypothesized that girls would not experience the same learning and affective benefits from *Ocean Adventure*, which was designed to appeal more to boys' narrative preferences based on a game preference

survey. Alternatively, we hypothesized that if girls learned more from *Decimal Point* because it reduced math saliency, and therefore girls' negative affect related to math, then girls would show greater learning from both *Decimal Point* and *Ocean Adventure* compared to boys, but not greater learning from the non-game. Additionally, their test anxiety would be lower and their test self-efficacy would be higher in both games compared to the non-game condition.

Results did not provide consistent evidence regarding these hypotheses. In general, there were no interactions between gender and learning condition. This suggests that the different narratives did not affect boys and girls differently, even though the narrative of *Ocean Adventure* was designed specifically to be more appealing boys, based on game preference survey data collected with the target population in a prior study (Nguyen et al., 2023). It may be that the preference differences that emerged when students read brief descriptions of the different narratives were not good predictors of how they would feel when playing the game with the fully developed narrative and graphics. Gender differences related to narrative preference might also be more likely to emerge in games with more immersive narratives. In both games, the narratives provide a framework for organizing learning games, but they lack realism—a positive predictor of narrative transportation (Green, 2010)—and are unlikely to harness the level of focused attention and emotional engagement proposed by narrative transportation theory (Green & Brock, 2000) that can occur when players are transported into the world of a game.

Alternatively, factors other than gender might be better predictors of game preference. Although prior research revealed clear differences in preference based on gender identity (Nguyen et al., 2023), preferences were predicted even more by gender-typed characteristics related to occupational interests, activities, and traits. A measure of gender-typed characteristics would also capture greater levels of gender-related diversity; this may be especially important given the absence of any students identifying outside the gender binary in the open-ended gender question, which is a limitation in our sample. We also failed to replicate prior research showing that girls benefitted more than boys from the game, but not from the non-game control (McLaren et al., 2017). In the current study, girls learned more than boys in both the game and non-game conditions, suggesting the learning benefits were not unique to the game environment. Future work should continue to investigate learning across genders in the game and non-game conditions to resolve these inconsistent findings and determine whether it is the game features or the broader features of computer-supported learning that are helping girls, in particular, to learn these materials.

Results regarding affective experiences generally provided some evidence supporting both an engagement and a math anxiety route to learning with digital games. On the engagement side, all students found the games more engaging and interesting than the non-game control. Narrative did not appear to affect students' engagement or interest. On the other hand, only *Decimal Point* increased students' test self-efficacy and reduced their test anxiety. *Ocean Adventure* results trended in the same directions but were not significantly different from the non-game. It is unclear why *Ocean Adventure* did not have the same effect on test self-efficacy and test anxiety, especially when students found it equally engaging and interesting as *Decimal Point.* One possibility is that the battle narrative of *Ocean Adventure* could create greater affective arousal and tension, which might in turn minimize benefits for test anxiety and test self-efficacy.

Finally, although the games produced significant effects on learners' affective experiences, these affective experiences generally did not predict learning gains. Only interest was significantly correlated with pre-to-posttest gains. This may explain why the game conditions did not lead to significant better learning outcomes compared to the non-game. However, there may still be merit to presenting math practice in ways that are more engaging and interesting, and in ways that lower test anxiety while increasing test self-efficacy. Such changes could affect students' willingness to engage in math practice, their persistence over time, and thus their long-term learning outcomes. If digital learning games lead to lasting changes in students' test anxiety and self-efficacy, it is very likely that those changes would impact learning and test performance over time.

Despite the lack of condition effects on learning, this paper makes several important contributions. Few papers have directly compared the effects of different game narratives in a

math learning game alongside a non-game digital control matched for content. This comparison allowed us to directly compare the effects of different narratives on learners' affective experiences and learning outcomes and has revealed surprising results regarding the effects of different narratives on test anxiety and test self-efficacy. It has also contributed some preliminary evidence that digital learning games might affect students both in terms of their engagement and by reducing testing anxiety and increasing test self-efficacy.

Understanding why *Decimal Point* supports boys and girls differently has both theoretical and practical implications. The comparison of the engagement and math anxiety hypotheses contributes to our theoretical understanding of the different mechanisms through which digital learning games can support students. In particular, relatively little work has investigated whether digital learning games can support learners by reducing math anxiety. We found tentative support for the hypothesis in this study, and future work should continue to investigate the impact of digital learning games on cognitive load and anxiety. In practice, this work could support games as a potential tool for teaching students whose learning is otherwise inhibited by anxiety. Investigating whether narratives choices affect students' learning also has important implications for how digital learning game options are presented to students. For example, if narrative has a minimal effect on learning and affective experience, then designers can focus on creating a single, engaging game rather than creating multiple options for students to choose among. On the other hand, if aligning game features with students' preferences enhances learning and engagement, then creating more game options could improve alignment with individual students' preferences and, in turn, improve student learning outcomes. We believe future work building on comparisons between *Decimal Point* and *Ocean Adventure* will allow us to continue to explore factors that predict students' narrative preferences, how games influence students' affective experiences, and whether those effects ultimately lead to changes in learning behaviors or outcomes.

Acknowledgements

This work was supported by the National Science Foundation Award #DRL2201796. The opinions expressed are those of the authors and do not represent the views of NSF. Thanks to Jimit Bhalani, John Choi, Kevin Dhou, Darlan Santana Farias, Rosta Farzan, Jodi Forlizzi, Craig Ganoe, Rick Henkel, Scott Herbst, Grace Kihumba, Kim Lister, Patrick Bruce Gonçalves McLaren, and Jon Star for important contributions to the development and early experimentation with *Decimal Point*. Thanks also to Rick Henkel, Sarah Di, Em Tyminski, and Kevin He for important contributions to the development and early experimentation with *Ocean Adventure*.

References

Aleksić, V., & Ivanović, M. (2017). Early adolescent gender and multiple intelligences profiles as predictors of digital gameplay preferences. *Croatian Journal of Education: Hrvatski časopis za odgoj i obrazovanje, 19*(3), 697–727.

Anastasiadis, T., Lampropoulos, G., & Siakas, K. (2018). Digital game-based learning and serious games in education. *International Journal of Advances in Scientific Research and Engineering, 4*(12), 139-144.

Arroyo, I., Burleson, W., Tai, M., Muldner, K., & Woolf, B. P. (2013). Gender differences in the use and benefit of advanced learning technologies for mathematics. *Journal of Educational Psychology*, *105*(4), 957-969.

Beilock, S.L., Gunderson, E.A., Ramirez, G. & Levine, S.C. (2010). Female teachers' math anxiety affects girls' math achievement. *Proceedings of the National Academy of Sciences, 1075,* 1060-3.

Ben-Eliyahu, A., Moore, D., Dorph, R., & Schunn, C.D. (2018). Investigating the multidimensionality of engagement: Affective, behavioral, and cognitive engagement across science activities and contexts. *Contemporary Journal of Educational Psychology, 53*, 87–105. doi: 10.1016/j.cedpsych.2018.01.002

Chen, H. P., Lien, C. J., Annetta, L., & Lu, Y. L. (2010). The influence of an educational computer game on children's cultural identities. *Journal of Educational Technology & Society, 13*(1), 94-105.

Clark, D.B., Tanner-Smith, E., & Killingsworth, S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. *Review of Educational Research, 86*(1), 79-122. doi: 10.3102/0034654315582065

Dele-Ajayi, O., Strachan, R., Pickard, A., & Sanderson, J. (2018, October). Designing for All: Exploring Gender Diversity and Engagement with Digital Educational Games by Young People. In *2018 IEEE Frontiers in Education Conference (FIE)* (pp. 1-9). IEEE.

Green, M. C. (2004). Transportation into narrative worlds: The role of prior knowledge and perceived realism. *Discourse Processes*, *38*(2), 247-266.

Green, M. C., & Brock, T. C. (2000). The role of transportation in the persuasiveness of public narratives. *Journal of Personality and Social Psychology*, *79*(5), 701-721.

Hamari, J., & Keronen, L. (2017). Why do people play games? A meta-analysis. *International Journal of Information Management, 37*(3), 125-141.

Haruna, H., Zainuddin, Z., Okoye, K., Mellecker, R. R., Hu, X., Chu, S. K. W., & Hosseini, S. (2023). Improving instruction and sexual health literacy with serious games and gamification interventions: an outlook to students' learning outcomes and gender differences. *Interactive Learning Environments, 31*(4), 2392-2410.

Hawkins, I., Ratan, R., Blair, D., Fordham, J. (2019). The effects of gender role stereotypes in digital learning games on motivation for STEM achievement. *Journal of Science Education and Technology, 28*, 628-637.

Hou, X., Nguyen, H. A., Richey, J. E., & McLaren, B. M. (2020). Exploring how gender and enjoyment impact learning in a digital learning game. In Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part I 21 (pp. 255-268). Springer International Publishing.

Hussein, M.H., Ow, S.H., Elaish, M.M., Jensen, E.O. (2022). Digital game-based learning in K-12 mathematics education: A systematic literature review. *Education and Information Technologies, 27*, 2859–2891 (2022). doi: 10.1007/s10639-021-10721-x

Khan, A., Ahmad, F.H., Malik, M.M. (2017). Use of digital game-based learning and gamification in secondary school science: The effect on student engagement, learning and gender difference. *Education and Information Technologies. 22*(6), 2767–2804.

Kim, D., Li, M. (2021). Digital storytelling: Facilitating learning and identity development. *Journal of Computers in Education, 8*(1), 33-61.

Linnenbrink-Garcia, L., Durik, A.M., Conley, A.M., Barron, K.E., Tauer, J.M., Karabenick, S.A., Harackiewicz, J.M. (2010). Measuring Situational Interest in Academic Domains. *Educational Psychology Measures, 70*(4), 647–671. doi: 10.1177/0013164409355699

Maloney, E. A., Schaeffer, M. W., & Beilock, S. L. (2013). Mathematics anxiety and stereotype threat: shared mechanisms, negative consequences and promising interventions. *Research in Mathematics Education*, *15*(2), 115-128.

Mayer, R.E. (2019). Computer games in education. *Annual Review of Psychology, 70*, 531–49 (2019). doi: 10.1146/annurev-psych-010418-102744

McLaren, B.M., Nguyen, H.A. (2023). Digital learning games in Artificial Intelligence in Education (AIED): A review. In B. du Boulay, A. Mitrovic, & K. Yacef (Eds.), Handbook of Artificial Intelligence in Education. Ch.20. doi: https://doi.org/10.4337/9781800375413.00032

McLaren, B.M. (2024). Decimal Point: A decade of learning science findings with a digital learning game. In: Ilic, P., Casebourne, I., Wegerif, R. (eds) Artificial Intelligence in Education: The Intersection of Technology and Pedagogy. Intelligent Systems Reference Library, vol 261. Springer, Cham. https://doi.org/10.1007/978-3-031-71232-6_9

McLaren, B., Farzan, R., Adams, D., Mayer, R., & Forlizzi, J. (2017). Uncovering gender and problem difficulty effects in learning with an educational game. In *Artificial Intelligence in Education: 18th International Conference, AIED 2017*, Wuhan, China, June 28–July 1, 2017, Proceedings 18 (pp. 540-543). Springer International Publishing.

Nguyen, H. A., Else-Quest, N., Richey, J. E., Hammer, J., Di, S., & McLaren, B. M. (2023). Gender Differences in Learning Game Preferences: Results Using a Multi-dimensional Gender Framework. In *Proceedings of the International Conference on Artificial Intelligence in Education,* pp. 553–564.

Nguyen, H. A., Hou, X., Richey, J. E., & McLaren, B. M. (2022). The impact of gender in learning with games: A consistent effect in a math learning game. *International Journal of Game-Based Learning (IJGBL), 12*(1), 1-29.

Mahdavinasab, Y., Mehrvarz, M. (2024). Adapting player styles to enhance learning and retention in high school science education. *E-Learning and Digital Media*, 20427530241239409.

Richey, J. E., Zhang, J., Das, R., Andres-Bray, J. M., Scruggs, R., Mogessie, M., Baker, R., & McLaren, B. M. (2021). Gaming and confrustion explain learning advantages for a math digital learning game. In *International Conference on Artificial Intelligence in Education* (pp. 342-355). Cham: Springer International Publishing.

Rocha, M., Dondio, P. (2021). Effects of a videogame in math performance and anxiety in primary school. *International Journal of Serious Games, 8*(3), 45-70.

Romrell, D. (2014). Gender and gaming: A literature review. In *Annual Meeting of the AECT International Convention*, pp. 170-182

Sitzmann, T. (2011). A meta-analytic examination of the instructional effectiveness of computer-based simulation games. *Personnel Psychology, 64*(2), 489-528.

Spencer, S.J., Steele, C.M., Quinn, D.M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35*(1), 4-28. doi: 10.1006/jesp.1998.1373

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295-312.

Yen, J. C., Wang, J., Chen, I. J. (2011). Gender differences in mobile game-based learning to promote intrinsic motivation. *Recent Researches in Computer Science, 15*, 279-284.

Yeo, J. H., Cho, I. H., Hwang, G. H., Yang, H. H. (2022). Impact of gender and prior knowledge on learning performance and motivation in a digital game-based learning biology course. *Educational Technology Research and Development, 70*(3), 989-1008.

Vogel, J. J., Vogel, D. S., Cannon-Bowers, J., Bowers, C. A., Muse, K., Wright, M. (2006). Computer gaming and interactive simulations for learning: A meta-analysis. *Journal of Educational Computing Research, 34*(3), 229-243.

Tsai, F. H. (2017). An investigation of gender differences in a game-based learning environment with different game modes. *Eurasia Journal of Mathematics, Science and Technology Education, 13*(7), 3209-3226.

Zhang, L., Lei, Y., Pelton, T., Pelton, L. F., Shang, J. (2024). An exploration of gendered differences in cognitive, motivational and emotional aspects of game-based math learning. *Journal of Computer Assisted Learning, 40*(6), 2633-2649.

From Roof Cracks to Wound Healing: How Structure-Behavior-Function Analogy Shapes Students' Reasoning in Complex Biology System

Meera PAWAR, Rajashri PRIYADARSHINI & Sahana MURTHY
Indian Institute of Technology Bombay, India
***meerapawar19@iitb.ac.in**

Abstract: Complex biological systems, such as Human Body Systems, require a holistic understanding of how structures interact dynamically to produce functional outcomes. The Structure–Behavior–Function (SBF) framework offers a lens for capturing such reasoning by emphasizing the interconnected roles of system components, their mechanisms, and purposes. To support students in developing this form of reasoning, this study employed an SBF-based analogical story that compared the biological process of wound healing to the repair of a cracked roof. This study explores how an SBF-based analogical story, of repairing a cracked roof comparing with wound healing shapes students' reasoning about biological systems. Thematic analysis revealed key reasoning strategies students employed while interacting with the analogy. They mapped characters and roles across domains, gradually shifting from surface level similarities to identifying deeper behavioral and functional equivalences. Students compared the systems at both whole and part levels, demonstrating their ability to reason about interdependent system elements. To organize their understanding of biological processes, many developed schemas, such as distinguishing between temporary and permanent repairs. Furthermore, students were able to identify and critically evaluate the limitations of the given SBF-based analogical story, reflecting a deeper, more critical engagement with the analogy rather than relying on straightforward one-to-one mapping.

Keywords: Structure-Behavior-Function Framework, SBF- Analogy, Complex Biology Systems,

1. Introduction

Students at the secondary level frequently encounter difficulties in making sense of complex biological systems. These systems involve multiple interacting components, dynamic behaviors, and emergent properties that operate across different spatial and temporal scales (Hmelo & Azevedo, 2006). However, many students tend to rely on linear, part-by-part reasoning, which makes it difficult for them to grasp the interconnected and dynamic nature of biological phenomena and systems such as that of the human body (Assaraf, et. al, 2013). Building on this, a further challenge lies in helping students make sense of unobservable, micro-level phenomena and understanding how these relate to structure, behavior, and function (Hemlo & Pfeffer, 2004). Students have shown difficulties in reasoning about complex biological systems particularly about underlying mechanisms. One way to promote a meaningful understanding of such systems is through the Structure-Behavior-Function (SBF) framework, which helps students in systematically organizing their thinking about complex systems (Hmelo, 2007).

According to the constructivist learning theory, learning is supported when students can connect new concepts to familiar knowledge (Brock, 2024). In this context, analogies are commonly used as scaffolds to make abstract and unfamiliar ideas more accessible through

acquiring relational schema (Gray & Holyoak, 2021). Analogical tasks have been proved to foster students' reasoning skills as well as their critical and creative thinking skills. Despite the widespread use of analogies in biology education, most emphasize only on structural and functional aspects, typically through visual or verbal formats (Rule & Furletti, 2004).

Combining pedagogical potential of analogies and explanatory power of SBF framework, we developed a technology-enhanced learning environment called BioAnalogica, featuring analogical stories designed explicitly around the SBF framework. These stories are intentionally aligned with biological target concepts such as, incorporating SBF elements along with narrative structures. The targeted complex biological system for this article is the human immune system. SBF based analogical story of roof crack repair was developed for the process of wound healing.

While our previous work focused on the design and learning outcomes of this intervention, the present study focuses on students reasoning process. This research article aims to investigate how analogical stories designed on SBF framework shapes students' reasoning about complex biological systems. With specific question, "*How does an SBF-based analogical story facilitate students' reasoning about the biological process of wound healing?"* By exploring how learners mapped elements of the analogical story to biological concept, we aim to uncover the underlying cognitive processes that shaped their SBF reasoning. we conducted a study in which 25 students engaged with the SBF based intervention. This paper contributes by highlighting key processes that students used the story to reason about wound healing.

2. Related Work

2.1 SBF Reasoning in Complex Biological Systems

Knowledge about complex biological systems requires students to comprehend different levels of organization: macro (system, organ) and micro (cell, molecules) and the interactions between them (Snapir, et. al 2017). Understanding how the human body functions requires reasoning about the roles of individual parts (such as cells, tissues, and organs, for example, white blood cells, lymph nodes, and skin in the immune system), the mechanisms by which they interact (such as how white blood cells identify and attack pathogens through processes like recognition, signaling, and phagocytosis), and the purposes they serve (for instance, protecting the body from infections and maintaining internal balance or homeostasis) (Assaraf, et. al, 2013). Structure-Bheaviour-Function framework gives a way to reason about such complex systems in terms of its structures, their functions and how they achieve its function (Hmelo & Azevedo, 2006).

Studies show its effectiveness when paired with complementary pedagogical approaches, for instance, Jordan et al. (2013) found that combining SBF ontology with guided questions and simulations enhanced students' understanding of system dynamics and knowledge transfer. Hmelo and Pfeffer (2004) observed that experts rely on integrated SBF representations, unlike novices who focus on static features, highlighting the framework's role in fostering expert-like reasoning. Additionally, SBF-based instruction combined with augmented reality has improved students' ability to visualize and connect multilevel biological elements (Gregorčič, & Torkar, 2022). These findings highlights SBF's potential in designing and analyzing effective biology learning interventions.

2.2 Analogies in learning Complex Biology Systems

Dedre Gentner's Structure-Mapping Theory (SMT) provides a fundamental framework for understanding how analogies support learning by helping students connect unfamiliar concepts (targets) to familiar ones (sources) (Gentner, 1996; Glynn, et al., 2007). A core principle of SMT is that learners tend to transfer coherent systems of relationships rather than isolated features, allowing the structural essence of the base domain to be meaningfully

applied to the target. This relational focus is especially critical for reasoning about complex systems, where underlying mechanisms and interactions are more important than superficial similarities.

Salih (2010) highlighted the potential of analogical to improve their reasoning abilities as well as their critical and creative thinking. Rule and Furletti (2004) introduced “form and function analogy object boxes” to teach high school students about human body systems by pairing physical objects (e.g., smoke detector, camera lens) with body parts (e.g., nose, eye) based on shared structural features and functions. Through hands-on activities, students explored similarities, differences, and generated new analogies. While effective in highlighting structure and function, the approach largely overlooked the behavioral dimension, critical for understanding dynamic system interactions. When it comes to teaching the human immune system, many existing interventions, such as game-based learning (e.g., *Humonology*) (Cheng, et. al, 2014), simulations, and classroom activities, frequently rely on metaphors and analogies drawn from defense, combat, or warfare. These analogies tend to be ad hoc and are rarely grounded in the SBF framework. In addition to lack in robust design of these analogies, a persistent challenge in analogical reasoning lies in supporting students to attend to these deeper structural and behavioral correspondences. Younger learners, in particular, often struggle due to focus on perceptual features (Richland et al., 2006).

One of the study shows, Han and Kim (2019) used a physical analogical model, comprising a syringe and balloon to explore students' SBF reasoning about the respiratory system. Their study demonstrated that such models can support a shift from understanding surface-level functions to reasoning about underlying behaviors and mechanisms. However, this analogical model lacked a system level perspective, dynamic progression and multi-level organization important for holistic view of the system. These studies emphasize the value of bringing in SBF reasoning in supporting students’ understanding of human body systems and highlights the instructional value of thoughtfully designed analogies to scaffold for multi-level, systemic understanding. Supporting students in developing SBF reasoning requires explicit scaffolding that helps them identify system structures, understand dynamic behaviors, and relate these to functional outcomes. Designing learning experiences, such analogies around the SBF framework can offer this structured support, guiding students to reason more effectively about complex systems (Pawar, 2024).

In the context of understanding complex biological systems, there has been limited exploration of how students develop Structure–Behavior–Function (SBF) reasoning. This study reorients the focus toward the process of SBF reasoning itself, using analogies not as the primary object of analysis but as a medium to support and reveal how learners identify structures, interpret dynamic behaviors, and relate them to system functions. By emphasizing SBF reasoning and examining how it unfolds through analogical support, we gain a deeper understanding of how students make sense of biological complexity and develop coherent system-level explanations.

3. Method

This research uses a qualitative research approach to understand how this SBF based analogical story helps students in reasoning about complex biology process. This is based on our previous work in which we designed and employed an SBF-based analogical intervention (BioAnalogica) involving roof crack repair story to explain the wound healing process carefully designed on the basis of SBF framework (Human Immune System) to secondary school students.

3.1 Participants

All participants were recruited through teachers, coaching centers, social media, and word-of-mouth during school vacations, limiting access to regular classrooms. Total 27 Grade 9-10 CBSE students from Maharashtra voluntarily joined, this grade was selected for their familiarity with basic biology (e.g., blood clotting, immunity) but not detailed wound healing mechanisms.

They had no formal exposure to S-B-F framework before the intervention. Before participation, informed consent was obtained from parents, and verbal assent was collected from students at the time of the study.

3.2 Materials

The instructional material used was animated story based videos of wound healing and roof crack analogy. Story content was reviewed by biology education researcher and two high school biology teacher for scientific accuracy and age-appropriateness. Figure 1 shows a few of the scenes from these both the videos.

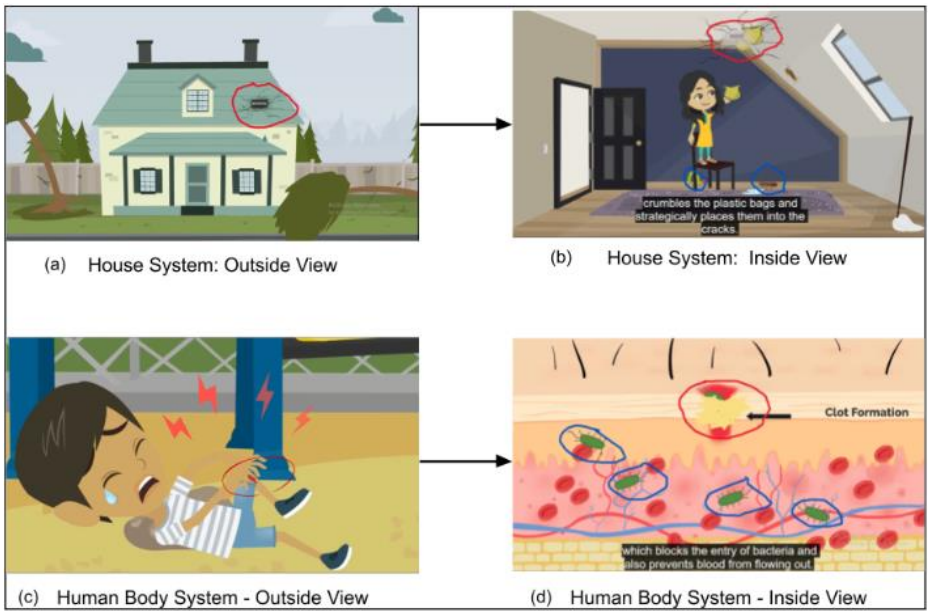


Figure 1: Visual representation of the SBF-based Roof Crack story (a, b) and wound healing process (c, d).

All the four visuals are as follows:

(a) Outside view of the analogy (Systemic view): A cracked roof on a house represents structural damage visible at the system level, analogous to a skin wound
(b) Inside view of the analogy (component view): Red circles indicate plastic patches applied internally to stop leakage (representing blood clotting), while blue circles depict insects entering through the crack (representing microbial invasion).
(c) Outside view of the human body (systemic view): A visible wound on the skin corresponds to the cracked roof, highlighting the structure and damage in the target domain.
(d) Inside view of the body (component view): Red circles show clot formation, and blue circles indicate bacterial entry at the wound site.

These visuals illustrate the Structure–Behavior–Function (SBF) mapping between the analogy and the target system. Images (a) and (c) represent the systemic views (house and human body). In contrast, images (b) and (d) provide internal components and their roles. In visuals (b) and (d), the red circles indicate plastic bags (Structure) used to block the roof crack (Function), which are analogous to platelets (Structure) responsible for blood clotting (Function) in the body. The blue circles mark the structural mapping of insects entering the damaged house and bacteria invading the wounded body. The behaviors are mapped and showed as continuous actions and interactions between all the components.

3.3 Data Collection

Each student participated in the study individually. The procedure began with a pre-test followed by watching an SBF based analogical story for approximately 5-10 minutes, then students interacted with the conceptual story of wound healing. This story included embedded reflection spot questions placed at logical transition points. These questions prompted students to map elements from the analogical story to the target concept and provide reasoning for the connections they made. After completing the concept story and reflection questions, students took a post-test. Total interaction with the intervention lasted for 30-45 minutes. This was immediately followed by a one to one semi-structured interviews lasting 15-20 minutes per session, to explore their reasoning in depth.

This study primarily focuses on qualitative analysis of student interviews to investigate how SBF-based analogies supported students' understanding of the wound healing process. While pre- and post-tests were conducted to assess conceptual gains, those results have been analyzed and published (Pawar, 2024) in a separate study and are not the focus of this paper. Sample interview questions included questions about the usefulness of the SBF story and the reasoning processes and approaches with it, such as, “How did the given analogical story help you understand wound healing?” and “Could you think of the differences between both the stories?”. These interviews provided rich insights into students’ reasoning strategies, analogical thinking, and metacognitive reflections, offering a deeper understanding of how analogies support conceptual learning in complex biological systems.

3.4 Data Analysis

All interview recordings were transcribed, and any responses in vernacular languages were translated into English for consistency. Students’ responses were segmented at the sentence level, which served as the unit of analysis. A qualitative thematic analysis approach was used with mixed coding (Clarke & Braun, (2017). The coding scheme was theoretically motivated by the Structure-Mapping Theory and the Structure-Behavior-Function (SBF) framework, which guided the identification of patterns in the student’s responses such as Mapping phase of their reasoning. However, the majority of the codes emerged inductively from the data, capturing students' context-specific reasoning patterns. The initial phase of analysis involved one researcher (First author) conducting open coding of 30 responses. These codes were then reviewed by a second expert researcher (Second author) with experience in conceptual learning and thematic analysis. An iterative consensus-building process followed with 87.1% agreement, during which codes were refined, categorized and merged based on emerging patterns and meaning. These codes captured both students’ analogical reasoning processes and their evolving understanding of the biological system. The codes were then grouped into four broader themes, which reflected the key mechanisms through which SBF based analogical stories helped students reasoning about the wound healing process. All the themes are explained with examples in detail in the findings section.

4. Findings

While mapping is a foundational strategy in analogical learning, the way students engaged in this process varied considerably in both depth and direction, revealing different underlying processes. These mapping strategies reflected how students interpret, prioritize, and adapt SBF elements from the analogy to the target biological system. The following subthemes detail the different approaches students used during this mapping process.

4.1 Students used chronological alignment, to understand system behavior, followed by role-based mapping to link structure and function

Students, while mapping an analogical story to the concept, mapped similarities in different ways. Learners aligned elements of the analogy and target concept by mapping structures,

assigning roles to characters, and recognizing sequence and system-level interactions, reflecting emerging SBF reasoning. For example, student, S1, when asked how the analogical story helped, he mentioned that, "*It helped me understand **as first Pooja put plastic bags to stop the rain coming inside**. That was the same as blood clot as platelets are put there. Platelets we can say that **platelets are the same as plastic bags**. Then we saw Anil removing all the insects just like how white blood cells did, then we saw the **plumber putting cement paddy, same as fibroplasm (Fibroblast) which put collagen***." This excerpt shows a stepwise mapping strategy, where the student first traces the chronological structure of the analogy, starting with the first event (*first Pooja put plastic bags- Character in the Analogy*), and then gradually connects each character (*plastic*) to its corresponding function (stop the rain) and then map these with the wound healing (*That was the same as blood clot as platelets are put there).* The transition from saying "the same as" to "we can say" reflected an emerging confidence in the relational alignment, moving from simply noticing resemblance to asserting equivalence from SBF perspectives. The character-role mapping represents structure-function mapping, and the overall chronological sequence describes the behavior of the system of how wound healing occurs stepwise.

Another mechanism observed was when students, while identifying the similarities, tried creating a hypothetical situation and extending the understanding of both analogy and transferring the same understanding to the concept. For example, S12 thought of a hypothetical situation of "what if?" By saying, "….*No, for the understanding it is fine, but I think when you put the plastic bag, **so then again when the rain comes, again the water would come out of it,** they might come out. But I **think it matches this thing very well, because even after like again, if we get wounded on the same place**, so that layer that has formed of platelet, that would get broken. So it is totally similar…".* This indicates that the student engaged in a deeper level of reasoning by constructing a hypothetical situation of dynamic behavior of the system, how the structures (Plastic bags) over time and changes their function (Again water would come) under different conditions within the analogue domain and extending it to the target concept. By imagining what might happen if the plastic bag were removed and it rained again, the student reasoned about the failure and recurrence of the damage, and mapped it to the biological event of re-injury and clot breakdown in the same wound area. This reflective process highlights not only structural and functional parallels but also an emerging awareness of dynamic, cyclical processes within the system, strengthening the mapping between the analogy and the target concept.

In another example, S11 remarked, "..*That story (Analogical story) gave me an idea of what is going to happen one after other, and understood steps …Because if, for example, the WBCs start working even before the bacteria arrive, and then the platelets come in afterward, **it won't make sense and the process won't work properly**."* Here, the student recognized the interconnectedness of the system components (Platelets, WBC) and the importance of their coordinated functioning. By emphasizing that a mismatch in the sequence would disrupt the overall mechanism, the student demonstrated an understanding of the system as a dynamic whole, where the correct temporal alignment of parts is essential for the intended function to emerge.

Through this manner, SBF-based analogical stories supported students' understanding of complex systems by helping them grasp system behavior through chronological and interconnected processes, learn structure-function pairing via character-role mapping, and learn dynamic behavior by reasoning through hypothetical scenarios.

4.2 Students shifted between macro-level and micro-level perspectives to map systemic functions and local mechanisms across the analogy and biological process

Another important strategy employed by students while engaging with SBF-based analogical stories involved changing their levels of perspectives, either adopting a zoomed-out, (macro) view of the system as a whole or a zoomed-in, (micro) view focusing on local parts and interactions. This form of perspective shifting revealed the flexibility with which students

approached the analogical mapping between the biological process of wound healing and the SBF-based analogical story. In the zoomed-out approach, students mapped broad systemic roles and processes between the analogy and the biological target. For example, when a student remarked, "*. It was explained that* ***the house is related to our body***,." they reflected a zoomed-out perspective, where the student approached the analogy by comparing entire systems rather than isolated parts or events. Instead of focusing on specific one-to-one mappings, such as platelets with plastic sheets or fibroblasts with cement, the student recognized a holistic parallel between the house and the human body.

Consider another example excerpt by S17, *"Because the roof , .. I mean,* ***the house was shown to be related to our body.*** *When the roof got damaged,* ***both water and insects entered the house****, ….and then later, Anil's father came".* In this example, we see that students first emphasize on the similarities of both analog (House) and target (Body) as a whole part as a system. An analogical system of the house was mapped to the body system, indicating a zoom- out perspective. After discussing the body and the house as whole systems, she shifted perspectives to examine what happens inside (*both water and insects entered*), focusing on how damage allows insects to enter and how they are removed (*Anil's father came)*, demonstrating a move from a zoomed-out to a zoomed-in view.

Hence, in this way, the SBF based analogical story prompted students to shift perspectives first zooming out to understand the system as a whole and then zooming in to explore the interconnected parts and their functions aligned. This perspective-shifting reveals the flexibility with which students approached the analogical mapping between the SBF-based analogical story the and biological process of wound healing.

4.3 Through analogical abstraction, students constructed simplified models of wound healing, organizing their understanding around structural roles, behavioral sequences, and functional goals

Another key theme that emerged was how students were able to make sense of the wound healing system by using the SBF-based analogical story. Through this, students formed a core idea or mental model of how the wound healing system works. For example, some students came up with a damage-repair model, where they understood the wound as a form of damage and the body's response as a repair process. Others talked about a temporary-permanent repair model, where the body initiates a quick, temporary fix followed by a more stable, long-term solution. These student-generated models suggest that learners were engaged in abstraction processes, moving beyond surface details to identify in-depth relationships. Rather than focusing only on visual or literal similarities, they recognized key structural, behavioral, and functional elements from the analogy and mapped them to the biological system. This shift indicates that the analogical story supported the understanding of how the system works both as a whole and at the level of individual parts. For example, when asked about the similarity between roof and skin, S13 explained "…*Because like after the* ***roof got damaged l****ike after that only the insects came inside and they tried to* ***fix it.*** *Okay. So like when* ***we get a wound,*** *the bacteria also comes inside and the cells, different cells* ***try to repair that****……".*

In this response, the student engages in a clear process of abstraction, moving beyond literal description to extract underlying relationships between the analogy and the biological system. She begins by identifying the structure-function relationship, linking the roof in the story to the skin in the human body. Both are recognized as protective barriers, and she understands that their function is compromised when physically damaged. Student then shifts to 1describing the behavior of the system: once the protective structure is breached, external agents (insects in the house and bacteria in the body) enter through the damaged area. This mapping reflects her ability to track cause-and-effect sequences, recognizing that the system's behavior changes dynamically after damage occurs. Finally, she refers to function again when she says "different cells try to repair that," pointing to the goal-directed actions that follow the intrusion. This shows her understanding that specific components of the biological system (e.g., immune cells and fibroblasts) are mobilized to restore the function of the damaged tissue. Here, student is not simply listing one-to-one correspondences between story elements

and body parts. Instead, she is abstracting higher-level concepts such as protection, disruption, invasion, and repair, and using them to reason about the wound healing process

Another instance of abstraction and model development was observed when, students internalized the analogy as a "temporary and permanent repair" schema, in which early actions were seen as provisional fixes and later actions as more stable, long-term fixes. Instance of this can be seen in the excerpt of S11, "…*Then, Pooja realized that insects were entering from there, so* ***she stuck plastic bags over it, similar to how platelets arrive and temporarily cover*** *the wound…".* In this response, the student exhibits an understanding of the wound healing process by distinguishing between immediate and long-term responses within the system. Student begins by describing the role of plastic bags and platelets as providing a temporary seal, indicating an awareness of the system's initial, rapid response to damage aimed at preventing further damage to the system. Indicating understanding of dynamic and temporal (Temporary at the beginning) behavior of structural components (Platelets) while fulfilling their functions of repairing (Covering the wound). Then the student transitioned to long term fixes and identified structures plumber in analogy story and corresponding cells fibroblast in the body by saying "*After that, they called a plumber, and* ***he applied cement putty to provide a permanent fix*** *to the roof, restoring it to its original state. In biology****, this is like fibroblasts coming in and secreting a protein called collagen, which provides a permanent solution to the skin***". This shows student's ability to move beyond surface-level analogies and recognize the progression from a temporary function to permanent functions aligning with the underlying biological mechanism. By articulating this two-phase process, the student engages in abstraction, mapping not just events, but functional roles and temporal dynamics within the system. This explanation reflects an emerging systems perspective, in which structural components interact sequentially to restore working of the system.

Hence, in this way, the SBF-based analogical story facilitated students' abstraction of the wound healing process by supporting model construction of system repair. It enabled learners to reason about the system's sequential behavior and goal-directed functions, highlighting coordination among components during healing.

4.4 Students identified and evaluated mismatches in character roles, system dynamics, and underlying mechanisms between analog and target

In this theme, we discuss how students demonstrated not only the ability to map similarities between the analogy and the biological system but also a critical awareness of where the analogy breaks down. Their reasoning extended beyond simple alignment to include identifying mismatches and limitations across structural, behavioral, and functional elements. This reflective engagement with the analogy indicates a process of evaluative thinking, where students monitored and questioned the accuracy of the mappings.

For instance, in one excerpt, a student S7 curious and grapples with the role of the character "Pooja," initially unsure how she fits into the wound healing process. Asks, "*Okay, so* ***what is Pooja supposed to be?*** *Hmm…",* This question emerges from the student's discomfort in not being able to locate a direct counterpart (Structure) for the character of Pooja in the wound healing concept, indicating active engagement with the structural elements of the analogy. The student kept thinking back and forth between the analogy and the target concept, and eventually attempts to bridge the gap by proposing a possible match for Pooja, by saying "***Maybe she could be the blood which brings the platelets***" this attempt to find a possible match goes beyond structural alignment and extends to the functional role, emphasizing the dynamic action of "bringing the platelets". In doing so, the student refines the conceptual model of wound healing through reflective mapping of characters. This shows students transferred understanding of structures, their functions and place of action from analogical story to concept of wound healing.

By noticing these discrepancies and, dissimilarities, students demonstrated a critical understanding toward the analogy itself, rather than accepting it as it is. This reflective engagement suggests that their analogical reasoning was not limited to straightforward one-to-one mappings, but instead involved iterative evaluation, questioning, and refinement. The

SBF based analogical story supported this process by guiding students to examine how structures, their dynamic behaviors, and resulting functions correspond or fail to correspond across the analogical and biological domains. Such reflective awareness is especially important in learning complex biological systems, as it helps students recognize limitations in analogies and reduces the risk of misconceptions.

5. Discussion and Conclusion

The SBF-based analogical story played a central role in shaping how students' reasoning process to understand the complex biological process of wound healing. Rather than simply helping them find surface-level similarities by embedding structural elements, behaviors, and functions into a familiar, everyday narrative, the SBF aligned story provided a concrete way for students to follow the sequence of events, understand elements involved at each stage, and understand why those steps mattered. Rather than viewing the process as a disconnected list of biological terms, students were able to make sense of the sequence and purpose behind each event. They reasoned about who performs which function at each stage (like platelets being like plastic bags, or fibroblasts sealing the wound permanently), which shows they were thinking about the roles and functions of each part. This helped them see how the system behaves over time, not just what things are, but how the whole thing works as a process. This kind of mapping supports what Gentner (1983) refers to as systematicity, where learners are more likely to understand complex systems when they can align not just individual elements, but the relationships and sequences among them.

Another way in which SBF- aligned analogical story supported reasoning about wound healing process was by helping students in developing mental models or core ideas of the process by taking processes “Damage - Repair” and “Temporary - Permanent fix” ideas. These models were conceptually understood with the help of the narrative used in the stories. This supports what Gentner and Stevens (1983) and Hmelo-Silver et al. (2007) describe as analogical model-building, where learners use analogies not only to understand isolated ideas, but to build coherent mental models of a system. Another way this study extends previous research is in how students approached the mapping process itself. Rather than treating it as a fixed one-to-one match, many students revised their initial mappings based on deeper functional or behavioral understanding. For instance, when a student initially aligned a character with white blood cells but later re-evaluated and changed their mapping, they were not just engaging in analogical refinement, they were actively refining their understanding of how structures relate to functions within the system (Holyoak, 2012).

This study provides insight into how SBF-based analogical stories help address the challenge of learners being distracted by irrelevant details in analogies. Unlike the findings by Richland et al. (2006), who reported that younger learners often become distracted by such details, students in our study acknowledged these aspects but were still able to engage effectively with the core ideas of the analogy. Notably, students did not accept the analogy uncritically; instead, they questioned its applicability—asking, for instance, “What is Pooja supposed to be?”. Such comments indicate that students were engaged in evaluation rather than merely searching for matches, consistent with Holyoak and Gentner’s framework. This engagement suggests that students were not only using the analogy as a learning scaffold but also thinking about how (Behavior) and why (Purpose/Function) it works (or doesn’t), a reflective process that signals deeper conceptual understanding and analogical maturity, especially for this age group.

This suggests that reflective reasoning about complex process through analogy can naturally emerge when analogies are designed to highlight structure, behavior, and function explicitly. Because the SBF framework makes these relationships clear, students could more readily detect misalignments at the level of function, behavior, or structure, whether within components or the entire system. These findings extend prior research in important ways. Previous studies have noted that analogies can sometimes lead to misconceptions if students overextend mappings (Glynn et al., 2007). However, in this study, students themselves identified potential misconceptions.

The strategies students used, such as chronological alignment, role-based mapping, and analogical evaluation, were strongly informed by the SBF-based design of the analogical story. By explicitly embedding structures, behaviors, and functions within a familiar narrative, the story made the biological system's dynamics more accessible. The sequence of events mirrored wound healing, encouraging students to reason about processes over time, while assigning roles to characters (e.g., Anil as WBC) supported functional understanding. The SBF framework's focus on causality and interconnections helped students build coherent mental models, refine their mappings, and critically assess the analogy's limits indicating that their reasoning was shaped by both the structure of the analogy and the underlying SBF design principles.

A key limitation of this research is that it focused on short-term reasoning gains without examining long-term retention or transfer of SBF-based understanding. Additionally, the study relied solely on interview responses; incorporating multimodal data such as drawings or think-a louds could provide richer insights into students' reasoning. Future research could explore how students apply these analogical reasoning strategies in new biological contexts over time to assess deeper conceptual understanding.

Acknowledgements

We would like to thank the schools and coaching centre for allowing their students to participate in the research and Coaching centre and IITB for granting access to their premises to conduct this research. And my colleague Alekh for their support during the research.

References

Assaraf, O. B. Z., Dodick, J., & Tripto, J. (2013). High school students' understanding of the human body system. Research in Science Education, 43, 33-56.

Brock, R. (2024). Constructivist approaches to learning and teaching science. In Learning to Teach Science in the Secondary School (pp. 131-141). Routledge.

Cheng, M. T., Su, T., Huang, W. Y., & Chen, J. H. (2014). An educational game for learning human immunology: What do students learn and how do they perceive?. *British Journal of Educational Technology*, *45*(5), 820-833.

Clarke, V., & Braun, V. (2017). Thematic analysis. *The journal of positive psychology*, *12*(3), 297-298.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive science*, *7*(2), 155-170.

Glynn, S. (2007). The teaching-with-analogies model. *PUB TYPE*, 195.

Gregorčič, T., & Torkar, G. (2022). Using the structure-behavior-function model in conjunction with augmented reality helps students understand the complexity of the circulatory system. *Advances in Physiology Education*, *46*(3), 367-374.

Gray, M. E., & Holyoak, K. J. (2021). Teaching by analogy: From theory to practice. *Mind, Brain, and Education*, *15*(3), 250-263.

Han, M., & Kim, H. B. (2019). Elementary students' modeling using analogy models to reveal the hidden mechanism of the human respiratory system. *International Journal of Science and Mathematics Education*, *17*, 923-942.

Hmelo-Silver, C. E., & Azevedo, R. (2006). Understanding complex systems: Some core challenges. *The Journal of the learning sciences*, *15*(1), 53-61.

Hmelo-Silver, C. E., Marathe, S., & Liu, L. (2007). Fish swim, rocks sit, and lungs breathe: Expert-novice understanding of complex systems. *The Journal of the Learning Sciences*, *16*(3), 307-331.

PAWAR, M., VASUDEVAN, S., & MURTHY, S. (2024, November). BioAnalogica: SBF-Based Analogical Stories to Enhance Understanding of Complex Biological Processes. In *International Conference on Computers in Education*.

Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of experimental child psychology*, *94*(3), 249-273.

Salih, M. (2010). Developing Thinking Skills in Malaysian Science Students via an Analogical Task. *Journal of Science and Mathematics Education in Southeast Asia*, *33*(1), 110-128.

Integrating Generative AI into Escape Room-Based Learning: Effects on Nursing Students' Clinical Skills, Flow Experience, and Self-Efficacy

Intan SETIANI[a*], Ching Yi CHANG[b], Riska SAPUTRA[c] & Jie Chi YANG[a,d]
[a]*Graduate Institute of Network Learning Technology, National Central University, Taiwan*
[b]*School of Nursing, College of Nursing, Taipei Medical University, Taiwan*
[c]*Graduate Institute of Digital Learning and Education, National Taiwan University of Science and Technology, Taiwan*
[d]*Research Center for Science and Technology for Learning, National Central University, Taiwan*
*intansetianis97@gmail.com

Abstract: This study investigates the impact of integrating Generative Artificial Intelligence (GenAI) into escape room-based learning on nursing students' clinical skills, flow experience, and self-efficacy. While escape rooms have proven effective in promoting collaboration and critical thinking in nursing education, their linear structure may limit personalized support. Conversely, GenAI tools such as ChatGPT offer adaptive feedback and individualized guidance but lack the immersive engagement of game-based learning. To address this, we designed a quasi-experimental study involving 47 first-year nursing students from a university in Taiwan. Participants were assigned to either an experimental group (AI-assisted escape room) or a control group (video-based learning). A four weeks, both groups engaged in sessions focused on breastfeeding skills, with pre- and post-intervention assessments administered using validated instruments. The results revealed statistically significant improvements in clinical performance, self-efficacy, and flow experience for the AI-assisted group compared to the control. These findings suggest that combining GenAI with escape room strategies creates a synergistic learning environment that enhances both cognitive and affective outcomes in nursing education. The study highlights the potential of AI-supported Escape Room learning to foster student-centered, competency-driven instructional models.

Keywords: Generative AI (GenAI), Escape Room, Clinical Skills, Nursing Education, Self-Efficacy, Flow Experience.

1. Introduction

In recent years, nursing education has undergone a paradigm shift towards more interactive, student-centred, and competency-based pedagogies. Among these, educational escape rooms have emerged as innovative game-based learning environments that promote collaboration, critical thinking, and motivation (Arrue et al., 2025; González-de laTorre, 2024). Unlike traditional case-based methods, escape rooms immerse learners in problem-solving scenarios under time constraints, fostering engagement and enhancing declarative knowledge. Nevertheless, such high-pressure environments can also elevate anxiety levels, and the direct correlation between enjoyment and knowledge acquisition remains ambiguous (Arrue et al., 2025).

Generative AI tools like ChatGPT are increasingly used in health education, offering adaptive, on-demand support for clarifying concepts, providing personalized feedback, and enhancing decision-making (El Arab et al., 2025; Han et al, 2025). Despite their growing utility, concerns regarding over-reliance, trust, and ethical considerations remain prevalent (Ramírez-Baraldes et al., 2025).

However, the integration of GenAI with escape room-based learning is still largely unexplored. Escape rooms foster motivation and knowledge acquisition but have fixed

learning paths, while GenAI offers personalized, real-time support without the immersive collaboration of physical spaces. Combining them could create a synergistic learning framework addressing cognitive, affective, and metacognitive aspects. Flow experience and self-efficacy are key to educational success, and enhancing AI-related self-efficacy requires not only AI literacy but also positive attitudes, practical application, and genuine interest in the technology. Bewersdorff et al. (2025) point out that to boost students' belief in their ability to use AI (self-efficacy), educational strategies must not only cover AI literacy but also concentrate on nurturing their attitudes towards AI, their practical application of it, and their overall interest in the technology.

This study aims to investigate the effects of integrating generative artificial intelligence (GenAI) into escape room-based learning on nursing students' clinical skills performance, flow experience, and self-efficacy, by comparing outcomes between an experimental group (escape room with GenAI support) and a control group (video-based learning). A quantitative approach using validated instruments will be employed to measure differences between the two groups.

- To what extent does the integration of GenAI into escape room-based learning enhance clinical skills performance among nursing students compared to video-based learning?
- Does the use of GenAI within an escape room setting improve nursing students' self-efficacy more significantly than traditional video-based instruction?
- How does the flow experience differ between nursing students engaged in GenAI-assisted escape room learning and those in a video-based learning environment?

2. Methods

2.1 Participants

The study involved 47 first-year nursing students from a university in northern Taiwan, randomly assigned to an experimental group (n=24, AI-assisted escape room) or a control group (n=23, video-based learning). Ethical approval was granted by the university's Institutional Review Board (IRB). Using a quasi-experimental design within a structured four-week instructional period, pretests, posttests, and questionnaires were administered one week before and after the intervention. Both groups completed these assessments simultaneously and within the same intervals.

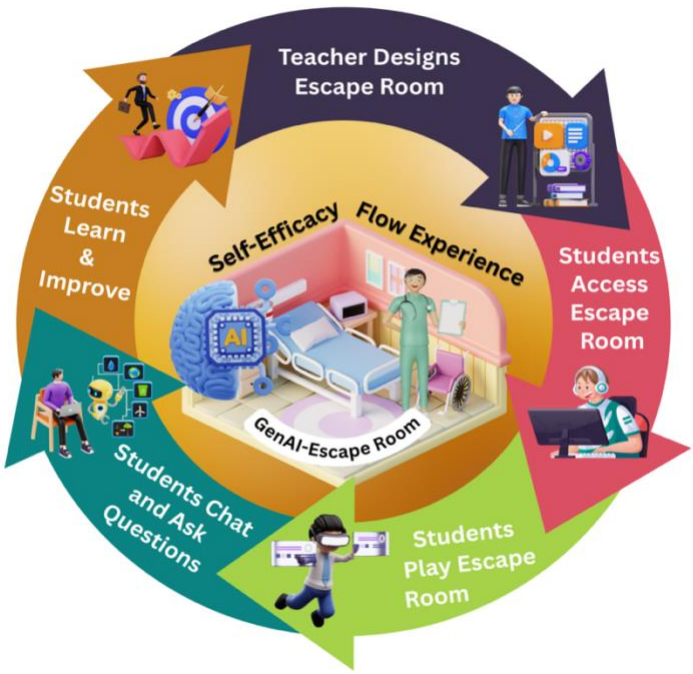


Figure 1. Experimental Design

2.2 Experimental Procedure

This study implemented a generative AI-assisted training intervention in a maternity nursing course through an Escape Room environment on the ZEP platform. The process began with the teacher designing a tailored escape room for nursing education, followed by students participating in the challenge. Generative AI was integrated to allow real-time interaction, enabling students to ask questions and receive clarification while solving puzzles and tasks. This combination aimed to enhance clinical skills, promote deep engagement or "flow," and increase self-efficacy. Before the intervention, the syllabus and learning goals were shared, and all participants completed a pretest along with self-efficacy and flow questionnaires. Participants were then randomly assigned to either the experimental group (AI-assisted Escape Room) or the control group (video-based learning).

The experimental group engaged in the AI-supported Escape Room, designed to foster learner independence and active participation, while the control group used traditional classroom methods, watching instructional videos and completing worksheets without AI or interactive tools. Both groups met once a week for four weeks, with each session lasting 100 minutes. After completing the sessions, participants took a post-test and repeated the questionnaires to measure any changes in clinical skills, self-efficacy, and flow experience. The study's main objective was to determine whether AI-supported Escape Rooms could yield greater improvements compared to traditional instruction.

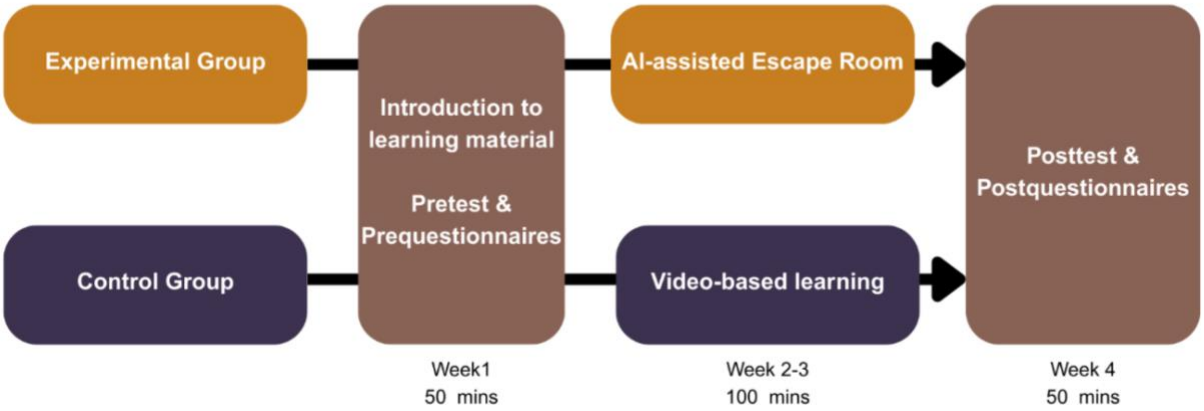


Figure 2. Experimental Procedure

2.3 Instruments and Data Analysis

In this study, we used 20-item multiple-choice clinical skills test, an eight-item self-efficacy questionnaire adapted from Pintrich et al. (1991), and an eight-item flow experience scale by Pearce et al. (2005). All items were rated on a 5-point Likert scale and demonstrated good reliability (Cronbach's alpha = .79–.88). Data analysis was performed in IBM SPSS 27 using descriptive statistics and nonparametric tests due to the small sample size. Mann–Whitney U tests compared pretest and posttest results for both groups. While no significant differences appeared at pretest, posttest results showed significant improvements in the experimental group's clinical skills, self-efficacy, and flow, confirming the effectiveness of the AI-assisted Escape Room intervention.

3. Results

3.1 Clinical Skills Achievement

The descriptive statistics showed that the AI-assisted Escape Room group had a pretest mean score of 78.96 (SD = 10.68) and a posttest mean score of 84.68 (SD = 8.43), while the video-based learning group had similar central tendencies. To compare the two groups, a Mann–Whitney U test was performed. The Mann–Whitney U test on the pretest scores for learning

achievement showed that there was no significant difference between the AI-assisted Escape Room (Mean rank = 25.58) and the video-based learning group (Mean rank = 22.35), U = 238.00, Z = −.821, p = .412. This indicates that both groups had comparable levels of achievement prior to the intervention. However, the posttest results revealed a statistically significant difference in learning achievement between the two groups. The AI-assisted Escape Room group (Mean rank = 29.85) significantly outperformed the video-based learning group (Mean rank = 17.89), U = 135.50, Z = −3.04, p < .01. These findings suggest that the use of AI-assisted Escape Room led to a significant improvement in students' post-intervention learning performance compared to the conventional video-based approach.

Table 1. *Mann-Whitney U test results for Clinical skills*

Variable	Groups	*N*	Mean rank	*Sum of ranks*	*U*	Z value
Clinical skills	Experimental group	24	29.85	716.50	135	-3.04**
	Control group	23	17.89	411.50		

p < .01**

3.2 Self-efficacy

Table 2. *Mann-Whitney U test results for self-efficacy*

Variable	Groups	*N*	Mean rank	*Sum of ranks*	*U*	Z value
Self-efficacy	Experimental group	24	30.83	740.00	112	-3.70***
	Control group	23	16.87	388.00		

p < .001***

The Mann–Whitney U test on the pretest scores for self-efficacy showed no significant difference between the AI-assisted Escape Room group (Mean rank = 24.94) and the video-based learning group (Mean rank = 23.02), U = 253.50, Z = −0.53, p = .599, indicating comparable self-efficacy levels before the intervention. In contrast, the posttest results revealed a statistically significant difference, with the AI-assisted Escape Room group (Mean rank = 30.83) scoring higher than the video-based learning group (Mean rank = 16.87), U = 112.00, Z = −3.70, p < .001. These findings demonstrate that integrating AI with Escape Room-based learning significantly enhanced students' self-efficacy in the experimental group.

3.3 Flow-experience

The Mann–Whitney U test on pretest scores for flow experience indicated no significant difference between the AI-assisted Escape Room group (Mean rank = 23.35) and the video-based learning group (Mean rank = 24.67), U = 260.50, Z = −0.36, p = .723, confirming that both groups had comparable flow levels before the intervention. In contrast, posttest results revealed a significant difference, with the AI-assisted Escape Room group (Mean rank = 31.69) reporting higher flow experience than the video-based group (Mean rank = 15.98), U = 91.50, Z = −4.07, p < .001. These findings suggest that the AI-assisted Escape Room strategy effectively enhanced students' flow experience during learning.

Table 3. *Mann-Whitney U test results for flow-experience*

Variable	Groups	*N*	Mean rank	*Sum of ranks*	*U*	Z value
Flow-experience	Experimental group	24	31.70	760.50	91	-4.07***
	Control group	23	15.98	367.50		

p < .001***

4. Discussion and Conclusions

This study concludes that integrating Generative AI into Escape Room-based learning significantly enhances nursing students' clinical skills, flow experience, and self-efficacy. By

merging immersive, collaborative, real-time feedback, the AI-assisted Escape Room fosters deeper cognitive engagement and greater learner autonomy than traditional video-based instruction. This student-centered, competency-based approach supports both skill acquisition and confidence in clinical decision-making. The significant gains in clinical performance among students in the experimental group extend the findings of Aktaş et al. (2024) and Gómez-Urquiza (2022), demonstrating that Generative AI amplifies the educational benefits of Escape Rooms through personalized scaffolding and timely clarification.

Building on this, a particularly noteworthy finding is the enhancement of both self-efficacy and flow experience among students in the AI-assisted group. Generative AI tools, such as ChatGPT, enabled students to receive immediate, context-sensitive responses to their queries during problem-solving tasks, contributing to a sense of control, competence, and emotional engagement. This suggests that the integration of AI can help learners navigate complex clinical scenarios with increased confidence, particularly when traditional game-based learning lacks individualized support. These results align with the research of Bouriami et al. (2025), Chang et al. (2024), Chen et al. (2025), and Marinho et al. (2025), who highlight the importance of personalization, cognitive-emotional balance, and immersive challenge-based learning in developing learner motivation. Furthermore, the AI layer addresses a key limitation of traditional Escape Rooms—their fixed structure—by introducing adaptive pathways that maintain difficulty while preventing cognitive overload (Hong et al., 2025; Wang et al., 2025). Thus, the integration of GenAI does not diminish student agency; rather, it enhances it by offering just-in-time support that preserves critical thinking and collaboration.

Despite these promising outcomes, several considerations must be addressed for future implementation. A key concern is the potential over-reliance on AI, which may diminish students' independent reasoning and self-regulation over time. As cautioned by Fan et al. (2025) and Gerdes et al. (2025), treating AI as a shortcut rather than a supportive tool could undermine critical thinking skills. Thus, instructional design should balance assistance with cognitive challenge to ensure sustained student engagement, while educators play a vital role in fostering ethical and responsible AI use. Although the quasi-experimental design and validated instruments strengthen the study's internal validity, its generalizability is limited by the single-institution context and modest sample size. Future research should involve more diverse populations and examine long-term effects, such as knowledge retention, real-world skill transfer, and ethical AI literacy development.

Based on these results, nursing educators and curriculum developers are encouraged to incorporate Generative AI tools into interactive learning models such as Escape Rooms. Well-designed AI-enhanced environments can support both personalized learning and collaborative skill development. Future research should investigate how AI-assisted Escape Room influences clinical performance across diverse topics and settings, while also examining learners' ethical awareness, emotional resilience, and sustained engagement. This will help ensure that AI-integrated pedagogies enhance learning outcomes while preserving professional competence, autonomy, and ethical technology use in nursing education.

References

Aktaş , N., Baykara, Z. G., & Öztürk, D. (2024). The effect of education provided with the escape room game on nursing students' learning of parenteral drug administration. *Nurse Education in Practice, 80*(104133), 104133. https://doi.org/10.1016/j.nepr.2024.104133

Arrue, M., Suárez, N., Ugartemendia-Yerobi, M., & Babarro, I. (2025). Let's play and learn: Educational escape room to improve mental health knowledge in undergraduate nursing students. *Nurse Education Today*, *144*, 106453. https://doi.org/10.1016/j.nedt.2024.106453

Bewersdorff, A., Hornberger, M., Nerdel, C., & Schiff, D. (2025). AI advocates and cautious critics: How AI attitudes, AI interest, use of AI, and AI literacy build university students' AI self-efficacy. *Computers and Education: Artificial Intelligence,* (100340), 100340. https://doi.org/10.1016/j.caeai.2024.100340

Bouriami, A., Boussaa, S., & El Adib, A. R. (2025). Active Strategies in the AI Era: Can ChatGPT Help or Hinder Nursing Education?. In *Advanced Nursing Practices for Clinical Excellence* (pp. 237-258). IGI Global Scientific Publishing. https://www.igi-global.com/chapter/active-strategies-in-the-ai-era/373781

Chang, C.-Y., Yang, C.-L., Jen, H.-J., Ogata, H., & Hwang, G.-H. (2024). Facilitating nursing and health education by incorporating ChatGPT into learning designs. *Educational Technology & Society, 27*(1), 215–230. https://doi.org/10.30191/ETS.202401_27(1).TP02

Chen, H.-L., Yohannes, A., & Hung, N.-L. (2025). Effects of escape room game-based civics education on junior high school students' learning motivation, critical thinking and flow experience. *British Journal of Educational Technology: Journal of the Council for Educational Technology, 56*(3), 1170–1190. https://doi.org/10.1111/bjet.13519

El Arab, R. A., Al Moosa, O. A., Abuadas, F. H., & Somerville, J. (2025). The role of AI in nursing education and practice: Umbrella review. *Journal of Medical Internet Research, 27*, e69881. https://doi.org/10.2196/69881

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology: Journal of the Council for Educational Technology, 56*(2), 489–530. https://doi.org/10.1111/bjet.13544

Gerdes, M., Bayne, A., Henry, K., Ludwig, B., Stephenson, L., Vance, A., ... & Winston, S. (2025). Emerging Artificial Intelligence-Based Pedagogies in Didactic Nursing Education: A Scoping Review. *Nurse Educator*, 10-1097. https://doi.org/10.1097/nne.0000000000001746

Gómez-Urquiza, J. L., Hueso-Montoro, C., Correa-Rodríguez, M., Suleiman-Martos, N., Martos-Cabrera, M. B., Gómez-Salgado, J., & Albendín-García, L. (2022). Nursing students' experience using an escape room for training clinical skills and competencies on emergency care: A qualitative observational study. *Medicine, 101*(30), e30004. https://doi.org/10.1097/MD.0000000000030004

González-de la Torre, H., Hernández-De Luis, M.-N., Mies-Padilla, S., Camacho-Bejarano, R., Verdú-Soriano, J., & Rodríguez-Suárez, C.-A. (2024). Effectiveness of 'escape room' Educational Technology in nurses' education: A systematic review. *Nursing Reports, 14*(2), 1193–1211. https://doi.org/10.3390/nursrep14020091

Han, S., Kang, H. S., Gimber, P., & Lim, S. (2025). Nursing students' perceptions and use of generative artificial intelligence in nursing education. *Nursing Reports, 15*(2). https://doi.org/10.3390/nursrep15020068

Hong, J.-C., Tai, T.-Y., & Liang, F.-Y. (2025). English vocabulary learning through gamification with Tic-Tac-Toe in Flippity-Connecto: The prediction of gameplay self-efficacy to anxiety, interest and flow experience. *British Journal of Educational Technology: Journal of the Council for Educational Technology*. https://doi.org/10.1111/bjet.13577

Marinho, A., Bittencourt, I. I., Dermeval, D., Santos, J., Challco, G., & Reis, M. (2025). How does the mechanics of competition and collaboration impact on student's flow experience in a gamified educational environment? An experimental study in higher education (under review). *Technology Knowledge and Learning, 30*(1), 399–424. https://doi.org/10.1007/s10758-024-09796-5

Pearce, J. (2005). Engaging the learner: how can the flow experience support e-learning?. In *E-learn: world conference on E-learning in corporate, government, healthcare, and higher education* (pp. 2288-2295). Association for the Advancement of Computing in Education (AACE).

Pintrich, P. R., Smith, D., Garcia, T., & McKeachie, W. J. (1991). *A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ).* National Center for Research to Improve Post-Secondary Teaching and Learning. https://eric.ed.gov/?id=ED338122

Ramírez-Baraldes, E. Ia, García-Gutiérrez, D., & García-Salido, C. (2025). Artificial intelligence in nursing: New opportunities and challenges. *European Journal of Education, 60*(1). https://doi.org/10.1111/ejed.70033

Wang, M., Zhang, D., Zhu, J., & Gu, H. (2025). Effects of incorporating a large language model-based adaptive mechanism into contextual games on students' academic performance, flow experience, cognitive load and behavioral patterns. *Journal of Educational Computing Research, 63*(3), 662–694. https://doi.org/10.1177/07356331251321719

Design and Testing of a VR Escape Room Game for Philippine Martial Law History

Eric Cesar E. VIDAL Jr.[a*], Nicko R. CALUYA[b], Johanna Marion R. TORRES[a], Jesus Alvaro C. PATO[a], and Kenneth King L. KO[a]
[a] *Virtual/Augmented/Mixed Reality Laboratory, Ateneo de Manila University, Philippines*
[b] *Visual Information Engineering Laboratory, Ritsumeikan University, Japan*
*evidal@ateneo.edu

Abstract: This paper presents a Virtual Reality Educational Escape Room game where players learn about the highly divisive Martial Law period in Philippine history. We describe the game's general design and the results of a user test to evaluate the game in terms of VR presence, immersion, and overall usability.

Keywords: Virtual Reality, Escape Rooms, Usability Testing, Philippine History

1. Introduction and Related Work

In recent years, Virtual Reality (VR) has been heavily employed by museums for historical learning applications, such as guided virtual tourism and simulated interaction with preserved cultural artifacts (Song & Evans, 2024). However, VR experiences that discuss serious subject matter are infrequent and provide restricted user interaction. Studies on controversial historical VR experiences such as *Wiktoria 1920* (Kazlauskaitė, 2023) and *The Anne Frank VR House* (Mulders et al., 2025) found that these experiences typically use a linear narrative progression and users do not perform any intentional actions besides moving to fixed spots, handling objects at predetermined times, etc., which may limit their potential for affective learning.

Several non-VR studies tackle the politically divisive topic of Philippine Martial Law in game-based learning (GBL) contexts. These include table-top role-playing games (De Vera, 2018) and semester-long game activities monitored via learning management systems (Despabiladeras et al., 2024); unfortunately, these require active facilitation by the instructor.

An alternative GBL method is Educational Escape Rooms or EERs, where the player solves learning puzzles in a locked room (Veldkamp et al., 2020). EERs may possibly improve learning over traditional methods through more direct bodily engagement, with tasks that are highly integrated with learning outcomes (Skulmowski & Rey, 2018). However, EERs still do require extensive facilitation by moderators and large floor spaces and equipment.

Thus, our team has developed a system for creating single-player VR EERs. The VR modality helps reduce the learning activity's facilitation requirements, as VR escape rooms require less floor space and setup time over successive play sessions compared to physical escape rooms. Our EER system is called Ateneo Virtual Reality Escape (AVRE), which runs on top of the open-source Godot Engine (Vidal et al., 2024). AVRE was used to develop a serious Philippine Martial Law History game, entitled *Heritage Hero: Secrets of the 'Golden Era'*, in collaboration with the Ateneo Martial Law Museum and Library. Here we describe the game's design and the results of a preliminary user test to show its potential effectiveness.

2. Game Design

The intended audience of the VR game are Filipino young adults (ages 15-25) with an overall play time of 30 to 60 minutes. The escape room is themed around an abandoned mansion belonging to a crony of former president and dictator Ferdinand E. Marcos, Sr., who declared Martial Law in 1972 and was ousted during the 1986 EDSA Revolution. The mode of the escape room is to solve the mystery of the abandoned mansion, which is divided into three areas, each of which tackles a specific aspect of the Martial Law regime (see Figure 1):

- The ***study*** focuses on the *press freedom* aspect. The player pieces together several propaganda materials about the era, as well as operates a confiscated printing press machine that symbolizes the control that the Marcoses had over mass media.
- The ***kitchen*** focuses on the *public health* aspect. This area allows the player to bake Nutribun bread, a key component of the public health program of the historical period.
- The ***bedroom*** focuses on the *economic development* aspect. The player deciphers confidential documents such as architectural construction blueprints and contracts relating to the infrastructure projects of the Marcoses, which were marred by controversies.



Figure 1. Main game areas. Left-to-right: study, kitchen, bedroom.

The game progression is not strictly linear and requires the player to revisit areas to solve puzzles fully. Unlike passive VR games, AVRE implements an object interaction system using hand controller tracking; the player has to perform twisting, levering, connecting, hitting, and wiping motions, directly linking player movement with intentional game actions to increase immersion and integration with learning objectives (Lee et al., 2023; Vidal et al., 2024).

These interactions are also supplemented with two special items: a *journal*, which tracks the player's current progress and gives an outline of the remaining tasks without directly giving hints to the player; and a *flashlight-radio*, operated via a manual crank (popular during the era), which can receive a "radio broadcast" that is, in reality, a hint to help solve the player's current puzzle. These additions serve to replace the moderator role in a traditional escape room game, allowing players to finish the game without outside help.

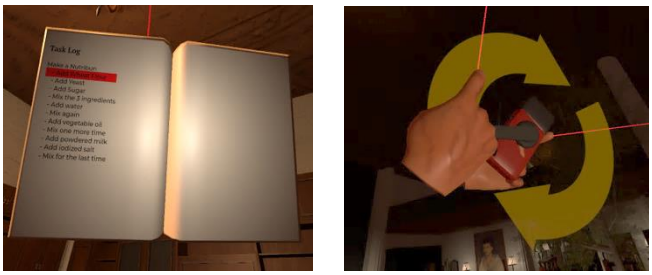


Figure 2. Special equipment to aid player progress: journal (left) and flashlight-radio (right).

3. Evaluation and Results

Fourteen (14) participants joined the user testing and were composed of Ateneo de Manila University students with an average age of 19.86 (± 0.86), all healthy adults who voluntarily provided consent and feedback. Self-assessment on a scale of 0 to 6 reveals that the participants have limited familiarity with VR games (2.64) and the Meta Quest 3 headset used for the study (1.57) but were more familiar with escape room games (3.00) and digital adventure games (4.29). The participants were also asked to answer an iGroup Presence Questionnaire (IPQ) (Tran et al., 2024) and a subset of the PLAY Game Heuristics (Desurvire & Wiberg, 2009). See Figure 3 for a visualization of the questionnaire results.

Following the recommendation of Tran et al. (2024) to correct each IPQ subscale from [0,6] to [-3,3], the ranking was determined to be ***low*** (-0.14) for *experienced realism*, ***moderate*** (0.54) for *involvement*, ***high*** (1.34) for *spatial presence*, and ***exceptional*** (2.28) for *general presence*. We note that the participants may have correlated the questionnaire items about experienced realism with higher-fidelity art assets (e.g., high texture and animation quality), which were admittedly limited by our team's art production capabilities. Despite this, the

general presence is still high, suggesting that participants were aware that they were immersed in a virtual world and would willingly suspend disbelief.

The scores of the PLAY heuristics were also treated similarly to the IPQ: since items in the questionnaire had both positive and negative statements, responses for the latter statements were reverse scaled; then, scores were averaged according to category. The game was evaluated with an exceptional level of immersion (5.35), high gameplay value (4.80), and high usability (3.92).

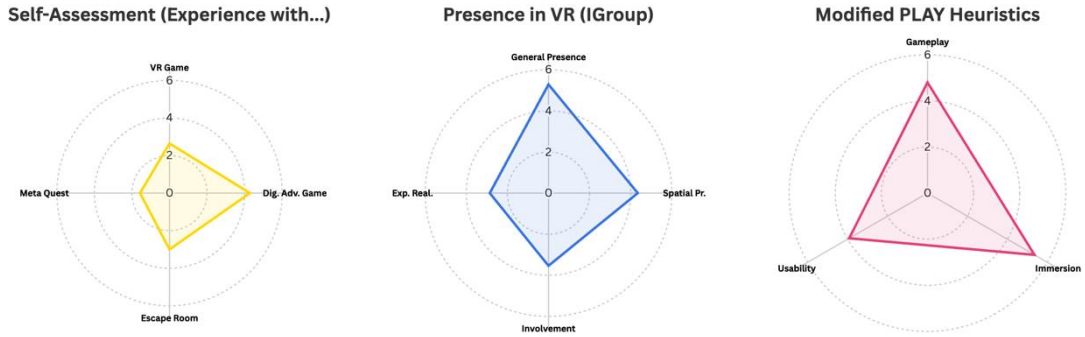


Figure 3. Summary of self-assessment (left), IPQ (middle) and PLAY heuristics (right).

The participants also gave open-ended evaluation responses. Some usability issues, such as players experiencing discomfort and nausea due to the VR modality, were brought up. Despite this, participants reported that the kitchen and bedroom areas are engaging due to real-world familiarity with the escape room puzzles therein, i.e., cooking and architecture. A majority of participants found the scope of the learning content sufficient, but some found it lacking and suggested other controversial topics to add (such as censorship and bribery during the featured historical period). These topics were included in the final game as optional quests.

References

De Vera, I. (2018). *Lipunan - An Educational Role Playing Game*. Ateneo Martial Law Museum & Library. https://learn.martiallawmuseum.ph/magturo/lipunan/

Despabiladeras, G., Lasquite, E., Atayde, E. J., Olvis, K., & Serafico-Reyes, N. M. (2024). Exploring game-Based Learning (GBL) through learning management systems in teaching Batas Militar (Martial Law) to elementary students' achievement. *TALA: An Online Journal of History*, *7*(2), 100–133.

Desurvire, H., & Wiberg, C. (2009). Game Usability Heuristics (PLAY) for Evaluating and Designing Better Games: The Next Iteration. In A. Ant Ozok & P. Zaphiris (Eds.), *Online Communities and Social Computing* (pp. 557–566). Springer Berlin Heidelberg.

Kazlauskaitė, R. (2023). Virtual reality as a technology of memory: Immersive presence in Polish politics of memory. *Memory, Mind & Media*, *2*(e7), e7.

Lee, J. D. L., Restoles, D. M. D., & Vidal, E. C. E. (2023). A Virtual Reality Object Interaction System with Complex Hand Interactions. *Lecture Notes in Computer Science*, *14058*, 122–134.

Mulders, M., Träg, K. H., Kaninski, L., Kirner, L., & Kerres, M. (2025). Virtual reality and affective learning in commemorative history teaching: effects of immersive technology and generative learning activities. *Journal of Research on Technology in Education*, 1–21.

Skulmowski, A., & Rey, G. D. (2018). Embodied learning: introducing a taxonomy based on bodily engagement and task integration. *Cognitive Research: Principles and Implications*, *3*(1), 6.

Song, Z., & Evans, L. (2024). The museum of digital things: extended reality and museum practices. *Frontiers in Virtual Reality*, *5*, 1396280.

Tran, T. Q., Langlotz, T., Young, J., Schubert, T. W., & Regenbrecht, H. (2024). Classifying presence scores: Insights and analysis from two decades of the Igroup Presence Questionnaire (IPQ). *ACM Transactions on Computer-Human Interaction: A Publication of the Association for Computing Machinery*. https://doi.org/10.1145/3689046

Veldkamp, A., van de Grint, L., Knippels, M.-C. P. J., & van Joolingen, W. R. (2020). Escape education: A systematic review on escape rooms in education. *Educational Research Review*, *31*, 100364.

Vidal, E. C. J. E., Caluya, N. R., Lee, J. D. L., Ko, K. K. L., Jocson, J. L. O., & Go, G. J. N. N. (2024, November 25). Improving Engagement in Museums Through Virtual Reality Educational Escape Rooms (VREER): A Framework and Usability Study. *International Conference on Computers in Education*. https://doi.org/10.58459/icce.2024.5017

C6
Technology Enhanced Language Learnin (TELL)

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Exploring EFL Learners' Perceptions of Generative AI through the Technology Acceptance Model: Insights from an AI Literacy Program

Michiyo ODA[a*], Ryo SAITO[b] & Tatsuya HORITA[c]
[a] *Faculty of Foreign Studies, Reitaku University, Japan*
[b] *Graduate School of Information Sciences, Tohoku University, Japan*
[c] *Graduate School of Education, Tokyo Gakugei University, Japan*
* mioda@reitaku-u.ac.jp

Abstract: This study investigates how English as a Foreign Language (EFL) learners perceive the use of generative AI (GenAI) tools in their language learning, using the Technology Acceptance Model (TAM) as an analytical framework. Drawing on final reports from 41 Japanese university students who participated in a four-week AI literacy program, this study qualitatively analyzed their reflections based on TAM constructs. The analysis revealed that students experienced a wide range of benefits from using GenAI, including support for all four English skills. Notably, 83% of the sentences in the behavioral intention category expressed a positive intention to use GenAI. In the attitude construct, critical perspectives were predominant, including concerns about over-reliance, ethical risks, and privacy. Furthermore, two subcategories, namely *Selective Use of GenAI* and *Expectations for Technological Advancement*, emerged beyond the TAM framework. These reflect learners' critical engagement and developing agency in deciding when and how to integrate GenAI tools into their studies. Additionally, students' prior frequency of GenAI use influenced their behavioral intentions. Despite perceiving GenAI as effective during the program, those with less prior use often showed lower willingness to continue using it, citing psychological concerns or feelings of guilt. This study underscores the importance of pedagogical interventions to help learners use GenAI ethically and effectively, with an understanding of its underlying mechanisms. It also proposes extending the TAM framework to incorporate critical and reflective dimensions of technology use.

Keywords: Generative AI, AI literacy, technology acceptance model, EFL learning

1. Introduction

Generative artificial intelligence (hereafter, GenAI), exemplified by ChatGPT, has rapidly gained widespread adoption in recent years due to advances in large language models (LLMs). Built on LLMs and powered by natural language processing technologies, these tools are capable of performing sophisticated language tasks such as translation, summarization, and text revision. In the first 10 weeks following the release of ChatGPT, educators were initially reluctant to adopt the technology, primarily due to concerns about potential student misuse (Alm & Ohashi, 2024). However, research on AI in language teaching and learning has since accelerated, with publications in this field skyrocketing. In this context, Law (2024) conducted a literature review of studies on GenAI in language education published between 2017 and 2023 and found that 98% of the reviewed publications appeared in 2023. The review also indicated that most studies focused on writing and that GenAI has been found to positively influence both learners' psychological states, such as motivation, and their language learning outcomes, owing to its immediacy and personalized support.

Although most studies reviewed by Law (2024) exhibited positive attitudes toward the implementation of GenAI, recent studies have shown that both teachers and learners express mixed feelings. While they acknowledged the benefits of GenAI, including prompt feedback,

individualized learning support, and help with brainstorming, they also expressed concerns about its potential risks related to academic integrity, accuracy, privacy, and ethics (Chan & Hu, 2023; Chukhno, 2024). Given the widespread impact of GenAI on education, Trust et al. (2023) argued that the most effective way to address its misuse is not through exclusion but through thoughtful inclusion. In this regard, Chan and Tsi (2024) emphasized the importance of developing AI literacy in order to effectively integrate GenAI into teaching and learning practices, and to address practical issues such as ethics and privacy.

1.1 AI Literacy and Language Learning

Significant efforts have been made over the past five years to define AI literacy (Zhao et al., 2024). Building on prior definitions, Long and Magerko (2020) described AI literacy as a set of competencies that enable individuals to critically assess AI, interact effectively with it, and apply it in various settings such as online, at home, and in the workplace. Expanding on this foundation, Zhao et al. (2024) proposed a definition of GenAI literacy comprising five core competencies: pragmatic use and critical evaluation of AI tools, awareness of safety and privacy risks, reflection on AI's impact on learning, understanding of socio-ethical implications, and appropriate, transparent use in context. In the context of English language learning, AI literacy entails understanding how to use AI tools effectively, recognizing their strengths and limitations, and critically integrating them to enhance language acquisition and communication skills (Alzubi, 2024).

Research on the impact of AI literacy on English as a Foreign Language (EFL) learning has recently gained attention. In the area of communication, Zhang et al. (2025) examined the relationship between AI literacy and psychological factors and showed that AI literacy positively influenced learners' self-efficacy and reduced foreign language anxiety. In the area of writing, Alzubi (2024) investigated factors affecting university students' GenAI literacy in EFL contexts and found that GPA and mastery in using GenAI tools as significant predictors. These findings underscore the importance of integrating AI literacy into EFL curricula.

Furthermore, studies indicate that individuals' perceptions of GenAI are shaped by their familiarity with and frequency of using such tools. For example, Chan and Hu (2023) found that students with greater knowledge and experience of GenAI were more likely to continue using it. Similarly, Tossell et al. (2024) reported that students' perceptions shifted from viewing GenAI as a tool for cheating to one of collaboration following usage. Bailey et al. (2021) also found that participation rates and self-rated L2 proficiency were positively correlated. Collectively, these findings suggest a need to explore how fostering AI literacy may influence the perceptions of learners who use GenAI less frequently.

1.2 Technology Acceptance Model

To analyze learners' perceptions of using new technologies, the Technology Acceptance Model (TAM) has been widely utilized. TAM is a theoretical framework proposed by Davis (1989) to predict users' acceptance of information technology. It comprises several constructs, including perceived usefulness (PU), perceived ease of use (PEU), attitude (AT), behavioral intention (BI), and actual use (AU). According to the model, AU is predicted by BI, which is, in turn, influenced by AT (Goh & Wen, 2020). Furthermore, AT is shaped by two key factors: PU, defined as the degree to which an individual believes that using a particular technology will enhance their performance, and PEU, defined as the degree to which they believe that using the technology will require minimal effort (Goh & Wen, 2020). Prior research has shown that PEU positively predicts PU, and that PU directly influences BI (Goh & Wen, 2020).

Researchers have also applied TAM in EFL learning studies involving various technologies, including GenAI. For example, studies introducing GenAI into EFL learning environments have shown that PEU positively predicts PU (Liu et al., 2023; Liu & Ma, 2023; Zou et al., 2024); PU positively predicts AT (Liu & Ma, 2023; Zou et al., 2024); PEU positively predicts AT (Liu & Ma, 2023); AT positively predicts BI (Liu & Ma, 2023; Zou et al., 2024); and BI positively predicts AU (Liu et al., 2023; Liu & Ma, 2023). While these studies have examined the relationships among TAM constructs in the context of GenAI use in EFL learning, few have

explored how EFL learners perceive GenAI from the perspective of these constructs. Investigating students' perceptions through the lens of TAM may offer valuable pedagogical insights into the strengths and limitations of implementing GenAI in EFL teaching and learning.

1.3 Research Purpose

Therefore, the purpose of this study is to gain insights that support the integration of GenAI into EFL teaching and learning by qualitatively examining learners' perceptions through the lens of TAM constructs. Specifically, this study examines learners who studied AI literacy in the EFL context. It also considers how the frequency of GenAI use influences learners' perceptions. This study is guided by the following research questions:

RQ1: How do EFL learners who studied AI literacy in the EFL context perceive the use of GenAI, including perceptions that align with or extend beyond the constructs of TAM?

RQ2: How does the frequency of GenAI use influence EFL learners' perceptions of GenAI, as interpreted through the TAM framework?

2. Methodology

2.1 Participants

This study was conducted at a private university located in eastern Japan. The participants were third- and fourth-year Japanese university students majoring in English who were enrolled in the course described in the *Procedure* section. Data were collected from the final reports submitted by the participants at the end of the course. Of the 43 reports submitted, two were excluded from the analysis because they did not sufficiently describe the use of GenAI. As a result, 41 reports were included in the analysis.

2.2 Instruments

In the pre-survey, participants were asked demographic questions, including their year in school, gender, and frequency of AI use. Frequency of AI use was measured using a self-reported five-point Likert scale ranging from 1 (Never) to 5 (Always), following the scale used by Chan and Hu (2023).

2.3 Procedure

The program consisted of four 100-minute sessions (see Table 1), conducted during the second half of a 14-week course during the fall semester of the 2024 academic year. The aim of the program was to develop students' AI literacy in the context of EFL learning by fostering their understanding of GenAI's characteristics and mechanisms—including its strengths, limitations, and ethical and safety considerations—and to promote the effective use of GenAI in their English studies.

In the second and third sessions, students were assigned follow-up tasks that required them to actively use GenAI for EFL-related activities. They were encouraged to make extensive use of GenAI tools in their English learning. Students applied GenAI for various tasks, such as sentence correction, grammar checking, and vocabulary test creation.

Before the program began, a pre-survey was administered. After the completion of the four sessions, students responded to the following two prompts as part of their final reports:

- What were your impressions of using GenAI in your English learning? Please describe any benefits and drawbacks you experienced, including specific examples.
- Do you intend to continue using GenAI in your English learning? Please explain your reasons in detail.

Students were informed of the study's purpose, and that anonymized data might be used for research purposes prior to completing the survey and report. They were also assured

that participation was voluntary and that non-participation would not affect their grades. Regarding the report, students who did not wish to participate were instructed to indicate their refusal in the report or to contact the instructor via email.

Table 1. *Overview of the Program*

Session	Topics and Activities
Week 1	Types of generative AI; how ChatGPT generates text; exploring ChatGPT (e.g., investigating functions, creating an account, and using it freely)
Week 2	Comparing GenAI tools; introduction of prompts for English learning using ChatGPT (e.g., comparing outputs from different GenAI tools; trying sample prompts provided by the instructor, including speaking practice, sentence correction, grammar checking, and vocabulary test creation)
Week 3	How to create effective prompts for English learning (e.g., sharing follow-up tasks, analyzing good/bad prompts, creating and testing prompts)
Week 4	Ethical and safety issues in GenAI use (e.g., sharing follow-up tasks, discussing ethical and safety concerns related to using GenAI)

2.4 Data Analysis

To investigate RQ1, content analysis was conducted using students' reports from the perspective of TAM. First, each report was segmented into individual meaning-preserving sentences, which were then classified according to the revised definitions of TAM constructs proposed by Liu and Ma (2023) to assess EFL learners' behavioral use of ChatGPT. This study further elaborated on their framework by incorporating additional details to better suit the research context, considering both positive and negative aspects (see Table 2). For example, the sentence "I plan to continue using GenAI for foreign language learning. The reason is that GenAI is effective in reducing study time, which is a great help in a busy student life" was divided into two parts: the first sentence and the second sentence. These were then categorized individually as BI for the former sentence and PU for the latter sentence, respectively. The first author initially categorized all segmented sentences, and the second author, who has expertise in English education and educational psychology, independently classified five reports (12.2% of the total) using the same segmented data. Cohen's kappa coefficients were calculated to assess inter-rater reliability, and the mean kappa value indicated an acceptable level of agreement (κ = .66). When discrepancies occurred between the two authors, they discussed the differences until reaching a consensus. The first author then revised the initial classifications where necessary, based on these discussions and with reference to the TAM definitions.

Table 2. *Definitions of the TAM constructs*

TAM Constructs	Definitions
Perceived Ease of Use (PEU)	The extent to which a learner perceives that using GenAI for English learning requires little effort (Positive) or substantial effort (Negative).
Perceived Usefulness (PU)	The extent to which a learner perceives GenAI as highly useful and as facilitating their English learning (Positive) or as not useful and not contributing to their English learning (Negative).
Attitude (AT)	The extent to which a learner is interested in using GenAI for English learning and evaluates its use positively (Positive) or evaluates the use of GenAI for English learning negatively (Negative).
Behavioral Intention (BI)	The extent to which a learner intends to use GenAI for English learning (Positive) or intends not to use it (Negative).
Actual Use (AU)	The extent to which a learner autonomously uses GenAI in their English learning (Positive) or does not use it (Negative).

Finally, the first author inductively grouped sentences with similar meanings within each TAM

construct and labeled them as subcategories.

For RQ2, Fisher's exact test was conducted to compare the number of students whose responses included sentences falling into each subcategory between low- and high-frequency GenAI use groups. Students who selected 1 to 3 on the pre-survey item measuring AI use frequency were categorized as the low-frequency group (n = 15), and those who selected 4 or 5 as the high-frequency group (n = 23). Of the 41 participants, three did not report their frequency of AI use and were therefore excluded from the analysis. The test was conducted using js-STAR XR+ (release 2.1.3, Japanese version).

3. Result

To explore RQ1, results of qualitative analysis of learners' perceptions based on TAM constructs are discussed. For RQ2, differences of subcategories between the low- and high-frequency groups are compared.

3.1 Learners' Perceptions of GenAI Use in EFL Learning

Table 3 presents the subcategories of PEU. Four subcategories were identified, encompassing both positive and negative postures, with a total of 32 sentences: 19 positive and 13 negative. The positive subcategories reflected students' experiences of ease in terms of accessibility and the conversational format of GenAI. In contrast, the negative subcategories highlighted difficulties with generating intended outputs and the complexity of prompt creation.

Table 4 shows the subcategories of PU, which had the highest number of categorized sentences. Seventeen subcategories were identified, encompassing both positive and negative postures, with a total of 283 sentences: 225 positive and 58 negative. Students perceived GenAI as functionally useful, particularly for receiving immediate feedback and enabling personalized learning. They also regarded GenAI as a valuable tool for supporting the development of English language skills, including reading, writing, speaking, and listening. In addition, students used it as a conversation partner and as a tutor for suggesting corrections and improving the grammar in their English compositions. However, they also experienced output errors and expressed concerns about incorrect or unnatural outputs in both spoken and written forms.

Table 5 outlines the subcategories of AT. Eleven subcategories were identified, encompassing critical, positive, and negative postures, with a total of 130 sentences: 108 critical, 21 positive, and one negative. Critical responses reflected thoughtful consideration regarding the use of GenAI, rather than clearly positive or negative views. These responses revealed students' awareness of ethical and safety concerns, accuracy issues, and the risks of over-reliance. Positive responses described GenAI as a learning partner and noted that it enhances motivation through immediate feedback. The sole negative response expressed a sense of guilt about using GenAI for English learning.

Table 6 presents the subcategories of BI. Three subcategories were identified, encompassing positive, critical, and negative postures, with a total of 72 sentences: 60 positive, five critical, and seven negative. Of the sentences categorized under the BI construct, 83.3% expressed a willingness to continue using GenAI, while 9.7% indicated reluctance.

Table 7 presents the subcategories of AU. Two subcategories were identified, encompassing positive, and negative postures, with a total of three sentences: two positive and one negative. Although few sentences were categorized here, they indicated both active use and non-use of GenAI in EFL learning.

Table 8 shows the subcategories beyond the TAM constructs. Two subcategories were identified, encompassing both critical and positive postures, with a total of 19 sentences: 14 critical and five positive. One major theme, reflecting a critical posture, was the selective use of GenAI, in which students described the importance of determining which parts of the learning process should involve GenAI and which should be conducted through other methods. Some students also expressed expectations for technological advancement to improve the functionality and accuracy of GenAI in the future.

Table 3. *Subcategory of Perceived Ease of Use (PEU)*

Subcategory	Posture	Frequency	Excerpt
Accessibility anytime anywhere	Pos.	13	“I can practice as much as I want, whenever I want, without worrying about time or the other person’s availability.”
Ease of use via conversational input	Pos.	6	“Because it (GenAI) is in a conversational format, I can respond to the answers I receive, and the replies come back as if continuing a conversation.”
Unintended output	Neg.	12	“When I asked for listening advice, GenAI gave general suggestions like watching subtitled movies, but not specific feedback based on my level or pronunciation issues.”
Complexity of prompt creation	Neg.	1	“To get responses close to what I want, I need to include a lot of information in the prompt.”

Note. Pos. = positive; Neg. = negative (gray highlight). Parentheses added by the author. Excerpts were translated from Japanese student reports by the author.

Table 4. *Subcategory of Perceived Usefulness (PU)*

Subcategory	Posture	Frequency	Excerpt
Immediacy and Efficiency	Pos.	36	“It’s helpful that AI provides instant feedback and corrections, allowing me to resolve questions on the spot.”
Personalized and Goal-Oriented Use	Pos.	30	“Since it (GenAI) suggests English expressions tailored to my level and learning goals, it (GenAI) allows for more practical learning.”
Vocabulary and Expression Enhancement	Pos.	28	“It (GenAI) also suggested alternative grammar and expressions, which provided a good opportunity to learn English phrases I wouldn’t have come up with on my own.”
Writing Correction and Grammar Improvement	Pos.	27	“When I wrote an English composition, it (GenAI) pointed out grammatical and vocabulary mistakes, which helped me revise my writing to sound more natural.”
Conversation Practice	Pos.	24	"In particular, for conversation practice, it (GenAI) allows for practical training by setting specific situations and engaging in dialogue-based learning."
Quiz and Material Generation	Pos.	23	"When I asked GenAI to create TOEIC questions and specified the part, format, score range, and number of questions to practice, I found that the output was nearly identical to the actual test."
General Responses	Pos.	20	“GenAI is a highly beneficial tool for foreign language learning and using it appropriately can enhance the quality of learning.”
Importance of Prompts	Pos.	12	"Customized prompts, such as table-formatted corrections, made learning more effective."
Idea Generation	Pos.	6	"ChatGPT is useful for generating ideas and sometimes offers suggestions from perspectives I hadn’t considered."
Well-Structured Output	Pos.	6	“It (GenAI) created a table that summarized the meanings, usage, example sentences, and differences of each word.”
Citation & Source Suggestion	Pos.	4	“Since many of the responses are search-based, they are easy to understand with specific examples and sources.”
Translation & Summary	Pos.	4	“GenAI also provided answers about slang, which machine translation could never handle.”
Listening Skill Improvement	Pos.	3	“By entering English text and using the text-to-speech function, I can learn correct pronunciation and intonation.”
Multilingual Support	Pos.	2	“I also found it impressive that GenAI supports multiple languages and can explain the nuances between them.”
Output errors and bias	Neg.	30	“When I wrote long texts, there were sometimes grammatically incorrect parts or expressions that sounded unnatural in meaning.”
Unnatural Output (Speaking)	Neg.	16	“Since the conversation partner is AI, I noticed a drawback that its responses sometimes became repetitive and emotionally flat during interactions.”

Unnatural Output (General)	Neg.	12	“For expressions involving humor or emotion, the suggestions made by AI were sometimes unnatural or did not fit the context.”

Note. Pos. = positive; Neg. = negative (gray highlight). Parentheses added by the author. Excerpts were translated from Japanese student reports by the author.

Table 5. *Subcategory of Attitude (AT)*

Subcategory	Posture	Frequency	Excerpt
GenAI as a supportive tool for humans	Critical	37	“I feel that by using AI as a supplementary tool, I can continue learning while compensating for its limitations.”
Over-reliance and Cognitive Risk	Critical	35	"Relying too much on AI may undermine the learner’s ability to think and express ideas independently."
Verifying information reliability	Critical	20	"Therefore, I felt it was important not to use AI-generated text as it is, but to carefully review it myself."
Understanding GenAI mechanisms	Critical	8	"Since GenAI produces output based on past information, we must understand how it works in order to effectively apply it to language learning."
Security and safety	Critical	8	"Once personal information is entered into GenAI, it may be learned by the system and could lead to a privacy breach if accessed by a third party."
GenAI as a learning partner	Pos.	7	"I hope to use AI not just as a supplementary tool, but as a partner that supports my learning, and to establish a well-balanced approach to language study."
Learning motivation	Pos.	5	"By receiving quick responses, I was able to resolve my questions immediately, which helped me stay motivated in my learning."
Free access	Pos.	4	"There are many advantages, such as being able to do a fair amount without spending any money."
No human consideration needed	Pos.	3	"Because it (GenAI) has no emotions, I can use it without worrying about what others think."
Being supported by GenAI	Pos.	2	"Looking back, I realized that I have often been supported by GenAI in my foreign language learning."
GenAI use guilt	Neg.	1	"The sense of guilt I feel about using it frequently has not changed since before taking the course."

Note. Pos. = positive; Neg. = negative (gray highlight). Parentheses added by the author. Excerpts were translated from Japanese student reports by the author.

Table 6. *Subcategory of Behavioral Intention (BI)*

Subcategory	Posture	Frequency	Excerpt
Willingness to Use	Pos.	60	"Therefore, I would like to continue actively using GenAI in my foreign language learning."
Cautious Use	Critical	5	"I would like to use it appropriately without overreliance, while fully recognizing its limitations."
Reluctance to Use	Neg.	7	"While I benefited from the course, I am personally not inclined to actively use GenAI in my future language learning."

Note. Pos. = positive; Neg. = negative (gray highlight). Excerpts were translated from Japanese student reports by the author.

Table 7. *Subcategory of Actual Use (AU)*

Subcategory	Posture	Frequency	Excerpt
Using	Pos.	2	"This course served as a turning point for me to begin actively using generative AI in my English learning."

Not Using	Neg.	1	"Since ChatGPT is rarely permitted in university classes, I did not use it before the course."

Note. Pos. = positive; Neg. = negative (gray highlight). Excerpts were translated from Japanese student reports by the author.

Table 8. *Subcategory Beyond the Scope of TAM Constructs*

Subcategory	Posture	Frequency	Excerpt
Selective Use of GenAI (Delegation vs. Other Methods)	Critical	14	"I felt that if learners are not consciously aware of which parts of the learning process they are delegating to GenAI, it ceases to be meaningful learning and simply becomes a shortcut."
Expectations for Technological Advancement	Pos.	5	"I believe that as GenAI continues to improve in accuracy and functionality, current issues such as the provision of inaccurate information and incorrect answers are likely to be resolved."

Note. Pos. = positive. Excerpts were translated from Japanese student reports by the author.

3.2 Frequency of GenAI Use and EFL Learners' Perceptions of GenAI

To examine differences in subcategories between low- and high-frequency GenAI use groups, Fisher's exact test was conducted. Significant differences were found in two BI subcategories: *Willingness to Use* and *Reluctance to Use*. No significant differences were observed in the remaining subcategories, including those under PEU, PU, AT, AU, and *Cautious Use* in BI. In the *Willingness to Use* subcategory, 11 of 15 students in the low-frequency group and all 23 students in the high-frequency group provided relevant responses (p = .0185, Φ = .337). In contrast, *Reluctance to Use* was expressed by 4 of 15 students in the low-frequency group, compared to none in the high-frequency group (p = .0185, Φ = .337). These results suggest that prior GenAI usage frequency significantly influenced students' behavioral intentions.

4. Discussion

Based on RQ1, this study examined how learners who studied AI literacy in the EFL context perceive GenAI for English learning, using TAM constructs. Three key findings emerged from the analysis.

First, students experienced a variety of functional benefits from using GenAI in their English learning, including not only personalized and immediate learning support highlighted in prior studies, but also support for all four English skills: speaking, writing, reading, and listening (see Table 4). This may help explain why most students expressed a positive intention to use GenAI in their English learning within the BI construct, despite perceiving both positive and negative aspects in the PEU and PU constructs. This suggests that although students encountered challenges in using GenAI, they were still willing to continue its use—possibly because they perceived the benefits as outweighing the drawbacks. This finding differs slightly from previous studies, which reported that AI tools are predominantly used for translation and grammar correction in academic contexts (Hossain et al., 2025). Hossain et al. (2025) pointed out that this was because students tend to use AI as a translator rather than as a tool to support their writing. These findings suggest the need for pedagogical intervention to guide learners in using GenAI as a tool for learning support rather than as a means of academic dishonesty. This study contributes to the existing literature by demonstrating that, when appropriately guided, learners can integrate GenAI across diverse aspects of English learning and recognize it as an effective tool for supporting language development.

Second, 108 out of 130 sentences, which were classified into five subcategories within the AT construct, were associated with critical perspectives. These subcategories include human-AI interaction, concerns about over-reliance, ethical and moral considerations, security and safety issues, and the importance of understanding GenAI's mechanisms. Hossain et al.

(2025) reported that while Turkish university students were familiar with GenAI, they lacked technical proficiency and a clear understanding of its underlying mechanisms. Their study suggested that such understanding cannot be acquired through use alone. The findings of the current study indicate that developing an awareness of how GenAI works, including both its benefits and limitations, may help nurture a more critical mindset among students. Therefore, explicitly addressing these aspects in instruction may help prevent over-reliance and foster ethical and reflective use of GenAI in English language learning.

Finally, this study identified two subcategories that did not fall within the traditional TAM constructs: *Selective Use of GenAI* and *Expectations for Technical Advancement*. While Section 1.2 outlined the internal constructs of TAM, the model also allows for external variables that influence these internal factors. Previous studies have proposed various external variables, such as emotions, motivational factors, social influence, and facilitating conditions (i.e., the perceived availability of necessary resources) (Liu & Wu, 2025). The subcategory *Expectations for Technical Advancement* reflects learners' anticipation of improved GenAI functionality as a potential external factor influencing future acceptance. In contrast, *Selective Use of GenAI* may appear to fall under BI, as it involves learners' decisions about when and how to use GenAI. However, this subcategory represents a more nuanced and critical intention—not merely a willingness or reluctance to use GenAI, but a deliberate effort to determine which aspects of learning should involve GenAI and which should not. This reflective stance suggests the emergence of learner agency in balancing technological support with active engagement, thereby promoting more effective language learning. It also aligns with one of the key components of GenAI literacy, namely *pragmatic understanding*, as proposed by Zhao et al. (2024). According to Zhao et al. (2024), *pragmatic understanding* refers to an individual's ability to effectively use GenAI by selecting the appropriate tool for a given task and applying it efficiently. These findings suggest that, in the context of GenAI-supported language learning, the conceptual scope of BI may need to be extended to incorporate this critical mindset.

For RQ2, the results showed that both low- and high-frequency groups perceived GenAI as equally useful and easy to use, and they demonstrated similar attitudes toward its use in English learning. However, the low-frequency group exhibited significantly less willingness to use GenAI compared to the high-frequency group. This suggests that, even though both groups perceived the benefits and drawbacks of GenAI similarly throughout the program, students with less prior experience were still more reluctant to adopt GenAI use after the program. Some students attributed their hesitation to feelings of guilt or concerns that GenAI might be "too effective," potentially undermining their own learning efforts. These findings indicate that regardless of perceived usefulness, some learners may selectively avoid using GenAI due to psychological or ethical reservations.

5. Conclusion

This study examined how EFL learners who studied AI literacy in the EFL context perceive the use of GenAI through the lens of TAM, and explored how the frequency of GenAI use influenced their perceptions. The findings demonstrated that participation in the AI literacy program enabled students to experience a variety of functional benefits and challenges associated with GenAI, while also fostering a critical mindset toward its use in English learning. These results highlight the need for pedagogical interventions that support learners in using GenAI as an effective learning tool and in developing an understanding of its underlying mechanisms, thereby promoting ethical and reflective engagement. Furthermore, the study suggests that in the context of GenAI-supported language learning, the conceptual scope of the TAM's BI construct may need to be extended to incorporate learners' critical and selective decision-making. Future studies may explore the relationships among the TAM constructs, including an extended definition of BI that incorporates critical and reflective dimensions of technology use. The primary limitation of this study is its small sample size, drawn from a single private university in Japan. Therefore, the findings may have limited generalizability to broader EFL populations and educational contexts.

References

Alm, A., & Ohashi, L. (2024). A worldwide study on language educators' initial response to ChatGPT. *Technology in Language Teaching & Learning*, *6*(1), Article 1. https://doi.org/10.29140/tltl.v6n1.1141

Alzubi, A. A. F. (2024). Generative artificial intelligence in the EFL writing context: Students' literacy in perspective. *Qubahan Academic Journal*, *4*(2), Article 2. https://doi.org/10.48161/qaj.v4n2a506

Bailey, D., Southam, A., & Costley, J. (2021). Digital storytelling with chatbots: Mapping L2 participation and perception patterns. *Interactive Technology and Smart Education, 18*(1), 85–103, https://doi.org/10.1108/ITSE-08-2020-0170

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, *20*(1), 43. https://doi.org/10.1186/s41239-023-00411-8

Chan, C. K. Y., & Tsi, L. H. Y. (2024). Will generative AI replace teachers in higher education? A study of teacher and student perceptions. *Studies in Educational Evaluation*, *83*, 101395. https://doi.org/10.1016/j.stueduc.2024.101395

Chukhno, O. (2024). Teachers and learners' perspectives on the use of generative AI in foreign language education. *Modern Information Technologies and Innovation Methodologies of Education in Professional Training Methodology Theory Experience Problems*, *73*, 53–60. https://doi.org/10.31652/2412-1142-2024-73-53-60

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340. https://doi.org/10.2307/249008

Goh, E., & Wen, J. (2020). Applying the technology acceptance model to understand hospitality management students' intentions to use electronic discussion boards as a learning tool. *Journal of Teaching in Travel & Tourism*. https://doi.org/10.1080/15313220.2020.1768621

Hossain, Z., Çelik, Ö., & Hiniz, G. (2025). Exploring EFL students' AI literacy in academic writing: Insights into familiarity, knowledge and ethical perceptions. *Kuramsal Eğitimbilim*, *19*, 157–181. https://doi.org/10.30831/akukeg.1538011

Law, L. (2024). Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. *Computers and Education Open*, *6*, 100174. https://doi.org/10.1016/j.caeo.2024.100174

Liu, G. L., Ron, D., & Ma, C. (2023). Exploring AI-mediated informal digital learning of English (AI-IDLE): A mixed-method investigation of Chinese EFL learners' AI adoption and experiences. *Computer Assisted Language Learning*, 1–29. https://doi.org/10.1080/09588221.2024.2310288

Liu, G., & Ma, C. (2023). Measuring EFL learners' use of ChatGPT in informal digital learning of English based on the technology acceptance model. *Innovation in Language Learning and Teaching*. https://doi.org/10.1080/17501229.2023.2240316

Liu, W., & Wu, H. (2025). Towards a unified model of generative intelligence acceptance (UMGA) in second language learning: A review and integration. *Am J Psychol*, *2*(1), 1–8.

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3313831.3376727

Tossell, C. C., Tenhundfeld, N. L., Momen, A., Cooley, K., & de Visser, E. J. (2024). Student perceptions of ChatGPT use in a college essay assignment: Implications for learning, grading, and trust in artificial intelligence. *IEEE Transactions on Learning Technologies*, *17*, 1069–1081. IEEE transactions on learning technologies. https://doi.org/10.1109/TLT.2024.3355015

Trust, T., Whalen, J., & Mouza, C. (2023). Editorial: ChatGPT: Challenges, opportunities, and implications for teacher education. *Contemporary Issues in Technology and Teacher Education*, *23*(1), 1–23.

Zhang, Q., Nie, H., Fan, J., & Liu, H. (2025). Exploring the dynamics of artificial intelligence literacy on English as a foreign language learners' willingness to communicate: The critical mediating roles of artificial intelligence learning self-efficacy and classroom anxiety. *Behavioral Sciences*, *15*(4), 523. https://doi.org/10.3390/bs15040523

Zhao, X., Cox, A., & Cai, L. (2024). ChatGPT and the digitisation of writing. *Humanities and Social Sciences Communications*, *11*(1), 1–9. https://doi.org/10.1057/s41599-024-02904-x

Zou, B., Wang, C., Yan, Y., Du, X., & Ji, Y. (2024). Exploring English as a foreign language learners' adoption and utilisation of ChatGPT for speaking practice through an extended technology acceptance model. *International Journal of Applied Linguistics*. https://doi.org/10.1111/ijal.12658

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Chatbot Personalisation for EFL Learning: Integrating BKT and Task Context

Steve WOOLLASTON*, Brendan FLANAGAN, & Hiroaki OGATA
Graduate School of Informatics, Kyoto University, Japan
*s.m.woollaston@gmail.com

Abstract: English proficiency is critical in today's globalised world, yet many EFL learners face challenges in traditional classrooms due to large class sizes, rigid pedagogy, and varying proficiency levels. While AI chatbots offer accessible, low-pressure language practice, their effectiveness is limited by a lack of personalisation. This paper proposes a novel framework for dynamically personalising EFL chatbot interactions by integrating Bayesian Knowledge Tracing with semantic task analysis. The system aggregates learner data across multiple task-specific chatbots (e.g., translation, writing) to model vocabulary and grammar mastery. By analysing task context, the framework generates recommendations tailored to each learner, ensuring both proficiency-appropriate and activity-relevant support. This work advances personalised language learning by bridging the gap between isolated chatbots and a unified, adaptive learning ecosystem.

Keywords: Chatbot, conversational agents, EFL, chatlogs, second language learning, LLM, TELL, personalisation, bayesian knowledge tracing, semantic similarity

1. Introduction

English proficiency is increasingly important for living and working in today's interconnected world (Morita, 2017). This is especially true in Japan, which plays an active role in international economic and global political affairs. However, Japanese learners of English as a Foreign Language (EFL) face significant challenges within traditional classroom settings, including large class sizes, a reliance on grammar-translation pedagogical approaches, diverse proficiency levels among learners, and varying levels of motivation (Herawati, 2023). These and other factors can make it difficult to provide the individualised support and practice necessary for effective language acquisition (Zhang & Huang, 2024).

Educational technology, particularly the use of chatbots, has emerged as a promising tool to supplement traditional EFL instruction. Chatbots provide accessible and real-time interactive language practice opportunities outside the classroom, allowing students to learn independently and at their own pace in a friendly and low pressure environment. Chatbots can provide unlimited conversational practice, writing support, feedback, and other language learning activities (Brinegar, 2023).

1.1 *Personalisation*

Despite their advantages, the effectiveness of chatbots can be limited by their general one-size-fits-all approach (Zhang & Huang, 2024). Learners vary significantly in their current knowledge and learning pace. Personalised learning is a strategy to tailor learning content and experiences to the unique profile of each learner (Zerkowska, 2024). The US Department of Education similarly defines it as "...instruction in which the pace of learning and the instructional approach are optimised for the needs of each learner. Learning objectives, instructional approaches, and instructional content (and its sequencing) may all vary based on learner needs." (King & South, 2017, p. 9). AI chatbots have struggled to provide effective personalisation as they have little or no knowledge of the learner's current proficiency, and thus fail to adapt effectively to the prior knowledge and specific queries of individual learners. For EFL chatbots, personalisation means adapting interactions to the learner's Zone of Proximal Development (ZPD), providing appropriately challenging input (vocabulary and grammar), and ensuring relevance to the specific demands of the current

task (Chaiklin, 2003). Such personalisation is widely endorsed for boosting learner engagement and attainment (Prain et al., 2013). Chatbot systems that can dynamically model learner progress and provide this targeted, individualised support is crucial for improving EFL practice and learning outcomes.

1.2 *Proposed Framework*

In this paper, we propose a novel personalisation framework designed to inform the responses of a suite of EFL learning chatbots currently deployed in high schools and a university in Japan. Each chatbot is task-specific, specialising in key language skills (active reading, translation, diary writing, and academic writing). Log data generated from these chatbots will be analysed in a unified backend system. This proposed system will employ Bayesian Knowledge Tracing (BKT) to model learners' mastery of vocabulary and grammar based on their interactions across all chatbots. By analysing task contexts (e.g., reading text excerpts, diary prompts and entries, translation prompts) via word embeddings, the system will generate real-time recommendations tailored to each learner's ZPD. These recommendations will target gaps in knowledge while ensuring relevance to the learner's current activity, such as suggesting vocabulary synonyms during translation practice or highlighting grammatical errors in diary entries.

1.3 *Contribution*

This research expands on Takii et al.'s (2021) vocabulary quiz recommendation system by introducing three key advancements to enhance personalised chatbot language learning. First, the framework enables **cross-chatbot learner modelling**, which builds a comprehensive, dynamic learner profile by aggregating interaction data from a variety of language learning activities like reading, writing, and translation, rather than being limited to a single task. Second, it is **task-contextualised**, combining proficiency estimates from Bayesian Knowledge Tracing (BKT) with semantic analysis of task content to ensure that recommendations are not only appropriate for the learner's skill level but also relevant to the specific activity. Finally, the system is designed with a **Japan-centric EFL alignment**, integrating nationally recognised CEFR-J grammar standards (Ishii & Tono, 2016) that are used in common school textbooks. This integration ensures both scalability and curricular relevance within the Japanese educational context.

2. Related Work

2.1 *Personalisation in Educational Technology*

Personalisation is a key goal in computer-assisted education, enabling educators to tailor instruction to meet the needs of diverse learners (Lin & Chang, 2023). This approach involves providing tailored recommendations and content based on areas where a learner struggles, which can be identified through their interactions with a system (Brinegar, 2023). Tailoring responses to learners' previous interactions can personalise learning experiences and potentially impact achievement (Huang et al., 2022). Examples include providing recommended reading materials based on learner characteristics like gender and interests, or tailoring explanations to a learner's proficiency level (Zhang & Huang, 2024). However, there are challenges in implementing effective personalisation. Despite positive effects reported in research, it can be unclear whether these effects result solely from the chatbot's personalised guidance or from the learners' strategic decisions (Lin & Chang, 2023).

2.2 *AI Chatbots in Language Learning*

Chatbots have gained prominence in language learning due to their ability to converse using natural language. They offer learners opportunities for dialogue-based practice and real-time

feedback (Wang et al., 2024). While rule-based chatbots often struggled to understand nuanced input and produce natural language that was not formulaic or lacked contextual understanding, modern chatbots utilising large language models (LLM) are much more adept at producing natural varied responses and have a broad general knowledge (Woollaston, Flanagan, & Ogata, 2024). They are also robust to input errors, which is important in EFL contexts where mistakes and errors are common (Mzury, 2023). Chatbots provide numerous benefits in language learning:

- **Accessibility:** Learners can practise their language skills with chatbots anytime, anywhere, which is not easily possible with a human partner proficient in the target language. They offer quick access to support, addressing questions in real-time when human teachers are unavailable (Huang et al., 2022).
- **Lessened anxiety:** Chatbots offer a non-judgemental learning environment, are patient with errors and questions, and can provide a safe environment to experiment with new language without fear of making mistakes (Zhai, 2023).
- **Self-pacing and repetition:** Chatbots allow for self-paced conversation and learning (Ait Baha et al., 2023). They are also endlessly patient, providing unlimited practice - an important part of language learning (Huang et al., 2022).
- **Task-specific:** Chatbots can support various language learning activities, including conversation practice, answering language learning questions, conducting assessment and providing feedback, scaffolding writing skills, reading comprehension support, and translation practice.

2.3 *Learner Modelling*

Learner Modelling (LM) is the cornerstone of adaptive educational systems (ALS) and a critical component of personalised learning (Abyaa et al., 2019). Effective personalisation requires dynamic adaptation to a learner's current knowledge and task-specific needs. This adaptation hinges on robust LM, which serves as the system's "beliefs" about a learner's knowledge, misconceptions, and progress (Kay et al., 2022). By continuously collecting and interpreting data—such as mastery and knowledge gaps—LM provides the foundation for tailoring instruction, feedback, and content sequencing to individual learners.

LM has focused on modelling knowledge and cognitive skills, leveraging educational theories like the Learning Curve and Forgetting Curve to track proficiency over time (Chen et al., 2017). Modern systems increasingly incorporate Open Learner Models, which make this data and insights to learners and teachers, fostering metacognitive awareness and self-regulated learning (Abyaa et al., 2019). More recently, the Open Knowledge and Learner Model (OKLM) has been proposed as a universal learner model framework that integrates Learning Analytics (LA) data from everyday learning activities with knowledge maps extracted from learning materials (Takii et al., 2024a). OKLM aims to enhance versatility and accuracy by linking learning activity logs to a knowledge map, allowing management and tracking of knowledge acquisition across diverse materials and contexts (Takii et al., 2024b).

2.4 *Relevance to Learning Context*

Cognitive psychology research shows that learning improves significantly when the content is directly relevant to the task (Bransford et al., 2000). This relevance boosts learner motivation and engagement by highlighting the practical value of the material (Ryan & Deci, 2000). It also fosters more meaningful and memorable learning as new information is integrated with existing knowledge within the specific context, ultimately improving retention (Jones, 1999). Furthermore, task-relevant learning content promotes more efficient learning by focusing on essential information. In EFL contexts, this principle underscores the importance of providing vocabulary and grammar recommendations that are directly applicable to the specific activity and context, thereby optimising the learning experience. Lexical and grammatical skills are best developed when relevant to the context, allowing for new information to be linked to prior knowledge (Jones, 1999).

By utilising semantic embeddings and distance metrics, language learning can be improved through the incorporation of meaning and context (Zerkowska, 2024). Semantic similarity measures, using models like Word2Vec, quantify the relatedness of text beyond keywords (Edapurath, 2022). These embeddings capture complex language nuances, enabling metrics like cosine similarity to determine contextual relevance (Hariprasath et al., 2024). This allows chatbots and personalised learning systems to provide accurate, meaningful responses and materials, going beyond simple keyword matching to infer context to a nuanced understanding.

3. Proposed Framework

3.1 *Chatbot Suite*

Several LLM-powered EFL chatbots have been developed and are currently deployed. **TAMMY** (Translation Assistant for MMasterY) is a chatbot designed to support English learning through translation tasks - a common task in English proficiency examinations in Asia (Ross, 2008). TAMMY allows learners to translate sentences between Japanese and English, receive feedback, ask questions, and engage in dialogue to improve understanding. In a pilot study, Woollaston et al., (2024) analysed the chat logs and a usability questionnaire. The findings indicated that TAMMY demonstrated good response validity and moderate success in guiding learners to accurate translations. Learners perceived the chatbot as friendly and easy to use, though their views on its usefulness and their intention to continue using it were neutral.

ARCHIE (Active Reading CHatbot for Interactive English) is a chatbot that scaffolds active reading strategies (e.g., summarisation, questioning, making connections) through 12 interactive activities (e.g., cloze passages, comprehension questions, role-play). Activities are tied to predefined texts (uploaded by the teacher), encouraging learners to engage deeply with content. Logs capture reading actions (e.g., page navigation, highlighting, memo-writing) and chatbot interactions. In a study to compare how high-proficiency and low-proficiency EFL learners interacted with ARCHIE, Woollaston et al., (2025) coded learner messages and employed Epistemic Network Analysis (ENA) to analyse learner behaviour with the assigned reading and the chatbot, They found that while all learners perceived the chatbot positively, high-proficiency learners interacted more frequently with the chatbot, exhibited more active reading strategies such as backtracking, and demonstrated less help-seeking behavior.

Penny and **Revision Rex** are writing support chatbots developed recently, and currently deployed at a Japanese high school and university, respectively. Penny assists with English diary writing. Each day, learners receive a prompt such as *"What is your favorite memory?"* or *"If you had one superpower, what would it be?"* After responding, they receive feedback and suggestions to improve their writing. Revision Rex supports academic writing in English, including essays and reports. It provides feedback and scaffolding in four key areas: vocabulary and word choice, surface features (spelling, punctuation etc.), grammar and sentence structure, and text organisation and cohesion.

Data about learner writing and interactions with each chatbot is stored in a database. With consent, these are stored securely and may be purged at any time upon request.

3.2 *System Architecture*

Figure 1 provides a summary of the proposed system architecture. Learner interactions and generated English text across different task-specific chatbots feed into a unified database. The system will employ Bayesian Knowledge Tracing (BKT) to model individual learner's vocabulary and grammar knowledge, drawing from log data and leveraging vocabulary and grammar preprocessing. The Recommendation Engine will then use this learner model, along with the task context, to tailor the chatbot's responses, aiming to provide personalised and relevant feedback and suggested next-learning steps aligned with the learner's ZPD.

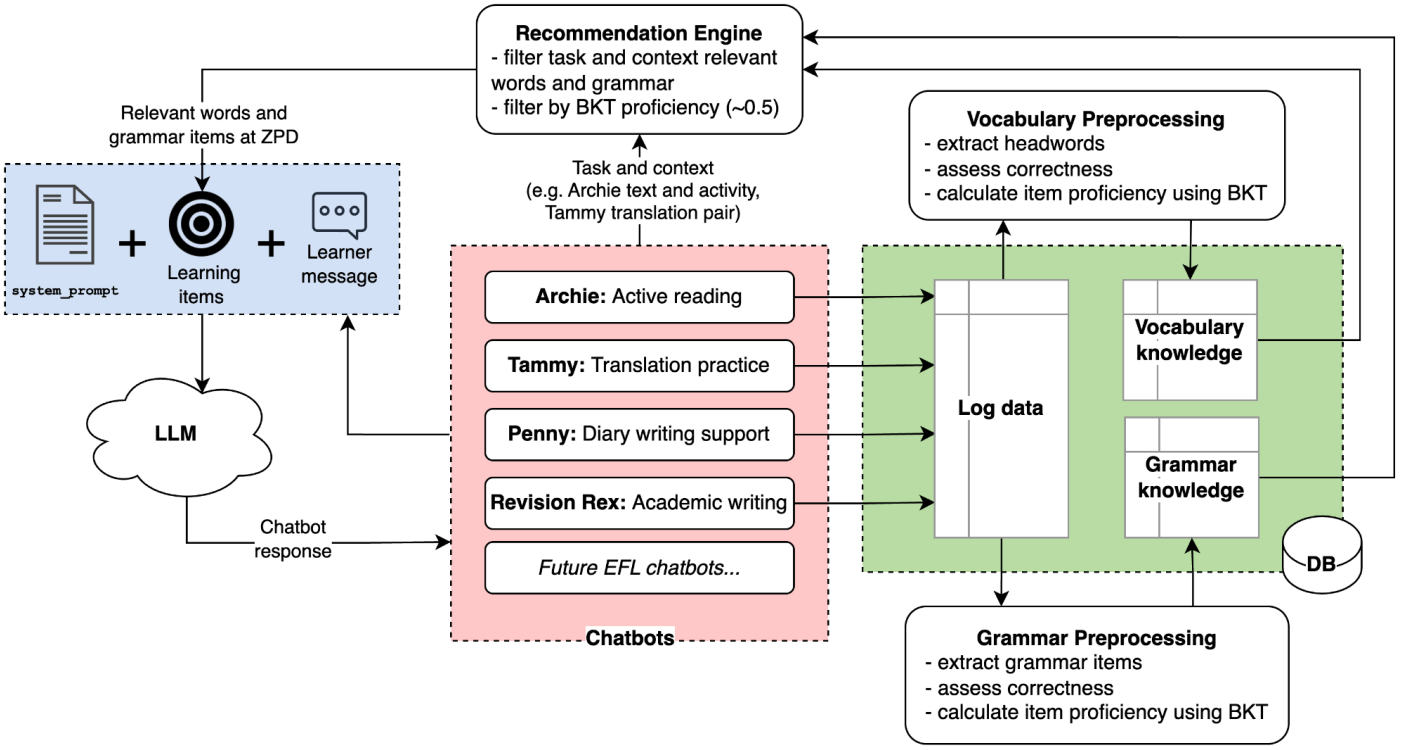


Figure 1. System architecture diagram

3.3 *Learner Modelling*

The system will dynamically model each learners' vocabulary and grammar knowledge by analysing learner-generated text from all chatbots. This includes messages sent to the chatbots and learner writing (e.g. diary entries in the diary chatbot Penny).

3.3.1 *Vocabulary Extraction and Evaluation*

For all learner submitted text, the system will execute a multi-step pipeline:

1. **Text Processing:** Inputs will be split into sentences, and each word will be checked for spelling errors. Misspelled words will be replaced with the most likely candidate (e.g., "acomodate" → "accommodate").
2. **Lexical Normalisation:** Each word will be lemmatised to its base form (headword) and tagged with its part-of-speech (POS). For example, "running" (verb) will be normalised to "run-VB," distinguishing it from noun or adjective uses.
3. **Usage Logging:** Each time a learner uses a normalised headword–POS pair, we record a log entry containing metadata such as the timestamp, the source chatbot, the activity context, and other relevant metadata (e.g., "ARCHIE; Summarisation; 'The Dangers of Social Media'"). We also save the learner's original sentence for context. As learners interact with the chatbots over time, each unique headword–POS pair accumulates one or more usage logs.

Whether a word is used correctly will be determined by evaluating two aspects:

- **Spelling:** Misspelled words will be automatically flagged as incorrect.
- **Contextual accuracy**: Correctly spelled words will be assessed for contextual accuracy using an LLM for classification. The model will evaluate whether the word aligns with the sentence's semantic and syntactic requirements (e.g., *Please <u>except</u> my invitation* vs. *Please <u>accept</u> my invitation*"). The model's classification (correct or incorrect usage) will be stored for later use by BKT.

3.3.2 *Grammar Extraction and Evaluation*

This section details the process of extracting and evaluating grammar usage in learner-generated text. This approach leverages the well-defined CEFR-J framework to identify specific grammatical structures and assesses the accuracy of their use within the context of the specific learning activity.

The Common European Framework of Reference for Languages (CEFR) provides a robust and widely accepted standard for describing language proficiency (Figueras, 2012). It provides a framework of language ability, from basic (A1) to fully proficient (C2), based on what language learners can do. In the Japanese context, the CEFR has been adapted into CEFR-J, which aims to apply the framework to the specific context of English language education in Japan, taking into account the unique challenges and needs of Japanese learners and aligning with Japanese educational policies and curriculum (Masashi, 2012).

A key component of CEFR-J is the identification of specific grammar items relevant to each proficiency level. These grammar items, derived from analyses of Japanese EFL textbooks and cross-referenced with other CEFR-aligned resources, are defined using regular expressions (regex). These regexes specify patterns based on word forms, lemmas, and parts of speech, enabling the automated extraction of grammar instances from English text. Table 1 provides a representative selection of these grammar items, illustrating their definition and corresponding example sentences. A comprehensive list of all 263 CEFR-J grammar items and their associated extraction regular expressions is publicly available at cefr-j.org (Ishii & Tono, 2016).

Table 1. *CEFR-J Grammar Item Examples*

CEFR-J Level	Grammar Item	Example sentence
A1.1	INDEFINITE ARTICLES (a/an + noun)	I have a book.
A1.1-A1.2	I am (... including question, negative)	I am happy. Am I next? I am not sad.
A1.1-B1.2	WH- QUESTION: How ADJ/ADV ...? (How + adjective/adverb ...?)	How quickly can you run?
A1.2-A1.3	TENSE/ASPECT: PRESENT PROGRESSIVE (present continuous)	She is reading.
A1.2-A2.2	VERB to DO (verb + to-infinitive)	I want to sleep.

Our grammar extraction and evaluation process involves the following steps:

1. **Sentence segmentation:** Learner-generated text is divided into sentences.
2. **LLM-enhanced error detection:** Each sentence is processed. Where grammar is problematic, an LLM will generate a grammatically corrected version with minimal changes, while preserving the original meaning. This step aids in highlighting potential errors for subsequent evaluation. Both the original and rewritten sentences are retained.
3. **Grammar item extraction:** For each sentence (or *minimally corrected* sentence), a list of all grammar items present in the sentence is generated using regex, as defined by the CEFR-J framework.
4. **Learner grammar tracking:** The system tracks learner grammar usage by logging each identified grammar item in their writing. For a learner's first use of a grammar point, it's recorded in their personal grammar record with its CEFR level. Subsequent uses trigger a usage log entry, noting the time, location (chatbot, activity, context), and an LLM-assessed correctness (correct/incorrect).

This method systematically identifies and evaluates grammar usage based on CEFR-J using LLM assistance for error highlighting and correctness assessment.

3.3.2.1. *English Vocabulary Profile*

While the CEFR-J framework was initially considered for vocabulary, its coverage (7,988 words, excluding C1–C2 levels) proves insufficient to accommodate advanced learners or support multilingual learners with diverse L1 backgrounds. To ensure inclusivity across proficiency levels and linguistic contexts, we will utilise the English Vocabulary Profile (EVP) as our predefined vocabulary list (Capel, 2015). The EVP offers two main advantages: complete coverage across all CEFR proficiency bands (A1-C2) and validation through extensive corpus analysis. This ensures our system can provide appropriate vocabulary recommendations for learners at any stage of their vocabulary learning.

3.4 *Modelling Vocabulary and Grammar Mastery*

To enable personalised and adaptive interactions across our suite of EFL chatbots, we will implement BKT to model learners' mastery of vocabulary and grammar in—or as close as possible to—real time. BKT's probabilistic framework will allow us to estimate the likelihood that a learner has mastered specific knowledge components (KCs)—such as a CEFR-J grammar rule or a vocabulary word—based on their observed performance during chatbot interactions. Unlike static pre-test assessments, BKT will dynamically update these estimates as learners engage with different tasks, ensuring that recommendations remain aligned with their evolving proficiency.

For vocabulary tracking, we will treat each lemmatised word–POS pair (e.g., "run-VB") as an individual KC. Since language knowledge is sensitive to forgetting over time, we will extend the classic BKT model to incorporate memory decay (BKT+F). This will allow us to adjust mastery estimates based on how frequently and recently a word has been encountered across different chatbot activities. Words that appear in multiple contexts will contribute to a more robust, generalised estimate of a learner's current knowledge.

The output of the BKT engine will be a continuously updated knowledge profile for each learner, which will drive personalisation across all chatbots. For vocabulary, this profile will prioritise words near the learner's ZPD—those with a $P(mastery)$ near 0.5—while ensuring relevance to their current activity (e.g., suggesting "leap" during a translation task if "jump" appears in their sentence). For grammar, the system will highlight errors on high-priority KCs, such as frequently missed A2-level rules during diary writing.

By modelling learner knowledge using BKT, our framework will advance beyond static rule-based personalisation, offering three key advantages:

1. Continuous updates to learner models as new data is generated across chatbots / tasks;
2. Effective management of errors, slips, and lucky guesses in learner language, and,
3. Scalability through shared KC definitions that unify data from diverse chatbots.

This approach will form the foundation for our system's ability to deliver tailored, context-aware support—a critical step toward addressing the personalisation gaps in current language learning chatbots.

3.5 *Aligning Task Relevance with Learner Model*

Our framework dynamically aligns learners' proficiency estimates (via BKT) with semantically relevant vocabulary and grammar suggestions through a three-stage pipeline:

1. **Proficiency Modelling:** BKT will track mastery probabilities for each knowledge component (KC)—lemmatised vocabulary (e.g., "run-VB") and CEFR-J grammar items (e.g., "past simple tense")
2. **Task Context Analysis:** When a learner engages in an activity (e.g., translating a sentence, drafting a diary entry), the system will extract keywords and grammatical structures from their input. Using word embeddings, the semantic similarity between these elements and the EVP will be computed. Grammar relevance will be ascertained by LLM. For example, in a translation task about sports, words like *compete* and *victory*

might be identified as contextually relevant vocabulary. For grammar, depending on the context (task and content), the system will suggest grammar items that are relevant (e.g. past tense usage during diary writing).

3. **ZPD Filtering:** The system will prioritise KCs with $P(mastery)$ near 0.5 and semantic similarity >0.6 to task keywords, ensuring suggestions like “stroll” (B1) for “walk” (A2) during a writing task. When processing grammar, the mechanism will prioritise corrective feedback based on the relevance to the current activity, such as emphasising tense consistency in narrative writing or article usage in descriptive tasks. This methodology ensures recommendations are simultaneously tailored to the learner's developmental stage and the specific linguistic demands of their current activity.

3.6 *Integration into Chatbot Dialogue*

When the Recommendation Engine identifies relevant vocabulary or grammar items based on the learner's model and the task context, this information will be used to dynamically shape the chatbot's responses via prompting. For vocabulary, the chatbot might subtly introduce target words in revisions, as synonyms, or within explanations, encouraging exposure, comprehension, and usage in context. For grammar, the chatbot will provide corrective feedback on errors, offer rephrased sentences demonstrating correct usage, or pose targeted questions that guide the learner to apply specific grammatical structures. This integration within the natural flow of the chatbot interaction will ensure that recommendations are timely, contextualised, and directly relevant to the learner's immediate learning activity.

4. Discussion

4.1 *Conclusion*

The proposed framework addresses three key EFL chatbot challenges: fragmented data, contextual misalignment, and relevance and scalability to the Japanese EFL context. Traditional isolated chatbots fail to track holistic progress, so our system aggregates data across activities (translation, writing, active reading) to reflect more comprehensive language skill integration. We employ BKT for its dynamic, probabilistic mastery estimation, which incrementally adapts to evolving proficiency. Unlike generic feedback systems, our semantic analysis ensures task-relevant suggestions, leveraging evidence that contextual learning boosts retention. The CEFR-J and EVP standards were chosen to balance Japanese curriculum specificity with comprehensive vocabulary coverage (A1–C2). By prioritising integration with existing tools (e.g., TAMMY, ARCHIE), the design ensures practical adoption in the Japanese context.

4.2 *Limitations*

The proposed framework faces several challenges, the most significant of which is **privacy and data security**. Aggregated learner data requires stringent safeguards to protect their privacy, such as data anonymisation and access control. Another issue is the potential for **bias and hallucinations** in LLM-generated feedback, as it may inherit biases from training data or produce incorrect corrections, risking learner trust (Zhai, 2023). For example, a chatbot might misclassify a creative but valid use of grammar as an error. The chatbots' **narrow skill focus** presents another challenge, as they currently concentrate on productive literacy skills, primarily writing, while excluding critical communicative skills like speaking and listening. Future chatbots should aim to expand the range of skills they develop. Lastly, the **LLM dependency** for error detection introduces stochastic variability; their judgments may lack the consistency, nuance, and experience of human tutors, especially when dealing with ambiguous or culturally specific language.

4.3 *Future Work*

To further enhance the framework's efficacy and applicability, several development and research directions warrant exploration. First, the development of interactive dashboards would empower learners and teachers to monitor progress, adjust recommendation thresholds, and validate AI-generated feedback—addressing current transparency gaps and introducing explainability. Expanding into multimodal learning, such as speech recognition for speaking and listening exercises, will significantly expand the range of language skills that can be effectively modeled and supported. Additionally, mitigating biases in LLM outputs through systematic audits and human-AI review (human in the loop) may improve accuracy. Longitudinal studies in real-world classrooms are also critical to empirically validate the system's impact on proficiency gains compared to non-personalised approaches.

By iterating on this foundation system, we hope this framework will develop into a universal adapter for language learning tools, bridging the gap between isolated chatbots and a cohesive task-contextualised personalised language learning ecosystem.

Acknowledgements

This proposal was supported by Council for Science, 3rd SIP JPJ012347, and JSPS Grant-in-Aid for Scientific Research (B) JP20H01722 and JP23H01001, (Exploratory) JP21K19824, JP22KJ1914, (A) JP23H00505, and NEDO JPNP20006.

References

Abyaa, A., Khalidi Idrissi, M., & Bennani, S. (2019). Learner modelling: systematic review of the literature from the last 5 years. *Educational Technology Research and Development*, *67*(5), 1105–1143.

Ait Baha, T., El Hajji, M., Es-Saady, Y., & Fadili, H. (2023). The impact of educational chatbot on student learning experience. *Education and Information Technologies*, *29*(8), 10153–10176.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn: Brain, mind, experience, and school* (Expanded Edition, Vol. 11). National Academies Press.

Brinegar, M. (2023). Chatbots as a Supplementary Language Learning Tool: Advantages, Concerns, and Implementation. *International Journal of Education and Social Science Research*, *6*(6), 223–230.

Capel, A. (2015). The English vocabulary profile. *English Profile in Practice*, *5*(1), 9–27.

Chaiklin, S. (2003). The zone of proximal development in Vygotsky's analysis of learning and instruction. *Vygotsky's Educational Theory in Cultural Context*, 39–64.

Chen, Y., Liu, Q., Huang, Z., Wu, L., Chen, E., Wu, R., Su, Y., & Hu, G. (2017). Tracking knowledge proficiency of students with educational priors. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. Conference on Information and Knowledge Management, Singapore.

Edapurath, V. N. (2022). Enhancing chatbot response relevance through semantic similarity measures. *Journal of Artificial Intelligence and Cloud Computing*, *1*(1), 1–5.

Figueras, N. (2012). The impact of the CEFR. *ELT Journal*, *66*(4), 477–485.

Hariprasath, Paramarthalingam, A., Sundaramurthy, S., & Cirillo, S. (2024). A study on word embeddings in local LLM-based chatbot applications. *2024 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, 232–237.

Herawati, A. (2023). *Personalized Learning in Teaching English as Foreign Language: Limiting the Challenges, Increasing Its Effectiveness*. 3–12.

Huang, W., Hew, K. F., & Fryer, L. K. (2022). Chatbots for language learning—Are they really useful? A systematic review of chatbot‐supported language learning. *Journal of Computer Assisted Learning*, *38*(1), 237–257.

Ishii, Y., & Tono, Y. (2016). A frequency survey of grammar items for the CEFR-J grammar profile. *Proceedings of the 22nd Annual Meeting of the Association for Natural*

Language Processing, 777–780.
Jones, C. (1999). Contextualise & personalise: Key strategies for vocabulary acquisition. *RECALL-HULL*, *11*(3), 34–40.
Kay, J., Bartimote, K., Kitto, K., Kummerfeld, B., Liu, D., & Reimann, P. (2022). Enhancing learning by Open Learner Model (OLM) driven data design. *Computers and Education: Artificial Intelligence*, *3*(2022), 100069.
King, J., & South, J. (2017). Reimagining the role of technology in higher education: A supplement to the national education technology plan. *US Department of Education, Office of Educational Technology*, *10*, 1–10.
Lin, M. P.-C., & Chang, D. (2023). CHAT-ACTS: A pedagogical framework for personalized chatbot to enhance active learning and self-regulated learning. *Computers and Education: Artificial Intelligence*, *5*, 100167.
Masashi, N. (2012). The development of the CEFR-J: Where we are, where we are going. *New Perspectives for Foreign Language Teaching in Higher Education: Exploring the Possibilities of Application of CEFR. Tokyo: Tokyo University of Foreign Studies*, 10–116.
Morita, L. (2017). Why Japan needs English. *Cogent Social Sciences*, *3*(1), 1399783.
Mzury, D. L. (2023). Exploring EFL learners' linguistic error and/or mistakes: Causes, and correction techniques. *Zanco Journal of Human Sciences*, *27*(4), 333–346.
Prain, V., Cox, P., Deed, C., Dorman, J., Edwards, D., Farrelly, C., Keeffe, M., Lovejoy, V., Mow, L., Sellings, P., Waldrip, B., & Yager, Z. (2013). Personalised learning: lessons to be learnt. *British Educational Research Journal*, *39*(4), 654–676.
Ross, S. J. (2008). Language testing in Asia: Evolution, innovation, and policy challenges. *Language Testing*, *25*(1), 5–13.
Ryan, R. M., & Deci, E. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *The American Psychologist*, *55*(1), 68–78.
Takii, K., Flanagan, B., & Ogata, H. (2021). EFL vocabulary learning using a learning analytics-based E-book and recommender platform. *2021 International Conference on Advanced Learning Technologies (ICALT)*, 254–256.
Takii, K., Liang, C., & Ogata, H. (2024a). OKLM: Open Knowledge and Learner Model using educational big data. *International Conference on Computers in Education*, 711–714.
Takii, K., Liang, C., & Ogata, H. (2024b). Open knowledge and Learner Model: Mathematical representation and applications as learning support foundation in EFL. *International Conference on Computers in Education*, 605–614.
Wang, F., Cheung, A. C. K., Neitzel, A. J., & Chai, C. S. (2024). Does chatting with chatbots improve language learning performance? A meta-analysis of chatbot-assisted language learning. *Review of Educational Research*. https://doi.org/10.3102/00346543241255621
Woollaston, S., Flanagan, B., Ocheja, P., Dai, Y., & Ogata, H. (2024). TAMMY: Supporting EFL Translation Practice With an LLM-Powered Chatbot. *Proceedings of the 32nd International Conference on Computers in Education*, *1*, 595–604.
Woollaston, S., Flanagan, B., Ocheja, P., Toyokawa, Y., & Ogata, H. (2025). ARCHIE: Exploring language learner behaviors in LLM chatbot-supported active reading log data with epistemic network analysis. *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 642–654.
Woollaston, S., Flanagan, B., & Ogata, H. (2024). Chatbots and EFL learning: A systematic review. *Joint Proceedings of LAK 2024 Workshops*, 89–98.
Zerkowska, A. M. (2024). *Personalized Language Learning in the Age Of AI. Leveraging Large Language Models for Optimal Learning Outcomes* [University of Gothenberg]. https://gupea.ub.gu.se/handle/2077/81934
Zhai, C. (2023). A systematic review on artificial intelligence dialogue systems for enhancing English as foreign language students' interactional competence in the university. *Computers and Education: Artificial Intelligence*, *4*, 100134.
Zhang, Z., & Huang, X. (2024). The impact of chatbots based on large language models on second language vocabulary acquisition. *Heliyon*, *10*(3), e25370.

Designing Metaverse-based language learning: How pre-service teachers applied the Triple-E framework

Thien-Bao NGUYEN[a], Darmawansah DARMAWANSAH[b*] & Van-Phu NGUYEN[b]
[a]*Department of Foreign Languages and Literature, National Cheng Kung University, Taiwan*
[b]*Global Foreign Language Education Program, Providence University, Taiwan*
*darmawansahlavoe@gmail.com

Abstract: This study investigates how 36 pre-service teachers in Taiwan applied the Triple-E Framework (Engage, Enhance, Extend) during a four-week design-based course using a Metaverse-based platform. Data sources included 264 peer evaluations of the Metaverse design and final group projects. Results show that pre-service teachers excelled in creating visually engaging environments and instructional goals, but faced challenges in aligning their designs with real-world transfer. Thematic analysis of peer feedback revealed that "Engage" was the most frequently recognized dimension, while "Extend" was the least observed. Inferential analyses indicated strong interrelationships between the Triple-E dimensions and a significant association between dimensions. This study contributes to the field by offering a framework-guided, data-informed model for developing digital pedagogical skills in teacher education.

Keywords: Triple-E framework, design-based learning, metaverse, pre-service teachers.

1. Introduction

Technology-enhanced language learning has become an integral component of teacher education programs, equipping future educators with the digital competencies needed to support modern language instruction. Research shows that when pre-service teachers are trained in technology-enhanced language learning, they are more likely to raise technological awareness and increase knowledge (see Almumen, 2023). However, teacher education programs often face challenges in providing adequate pedagogical frameworks and hands-on experience with emerging technologies (Kessler, 2018). Design-based learning has emerged as a prominent pedagogical approach within teacher education, emphasizing creativity, critical thinking, and practical application of educational theories.

This study aims (1) to investigate how pre-service teachers integrated the Triple-E Framework into their preliminary hands-on Metaverse-based language learning design during their initial stages of teacher training, (2) to analyze their peer feedback for evidence of explicit or implicit recognition and reflection of the Triple-E dimensions, and (3) to identify and document specific Triple-E Framework design patterns within the final design of their Metaverse environments. As a result, three research questions are raised:

1. How did pre-service teachers apply the Triple-E Framework in their Metaverse-based language learning design tasks?
2. To what extent did peer feedback reflect the application of the Triple-E dimensions?
3. How were the Triple-E design patterns in their final Metaverse environment designs?

2. Related Work

The Triple-E framework, developed by Kolb (2017), is defined as a measurement scale for educators to evaluate how technology tools integrated into lessons help students engage in, enhance, and extend learning goals. According to Al-Khalidi (2021) the Triple-E framework is

a practical tool that combines instructional strategies, learning goals, and purposeful technology selection.

This framework has been validated through previous research, such as Schatzke (2019). She investigated the content and concurrent validity, as well as the reliability, of the Triple-E rubric as a tool for lesson design. Similarly, Al-Khalidi (2021) investigated how the Triple-E framework was employed as an evaluation tool by ESL (English as a Second Language) teachers to assess technology integration in their lesson planning. This study indicated that teachers held positive views towards using the Triple-E rubric for assessing their lesson plans.

Immersive learning environments, particularly within the Metaverse, offer innovative ways to engage learners by providing interactive and realistic experiences. As cited in Hwang and Lee (2024), the metaverse, an immersive 3D virtual world, contributes to revolutionizing many fields, including education. These virtual settings leverage experiential learning principles, supporting active participation, reflection, and knowledge application (Kolb, 2015). Although recent research highlights increased learner motivation and improved cognitive engagement in Metaverse environments due to authentic simulations and multisensory experiences (Adelana et al., 2023), Adelana et al. (2023) accentuate the need for effective teacher training and policy formulation to address challenges associated with adopting emerging educational technologies like virtual reality.

Hence, research focusing on how pre-service teachers design immersive learning environments to facilitate their teaching lessons needs more attention. The current research examines pre-service teachers' designs of the metaverse to enhance language teaching.

3. Methodology

3.1 Participants

The participants of this study were 36 undergraduate students majoring in foreign language education at a private university in central Taiwan. These students attended an elective course, namely "Fundamentals in Interactive Design".

3.2 Experimental Procedure

The experimental duration lasted four weeks. The students met in person in a classroom for three hours each week. In the first week, the instructor introduced the Triple-E framework and the metaverse platform ZEP (https://zep.us/en) to the students. Students were familiarized with how to integrate the components of the Triple-E framework into a metaverse design to facilitate language learning.

In the second week, students were trained in designing a simple metaverse using the Triple-E framework. In the third week, the students were divided into 10 groups of three to four students. Each group selected a topic to design a metaverse. In the last week, all the groups presented their metaverse design and received feedback from the teacher as well as other students. The feedback based on the criteria of the Triple-E framework was given in the form of a self-designed questionnaire, together with an open-ended question at the end to collect qualitative data for the study.

3.3 Measuring Tools

Figure 2 illustrates a sample student-designed metaverse environment used to facilitate immersive language learning experiences. In the top panel, learners navigate the environment using customizable avatars, promoting a sense of presence. Non-player characters (NPCs), represented by AI-guided agents, provide interactional prompts and simulate real-world interlocutors. Other visual elements, such as signage, pathways, and terminals, serve as contextual backdrops that replicate authentic settings, here, a simulated airport.

In the lower panel, the environment further integrates task-based language learning strategies through embedded quizzes and decision-making challenges. Real-time video and audio features enable synchronous communication among peers and instructors, supporting oral language practice and collaborative interaction. Users' emotional reactions, expressed through avatar animations, while the real-time chat function allows for text-based negotiation of meaning. Together, these multimodal elements support learner engagement (via immersive interaction).

This study's measurement tools comprise 264 peer feedback forms in the second week and nine group-designed Metaverse artifacts. Examples of the final Metaverse design by students are presented in Figure . Thematic analysis was conducted on the feedback data to identify students' learning design experiences. The first and second authors coded 20 percent of the data separately. The inter-rater reliability using Cohen's Kappa was 0.9, indicating high agreement. The rest of the data coding was done by the first author. Table 1. Peer-Feedback Questionnaire presents the self-design questionnaire, based on the components of the Triple-E framework, to collect peer feedback, with 5 indicating "Very Good," 4 indicating "Good," 3 indicating "Average," 2 indicating "Okay," and 1 indicating "No improvement."



Figure 2. The sample of students' design of Metaverse-based language learning

Table 1. *Peer-Feedback Questionnaire*

Item	Score				
1. Interaction with the Environment How well can learners explore and interact with the virtual space? Are there tasks, objects, or NPCs to engage with?	1	2	3	4	5
2. Engagement Elements (Avatars, Challenges, Rewards, Storytelling) Does the design include fun elements like avatars, rewards, challenges, or a story to make learning interesting?	1	2	3	4	5
3. Triple E-Framework: Engage (Capturing Attention) Does the design grab learners' attention and keep them interested in learning?	1	2	3	4	5
4. Language Learning Support (Enhance Learning) Does the design help learners practice language through clear tasks, feedback, and activities?	1	2	3	4	5
5. Real-World Connection (Extend Learning) Can learners use what they learn in real-life situations (e.g., ordering food, traveling, job interviews)?	1	2	3	4	5

What other feedback based on the Triple-E framework do you want to give to your friends?

3.4 Inferential Analysis

To complement the rubric-based evaluations of the student-designed metaverse environments, a series of inferential statistical analyses was conducted to further examine patterns in the data. First, Spearman's rank-order correlation was employed to analyze the relationships among the five items included in the Likert-scale peer evaluation form (n = 264). Second, a chi-square test of independence was conducted to investigate the association between the frequency of Triple-E dimensions identified in peer comments. Third, the Friedman test was applied to assess statistically significant differences in the perceived quality of Engage, Enhance, and Extend elements within the nine final group projects. Following a significant Friedman test result, Wilcoxon signed-rank tests were used for pairwise comparisons between dimensions. A Bonferroni adjustment was applied to control for Type I error in the multiple comparisons.

4. Findings

4.1 How did pre-service teachers apply the Triple-E Framework in their Metaverse-based language learning design tasks?

Quantitative peer evaluation data from 264 responses during Week 2 reveal that students evaluate their peers' metaverse designs as creating engaging and visually appealing environments, as well as integrating the real-world application of language learning tasks in the initial stages. The average peer ratings (on a 5-point scale) are presented in Table .

Table 2. *Descriptive Statistics of Peer Evaluation*

Item	N	Minimum	Maximum	Mean	Std. Deviation
Item1	264	3	5	4.27	.692
Item2	264	2	5	4.19	.733
Item3	264	2	5	4.21	.728
Item4	264	2	5	4.18	.773
Item5	264	2	5	4.18	.758

A Spearman's rank-order correlation was conducted to examine the relationships among the five Likert-scale items. All items were strongly and positively correlated, with Spearman's rho values ranging from .77 to .85, all statistically significant at $p < .001$. The strongest correlation was observed between Item 4 and Item 5 ($\rho = .85$, $p < .001$), suggesting substantial overlap in the constructs they measured. Internal consistency reliability for the five items was excellent, with a Cronbach's alpha of $\alpha = .95$.

4.2 To what extent does peer feedback reflect the application of the Triple-E dimensions?

To explore how peer comments reflected the Triple-E Framework, 264 open-ended responses were coded based on the presence of elements related to Engage, Enhance, and Extend. The largest segment represents students who addressed only one dimension of Engage (39%). A significant portion (34.1%) addressed two elements of Engage and Enhance. A smaller proportion of 18% reflects feedback covering only the Enhance dimension. Notably, 20% of the comments did not relate to the Triple-E dimensions. Only 1% emphasized the Extend dimension. No comments focused on all the Triple-E dimensions.

Regarding peer comments that were thematically coded according to the Triple-E Framework, comments coded under the Engage dimension often emphasized aesthetic appeal and interactivity. Comments categorized under Enhance demonstrated awareness of how designs supported conceptual understanding and language practice. The Extend

dimension, although least frequently mentioned, was reflected in feedback recognizing real-world application.

A chi-square goodness-of-fit test was performed to examine whether the observed frequencies of eight coding categories, including Engage (1), Enhance (2), Extend (3), Engage and Enhance (4), Engage and Extend (5), Enhance and Extend (6), Engage and Enhance and Extend (7), and None (8), were significantly different. The test revealed a statistically significant difference, χ^2 (5, N = 264) = 165.50, $p < .001$, indicating a preference of Engage (N=102), Enhance (N=48), Engage and Enhance (N=59), and None (N=52) over Extend (N=2), and Engage and Extend (N=1).

4.3 How were the Triple-E design patterns in their final Metaverse environment designs?

The final Metaverse learning environments, developed by nine student groups, were evaluated using a rubric based on the Triple-E Framework. Each group project was scored on a scale of 0–2 in three categories: Engage (attention-capturing design), Enhance (language learning support), and Extend (real-world application).

The Engage dimension yielded the highest overall performance, with five groups achieving the maximum score of 2. This finding suggests that most students were successful in designing environments that visually captured learners' attention and integrated interactive elements such as avatars, quests, or gamified tasks. The relatively strong performance in this dimension reflects a focus on surface-level engagement and the motivational aspects of digital design.

In contrast, fewer groups reached the top score in the Enhance dimension. Only four out of nine groups achieved a score of 2, indicating that while some students embedded meaningful instructional support, many designs demonstrated only partial or limited enhancement. The Extend dimension was the most underdeveloped, with just one group attaining a top score of 2 and three groups receiving a score of 0. This trend highlights a recurring challenge in creating virtual learning activities that explicitly promote real-world application or knowledge transfer beyond the digital environment.

A Friedman test was conducted to assess differences between Engage, Enhance, and Extend ratings across nine final student designs. The test revealed a statistically significant difference among the three dimensions, $\chi^2(2, N = 9) = 9.91$, $p = .007$. Mean scores indicated that Engage (M = 1.56) and Enhance (M = 1.44) were rated higher than Extend (M = 0.67).

Post-hoc Wilcoxon signed-rank tests revealed that both Engage (W = 0.00, p = .023) and Enhance (W = 0.00, p = .020) scored significantly higher than Extend. At the same time, there was no significant difference between Engage and Enhance (p = .564). The results revealed that students' designs were significantly more effective at engaging and enhancing learning than at extending learning opportunities. No significant difference was found between engagement and enhancement performance, but extending learning was clearly the weakest dimension.

5. Discussion and Conclusion

This study examined how pre-service teachers implemented the Triple-E Framework in designing Metaverse-based learning environments, offering valuable insights into the evolving digital pedagogy of teacher candidates. Peer evaluation data from Week 2, alongside the rubric-based analysis of final group projects, consistently pointed to the Engage dimension as the most effectively realized aspect of students' work. Across both data sources, pre-service teachers demonstrated strong capabilities in creating visually stimulating, interactive spaces that successfully captured learner attention through elements such as avatars, narrative-driven tasks, and gamified activities.

However, a statistically significant disparity was observed across the three dimensions, Engage, Enhance, and Extend, with the Extend dimension receiving the lowest evaluations. While many student designers showed competence in developing tasks that supported language learning and cognitive engagement (Enhance), they faced notable challenges in designing experiences that promoted knowledge transfer to real-world contexts (Extend).

These results highlight the importance of more targeted scaffolding within teacher preparation programs, particularly in supporting pre-service teachers to conceptualize and implement designs that extend learning beyond the virtual environment. The findings echo earlier qualitative and quantitative analyses, reinforcing a recurring pattern: while students excel at crafting engaging and instructional virtual experiences, they often struggle with designing for real-world applicability. This highlights a critical gap in the pedagogical reasoning that underpins immersive learning design and signals the need for additional training, guided examples, and reflective activities focused on the Extend component of the framework.

Despite some limitations, including a relatively short intervention period, reliance on self- and peer-assessment, and a simplified coding rubric, the study contributes to the growing body of research on immersive learning and digital pedagogy. It affirms the utility of the Triple-E Framework as a scaffold for instructional design in Metaverse environments and demonstrates how peer feedback can serve as a formative tool in the design-based learning process. Future research should explore how extended exposure to iterative design tasks and longer-term collaborations might enhance pre-service teachers' ability to integrate the Extend dimension effectively. Longitudinal studies could also investigate how these design skills translate into actual classroom practice, particularly in language education settings where immersive and experiential learning hold increasing relevance.

Acknowledgements

This study is supported in part by the National Science and Technology Council of Taiwan under contract number NSTC 114-2410-H-126-002-MY2

References

Adelana, O. P., Ayanwale, M. A., Ishola, A. M., Oladejo, A. I., & Adewuyi, H. O. (2023). Exploring pre-service teachers' intention to use virtual reality: A mixed-method approach. *Computers & Education: X Reality*, *3*, 100045. https://doi.org/10.1016/j.cexr.2023.100045

Al-Khalidi, I. (2021). Using the Triple E framework as an instructional design tool for the assessment of technology integration in a variety of TESOL contexts. *The Asian ESP Journal*, *17*(2), 48–73.

Almumen, H. (2023). Technology and multimodality in teaching pre-service teachers: Fulfilling diverse learners' needs. *Technology, Knowledge and Learning*, *28*(2), 745–767. https://doi.org/10.1007/s10758-021-09550-1

Hwang, Y., & Lee, J.-Y. (2024). Exploring pre-service English teachers' perceptions and technological acceptance of metaverse language classroom design. *SAGE Open, 14*(4). https://doi.org/10.1177/21582440241300543

Kessler, G. (2018). Technology and the future of language teaching. *Foreign Language Annals*, *51*(1), 205–218. https://doi.org/10.1111/flan.12318

Kolb, D. A. (2015). *Experiential learning: Experience as the source of learning and development* (Second). Pearson Education.

Kolb, L. (2017). *Learning first, technology second: An educator's guide to designing Authentic lessons*. International Society for Technology in Education.

Schatzke, S. E. (2019). *A validation study of the Triple E Rubric for lesson design: A measurement tool for technology use in the classroom* [Doctoral dissertation, University of North Texas]. UNT Digital Library. https://digital.library.unt.edu/ark:/67531/metadc1505198/

Web Application to Help Learn English Using Medical Bilingual Corpus - Updates to Improve Learners' Writing Skills -

Riku YASUDA[a*], Yoshinori MIYAZAKI[b], Motoko ASANO[c] & Miho FUJIEDA[c]
[a]*Graduate School of Informatics, Shizuoka University, Japan*
[b]*College of Informatics, Shizuoka University, Japan*
[c]*Faculty of Medicine, Osaka Medical and Pharmaceutical University, Japan*
*yasuda.riku.20@shizuoka.ac.jp

Abstract: In the field of medical education, a deep understanding of domain-specific expressions is essential for reading and writing academic papers in English. To address this challenge, we are developing an English learning support web application with a bilingual corpus of medical texts. This paper presents two enhancements to the system designed to support learners' academic writing skills. First, we introduced a writing support function that suggests similar sentences based on user-generated English input. Second, we enhanced the accuracy of grammar-based categorization by integrating a BERT-based grammatical detection algorithm that leverages contextual understanding. These improvements aim to provide a more effective and domain-sensitive writing support experience for users engaging with medical English materials.

Keywords: Medical English, bilingual corpus, web application, reading and writing aid

1. Introduction

In the field of medical education, the ability to effectively utilize English is particularly important for acquiring up-to-date medical knowledge, such as through clinical conferences or reading recent academic papers in English. This reflects the domain of English for Medical Purposes (EMP), a specialized branch within English for Specific Purposes (ESP). Based on this background, our research group developed MEESUS (Medical English Education SUpport System) to support EMP learning among medical English learners, using a bilingual corpus of medical research papers (titles and abstracts) (Asano et al., 2022). The development process involved not only the provision of data by English instructors from a medical university but also incorporated ideas proposed by faculty members in the fields of medicine, engineering, system development, English education, and practicing physicians. As a result, the project was shaped by a multidisciplinary task force that integrates diverse expertise across academic domains. The current version of the system is designed with a focus on supporting reading and writing in English; however, this study aims to expand its functionality to include writing support, with the goal of providing more comprehensive assistance for EMP learning.

2. Literature Review

Related studies to this research include previous work on concordancers and section-based language support, which are key features of the proposed system.

AntConc, developed by L. Anthony, is a free text analysis tool designed for researchers in linguistics and literary studies, and it includes functions such as concordance searches for words and phrases (Anthony, n.d.). JECPRESE is a software tool that places the search term in the center of a keyword-in-context (KWIC) display when retrieving expressions from a corpus. It is designed for transcribed oral presentation data and allows users to search both English and Japanese sentences depending on the speaker's communicative intent (Kunioshi

et al., 2012). AWSuM is a writing support system that utilizes rhetorical “moves” (i.e., communicative intentions) to suggest frequently used word bundles by section of a research article. While it also includes concordancer functions and provides search capabilities, it does not support bilingual (English-Japanese) output (Mizumoto et al., 2017).

3. Medical English Learning System

Figure 1 shows the data structure of the bilingual corpus, which consists of the titles and abstracts of medical research papers. The corpus includes 3,920 such documents (title and abstract pair) and a total of 49,884 English sentences, all aligned with Japanese translations (NEJM Group, n.d.; Nankodo Co., Ltd., n.d. https://www.nejm.jp/). All corpus data has been reviewed and verified by medical faculty experts to ensure accuracy and reliability. Unlike generative AI tools, this expert-validated dataset provides trusted, domain-specific expressions without such risks. Each English document is segmented into six sections: title, background, methods, results, conclusion, and conclusion (research information). By utilizing section-specific information linked to each sentence, learners can study expressions within distinct rhetorical contexts. In Figure 1, for example, the sentence with ID = 1900236 appears as the fourth sentence of the fourth document and belongs to the “conclusion” section.

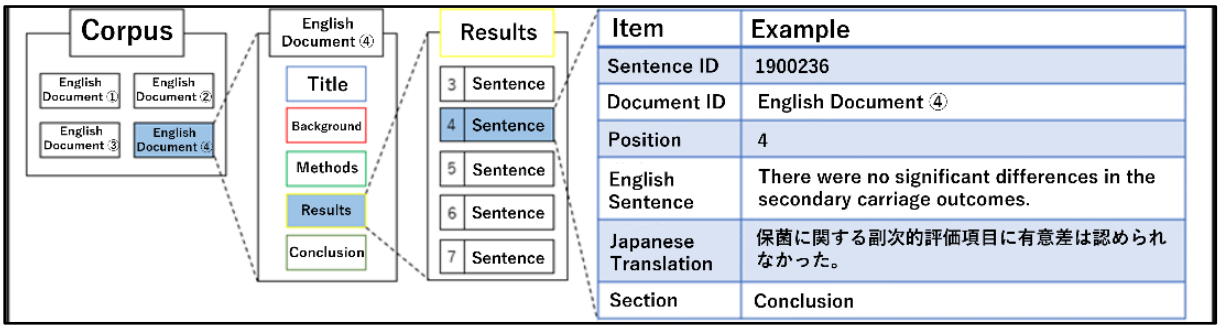


Figure 1. Data Structure and Examples

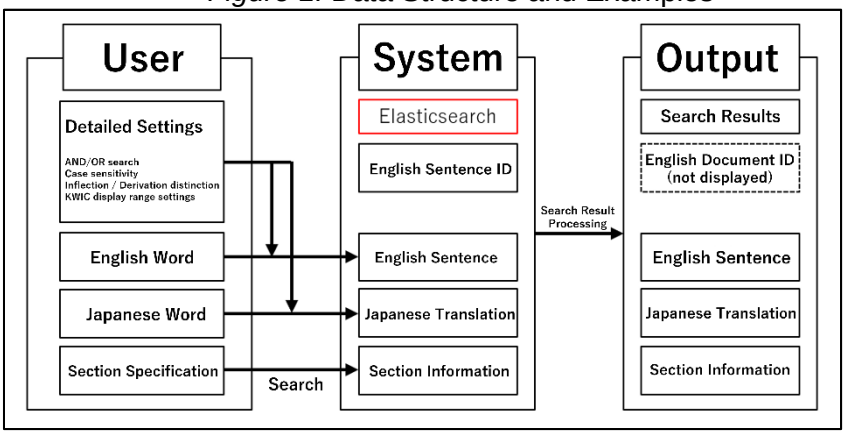


Figure 2. System Architecture

An overview of the system is presented in Figure 2. The search engine is built using Elasticsearch (Elasticsearch, n.d.). User input is converted into an Elasticsearch query, which is then used to retrieve relevant data from the corpus structured (Figure 1). The system offers various search methods, including single search, combined search, and cross-lingual search that allows users to input a mix of Japanese and English in each search field. Search options include specifying the document section (e.g., background or methods), setting the number of words for KWIC display, and adjusting parameters for variations in expressions (Figure 3). The system also implements a suggestion feature. This feature uses Levenshtein distance to recommend English words that are similar to the user's input. For example, if a user mistakenly enters “hospitul,” the system will display the interface shown in Figure 4. By clicking the “Did you mean” prompt, the system suggests possible correct spellings. This function enables

users to receive appropriate word suggestions even when typographical errors are present in the search terms. Search results are initially displayed in KWIC format, with each hit shown at the sentence level. Users can then access additional features, such as viewing the full English sentence with its Japanese translation, or displaying the entire document to which the hit sentence belongs. An example of these functions is shown in Figure 5.

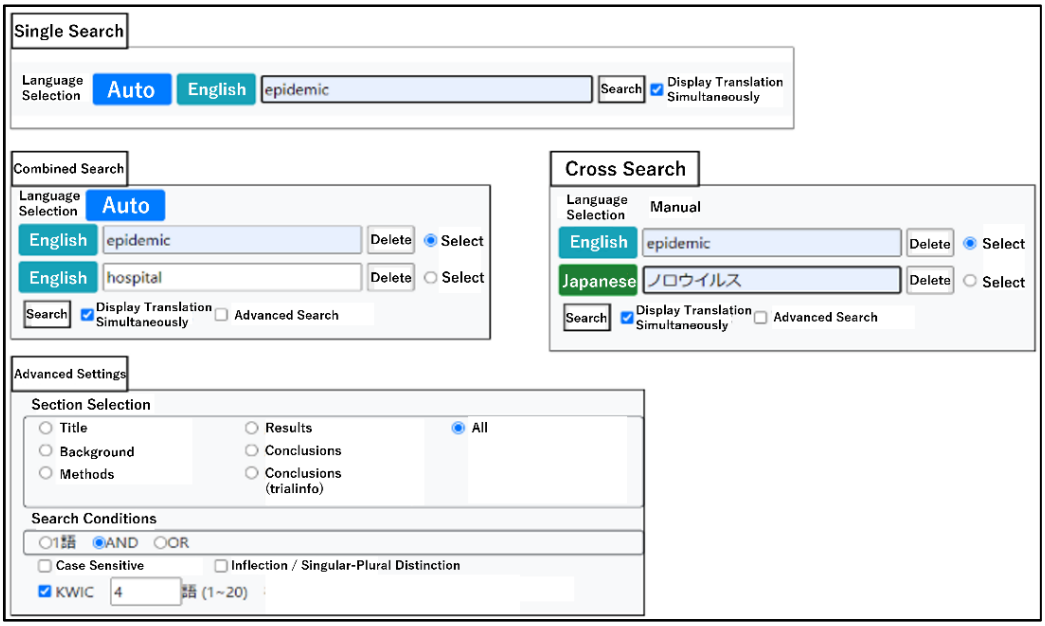


Figure 3. Various Search Methods

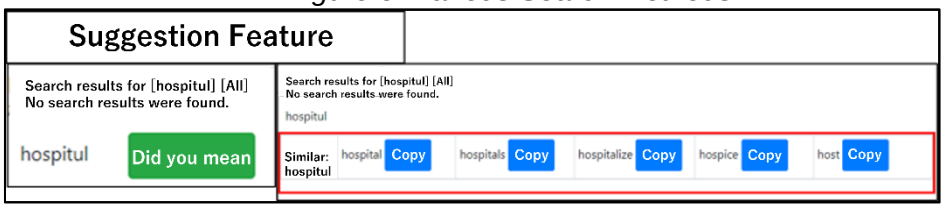


Figure 4. Suggestion Feature

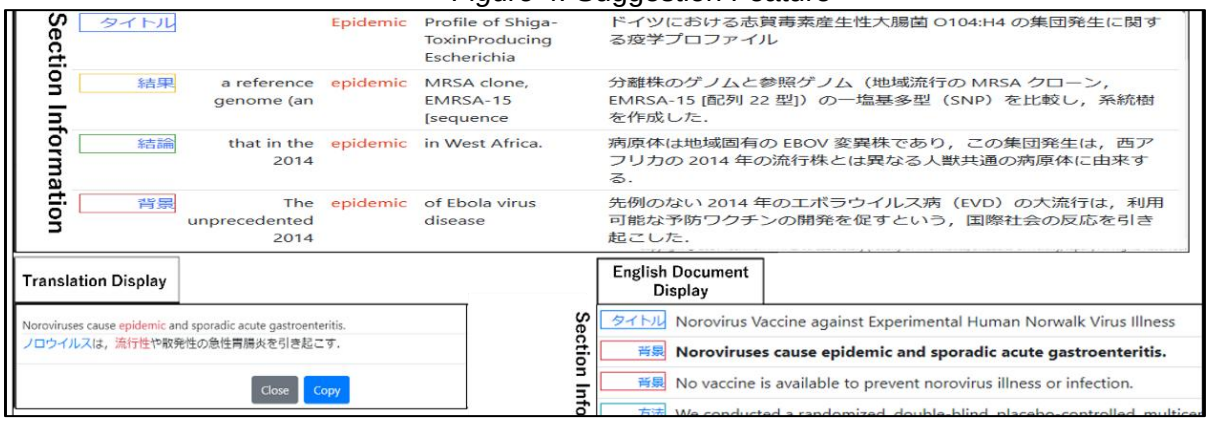


Figure 5. Search Results

4. Additional Features

4.1 Similar Sentence Search Function

In order to support academic writing more effectively, the system includes a function that presents similar sentences based on either search results or user-inputted English sentences. This feature was developed in response to user feedback.

The authors incorporated the EWSS system (Miyazaki et al., 2014) to add a feature that searches for sentences similar to either user-inputted text or search result sentences. By clicking the "Similar Sentence Search" button or selecting a sample sentence from the search

results, a modal window appears. The Similar Sentence Search function allows users to toggle between general English and medical-specific English, and the input sentence is automatically populated in the search field, making it easy to conduct a similarity-based search. Users can also view their search history for repeating a previously searched sentence.

In terms of system operation, the input sentence is analyzed using the morphological analyzer TreeTagger to extract the base forms and part-of-speech information of each word. Synonyms are also utilized to generate n-grams—sequences of n consecutive words—from the sentence. Based on this information, the system calculates sentence similarity in a vector space, allowing it to retrieve example sentences that closely match the user's intended meaning. Figure 6 shows the search results screen. The input sentence is displayed at the top, with content words (nouns, verbs, adjectives, and adverbs) highlighted. Synonyms for each content word in the input sentence are also shown, based on WordNet. By clicking the checkbox labeled "Add," users can save selected sentences to "My List."

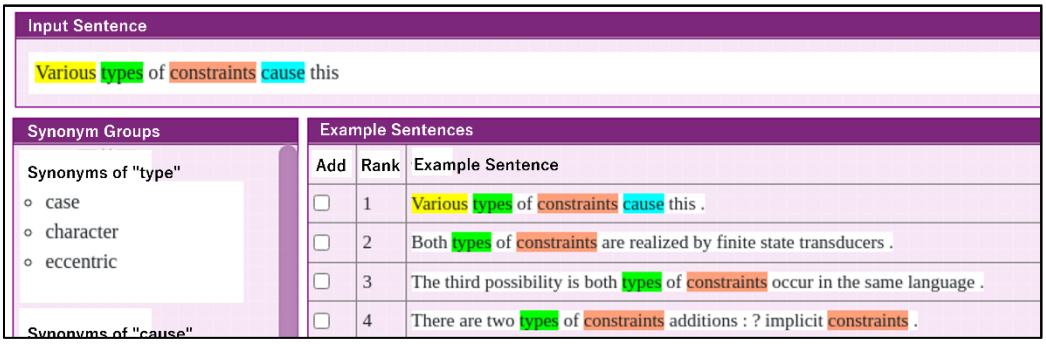


Figure 6. Search Results from the Similar Sentence Search Function

For the current system, we adapted the dataset used in MEESUS to specifically support learning in the domain of EMP. Simply adapting the dataset was not sufficient to present users with the medical-specific English sentences they were seeking. Therefore, we applied customized weighting to selected medical terms. To improve the scoring mechanism, we utilized a predefined list of medical terms extracted from domain-specific resources. When a sentence in the dataset contains a medical term that also appears in the user's input, the similarity score for that sentence is specially increased. By utilizing this feature, users are expected to enhance their writing skills in specialized medical contexts.

4.2 Grammar-Based Display Enhancement

Interrogative Form	0%	To-Infinitive	9%	Modal Verb	3%
Passive Voice	6%	Bare Infinitive	100%	Relative Pronoun	45%
Subjunctive Mood	100%	To-Infinitive + Bare Infinitive	11%	Existential "There" Construction	17%
Progressive Aspect	22%	Comparative	0%		
Perfect Aspect	0%	Gerund	83%		

Figure 7. Error Rates in Grammatical Category Detection by Previous Studies

To further enhance the usefulness of the Similar Sentence Search function, we focused on improving the accuracy of grammar-based categorization. In follow-up studies to Miyazaki et al. (2014), grammatical item detection methods using regular expressions in Python and BACT were employed. However, these approaches frequently misidentified complex grammatical constructs such as the subjunctive mood, bare infinitives, and gerunds (see Figure 7). These misclassifications were primarily due to the reliance on simple pattern matching, which lacked contextual and semantic understanding. To significantly improve the accuracy of grammatical item detection, we combined advanced Python libraries such as spaCy and NLTK with BERT to analyze English sentences. Specifically, spaCy's dependency parsing functionality was used to clarify syntactic relationships within sentences, while the machine learning model enabled context-aware classification of grammatical elements. In the case of gerund detection in particular, the approach goes beyond simple sentence pattern identification by taking into

account the overall meaning of the sentence and its relationship with verbs. This allows for more nuanced distinctions that were difficult to achieve with previous methods. Furthermore, to enhance detection accuracy, we adopted a supervised learning approach and trained the model using a large-scale corpus. This approach is expected to achieve high detection accuracy even for grammatical items such as the subjunctive mood and bare infinitives, which have previously yielded limited success with traditional methods.

4.3 Comparative Evaluation of the Grammar-Based Display Accuracy

Figure 8 presents the results of a comparison of grammatical item detection accuracy between the proposed method and previous studies. The blue bars represent the results from prior research, while the orange bars show the performance of the proposed method. For the comparison experiment, 200 sentences were randomly selected from a technical literature corpus, following the same procedure used in previous studies. To ensure a fair evaluation, each sentence was manually annotated with grammatical items, and detection accuracy was assessed based on the agreement between these annotations and the model outputs.

	Previous Studies	Proposed Method			Previous Studies	Proposed Method
Interrogative Form	100%	100%		To-Infinitive + Bare Infinitive	89%	96%
Passive Voice	94%	100%		Comparative	100%	100%
Subjunctive Mood	0%	100%		Gerund	17%	94%
Progressive Aspect	78%	100%		Modal Verb	97%	100%
Perfect Aspect	100%	100%		Relative Pronoun	55%	91%
To-Infinitive	91%	95%		Existential “There” Construction	83%	100%
Bare Infinitive	0%	100%				

Figure 8. Comparative Accuracy of Grammatical Category Detection

The results confirm that the proposed method significantly improved detection accuracy for grammatical items that had been challenging in previous studies, particularly the subjunctive mood, bare infinitives, and gerunds. This substantial enhancement is primarily attributed to BERT’s ability to understand context, allowing the model to consider relationships between surrounding words for more accurate classification. Overall, detection accuracy also improved across other grammatical categories. The average detection accuracy increased from 77% in previous studies to 96% with the proposed method. Although this experiment was conducted using technical literature, BERT's pretraining on a wide variety of text genres suggests a certain level of generalizability to the medical domain. However, to achieve even higher accuracy in medical contexts, fine-tuning with additional domain-specific data will be necessary, and this remains an area for future work.

5. Experimental Evaluation and Analysis

Undergraduate and graduate students from the Faculty of Informatics at a certain university were invited to use the system, and we collected both usage logs and questionnaire responses. The experiment involved 10 participants, and based on the results, we analyzed the effectiveness of the newly added features. Each participant completed four tasks using the system, followed by a post-experiment usability questionnaire. The task details are:
Task 1. Writing about the topic of a research paper, Task 2. Investigating a research paper, Task 3. Paraphrasing task, and Task 4. Error correction task.

The usage statistics for the Similar Sentence Search options follows, General English: 32 (times), Medical-Specific English: 5 (times), and Number of My List Usages: 19 (times).

It was found that the general English search option was used more frequently. This is likely because the example sentences for the paraphrasing task were written in general English. The ratio of "My List" usage to the total number of searches was approximately 51%, indicating that the feature was actively used to support task completion. According to the questionnaire responses on feature usage, the Similar Sentence Search function was used

by all participants in both Task 3 and 4. This suggests a strong demand for the feature in the tasks. The responses to the questionnaire item “Do you feel that using the system reduces the difficulty of writing academic texts?” showed that 30% of participants strongly agreed, 30% agreed, and 30% slightly agreed. Only 10% slightly disagreed, and no one strongly disagreed. Tables 3 and 4 present examples of feedback regarding the usability of this feature. 90% gave positive feedback, and the questionnaire responses indicated a 100% usage rate, suggesting that the system's support was highly valued. Below are the concrete comments: “I found it useful for rewriting English sentences,” “It is helpful when I want to express something in a different way,” “I would appreciate it if the system also displayed similar sentences based on grammatical rules, in addition to similar vocabulary.” The feature was particularly well-received for tasks such as paraphrasing, where it effectively assisted in developing writing skills. However, some users noted that the current version merely lists similar sentences, and suggested the system could categorize the results by grammatical rules.

6. Concluding Remark

In this study, we conducted an experiment using a developing medical English learning support system. Based on the feedback obtained, we implemented enhancements focused on supporting academic writing. The Similar Sentence Search function was designed to support writing skills by presenting similar sentences based on either search result sentences or user-inputted text. With some problems detected, we adapted a medical literature dataset and applied customized weighting to medical terms, to provide more domain-specific sentence suggestions. In addition, we aimed to further enhance writing assistance by improving the accuracy of the grammar-based categorization function. By combining BERT’s contextual understanding with syntactic parsing using spaCy, the overall average detection accuracy increased from 77% to 96%. The results of the pilot experiment using the enhanced system showed that the Similar Sentence Search function had a 100% usage rate, and 90% of participants rated it positively in the questionnaire, suggesting that the system successfully offers specialized writing support tailored to users’ needs. As for future directions, we plan to conduct a field-based validation study in actual educational settings to objectively evaluate the effectiveness of the newly implemented features for qualitative and quantitative feedback. This will allow us to assess the effectiveness, usability, and practical value of the system as a learning support tool. The findings will help identify areas for further improvement and reveal new user needs, which will inform the development of the next version of the system.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 23K02800.

References

Anthony, L. (n.d.). AntConc [Computer software]. Retrieved August 13, 2025, from https://www.laurenceanthony.net/software/antconc/

Asano, M., Nakano, M., Miyazaki, Y., & Fujieda, M. (2022). Introducing a bilingual corpus database system of medical abstracts for exploring academic connotations of words: A case study of first-year medical students. *Journal of Medical English Education*, *21*(1), 18–26.

Kunioshi, N., Noguchi, J., Hayashi, H., & Watanabe, Y. (2012). An online support site for preparation of oral presentations in science and engineering. *European Journal of Engineering Education*, *37*(6), 600–608. https://doi.org/10.1080/03043797.2012.738171

Miyazaki, Y., Tanaka, S., & Koyama, Y. (2014). Development of a corpus-based web application to support writing technical documents in English. In *Proceedings of E-Learn 2014* (Vol. 2014, No. 1, pp. 1371–1380).

Mizumoto, A., Hamatani, S., & Imao, Y. (2017). Analyzing applied linguistics research articles by integrating moves and lexical bundles: Toward the development of an academic writing support tool. *Journal of English Corpus Studies*, *23*, 21–32. [in Japanese]

Investigating the Impact of AI Generated Mnemonic Illustrations on Learning Japanese Kanji

Mayur BHURLE*, Vishwas BADHE & Ramkumar RAJENDRAN
Centre for Educational Technology, Indian Institute of Technology Bombay, India
*myrbhurle@gmail.com

Abstract: Learning Japanese Kanji presents a significant challenge for novice learners due to the vast number of characters, their visual complexity, and a lack of relatable learning resources. While various strategies exist, the use of illustrations in Kanji learning, despite their proven benefits, remains underexplored. This study investigates the impact of integrating AI-generated mnemonic illustrations with textual mnemonics on learners' Kanji attainment, cognitive workload, and learning attitudes. In a quasi-experimental study with novice learners, a control group using only text-based mnemonics was compared against experimental groups that used AI-generated illustrations. Quantitative analysis, corroborated by qualitative data, revealed that while the illustrations did not produce a statistically significant improvement in post-test scores, they had a significant positive impact on reducing the perceived cognitive workload. This indicates that despite theoretical support for improved learning outcomes when adding images to text, this principle may not directly enhance recall for logographic languages like Kanji. Nevertheless, the findings suggest that while AI-generated illustrations may not directly boost short-term recall, they effectively make the learning process less cognitively demanding.

Keywords: Japanese Kanji, Mnemonic Illustrations, Kanji Learning Strategies, Cognitive Workload, Learning Attitude, Generative AI

1. Introduction

Kanji, one of three Japanese writing systems, is crucial for achieving literacy in the Japanese language (Hiraga, 2006). The global popularity of learning Japanese has surged, with over 3.7 million learners worldwide facing the challenge of mastering Kanji's vast character set, visual complexity, and a scarcity of tailored learning resources (Japan Foundation, 2021; Shekarabi & Tajfirooz, 2023). While learners employ various strategies such as mnemonics, morphological analysis, and context-based learning (Mori & Shimizu, 2007), many existing methods underutilize illustrations. This is a notable omission, given that cognitive theories like Dual Coding Theory (Paivio, 1986; Patel, 2023) and the Cognitive Theory of Multimedia Learning (Mayer, 2024) suggest that combining images with text enhances memory and understanding.

Mnemonic strategies are particularly suitable for novices but can fail when the mnemonic itself is abstract, causing learners to forget the Kanji's meaning (Rose, 2013). This can increase cognitive workload and negatively impact student attitudes (Sweller, 2023). To address this, our study combines mnemonics with mnemonic illustrations, which visually represent key concepts to make them more concrete. Manually creating such illustrations is not only time-consuming but also fails to cater to diverse learners. This research utilizes text-to-image Generative AI (GenAI), which facilitates the creation of personalized mnemonic illustrations, allowing learners to connect with visuals that are culturally and personally relevant. This study investigates the impact of these AI-generated illustrations on novice learners' Kanji attainment, cognitive workload, and attitudes, focusing specifically on the identification of Kanji meanings and not on the pronunciation or writing aspect of Kanji. The

findings offer tangible instructional strategies, empowering educators and students to co-create customized visual aids in the classroom that cater to diverse learning styles and backgrounds.

2. Related Work

Kanji consists of logographic characters that range from simple forms, such as 一 (meaning “one”) with a single stroke, to complex ones like 員 (meaning “member”) with multiple strokes (Seely & Henshall, 2016). Proficiency requires memorizing hundreds of these characters, and learners adopt a variety of strategies to achieve this. Mori and Shimizu (2007) identified six primary self-reported strategies: mnemonic strategies, morphological analysis, rote memorization, context-based learning, metacognitive strategies, and helplessness. Among these, mnemonic strategies are particularly effective for novice learners (Rasiban, 2016). Recent pedagogical innovations include Naritachi, which focuses on Kanji formation (Aneros et al., 2020), Cognitive Illustration based on formation processes (Sutedi & Juangsih, 2024), and gamification approaches (Sonjaya et al., 2024). Despite these advances, a gap remains in effectively leveraging illustrations, even though frameworks such as Dual-Coding Theory (Paivio, 1986) and the Cognitive Theory of Multimedia Learning (Mayer, 2024) suggest that integrating illustrations with text can improve learning outcomes.

Illustrations in education can be decorative, representational, organizational, or mnemonic, each serving a distinct role. Among these, mnemonic illustrations have demonstrated the strongest positive effects on text learning (Carney & Levin, 2002). They are designed with systematic mnemonic components to enhance recall, often through keyword integration, where important terms from the text are visually represented in a way that is relevant and meaningful within the learning context (Erlan, 2024; Carney & Levin, 2002; Levin, 1983). Personalization remains a challenge, as static resources such as Kanji Pict-O-Graphix (Rowley, 1992) cannot accommodate the diverse needs of learners. Generative AI (GenAI) offers a promising solution by enabling the creation of personalized mnemonic illustrations. Although research on GenAI for Kanji acquisition is still emerging, its potential is evident from applications such as visual storytelling (Han & Cai, 2023) and AI-assisted video production for Japanese learning (Takeda-Kolb & Ohsawa, 2024). This study addresses this gap by investigating the impact of AI-generated mnemonic illustrations on the learning outcomes of novice Japanese as a Foreign Language (JFL) learners. The research is guided by the following questions: (RQ1) How do AI-generated mnemonic illustrations impact JFL learners' attainment of Kanji characters?; (RQ2) What is the impact of AI-generated mnemonic illustrations on the cognitive workload experienced by JFL learners during Kanji learning?; and (RQ3) How do AI-generated mnemonic illustrations affect JFL learners' attitudes towards learning Kanji?

3. Learning Intervention (Kanji Quest)

The learning intervention, "Kanji Quest," involved a set of 40 Japanese Kanji characters selected based on specific criteria. To ensure variety, ten characters were chosen from each of the four main types (ideographic, pictographic, phonetic, combined ideographic) as defined by Hiraga (2006).

Table 1. *Selected Kanji characters for Kanji Quest*

Kanji Type	Simple Kanji	Complex Kanji
Ideographic	一, 四, 末, 下, 三, 上, 八, 二, 十	百
Pictographic	元, 月, 犬, 火	員, 見, 金, 雨, 会, 京
Phonetic	円, 千, 仕	時, 作, 思, 地, 気, 体, 社
Combined Ideographic	友, 北, 立, 父	料, 夜, 明, 歩, 旅, 名
Total	20	20

All Kanji were from JLPT levels N4 and N5, suitable for beginners. The set was balanced for difficulty, containing 20 "simple" Kanji (≤5 strokes) and 20 "complex" Kanji (6–10 strokes), as stroke count impacts learning difficulty (Matsunaga, 2016). A Japanese language expert validated the selection. The complete list of selected Kanji is presented in Table 1. A website, Kanji Quest, was developed in two versions to test the intervention. The first version presented the Kanji, its meaning, and a textual mnemonic adapted from Seely and Henshall (2016). The second version included the same information plus an AI-generated mnemonic illustration (as shown in Figure 1).

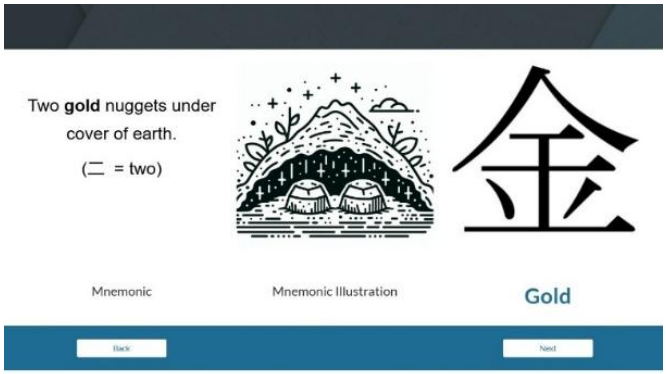


Figure 1. Depiction of the Kanji Quest interface.

These illustrations were produced in a two-step process. First, a custom tool, "Prompt Builder for Mnemonic Illustration" using the MistralAI API, converted mnemonics into structured text-to-image prompts based on prompt engineering guidelines (Garg et al., 2025; Garg & Rajendran, 2024). Second, these prompts were input into a Text-to-Image AI tool (MS CoPilot) to generate the final illustrations. Given the scale of the intervention, all illustrations were generated prior to the study. While not personalized for each individual in real-time, the prompts were specifically designed to create visuals that would be culturally and contextually relevant to the study's participant demographic. A Japanese language expert validated all resulting images for relevance and appropriateness. This approach ensured all students in the experimental groups viewed the same validated set of images.

4. Research Design and Data Collection

This study used a quasi-experimental design (Shadish et al., 2002) to compare three Kanji learning strategies. After obtaining ethical approval and student consent, 84 engineering undergraduate students (average age 20; 66 male, 18 female) with no prior Japanese knowledge were randomly assigned to one of three groups. The control group (Group M, n = 27) used only textual mnemonics. The first experimental group (Group MI, n = 26) used mnemonics with AI-generated illustrations. The second experimental group (Group MID, n = 31) used mnemonics, AI-generated illustrations, and also drew each Kanji character. The drawing component was included because writing activates sensorimotor systems, which has been shown to enhance learning and memory across age groups (Kontra et al., 2012; Macedonia, 2014). The overall research design, outlining the sequence of activities, is depicted in Figure 2.



Figure 2. Research Design.

The procedure began with a pre-test to assess participants' existing Kanji knowledge. Any student demonstrating prior knowledge was excluded. The remaining participants received a brief introduction to the Japanese language, followed by a 20-minute learning session using their assigned version of the Kanji Quest platform. Group M used a version

containing only mnemonics and Kanji characters, and participants were instructed to read the mnemonic and try to connect it to the meaning of the character. Group MI used a version of Kanji Quest (shown in Figure 1) that included illustrations alongside the mnemonics and Kanji. These participants were instructed to use both the mnemonic and the illustration to connect the Kanji character to its meaning. Group MID followed the same instructions as Group MI, but in addition, participants were asked to draw each Kanji character without concern for stroke order. After the learning session, participants were given a 15-minute break. They then completed a post-test, the NASA Task Load Index (NASA-TLX) survey to measure perceived cognitive workload (Hart & Staveland, 1988), and an Attitude Questionnaire (Mori & Shimizu, 2007; Shimizu & Green, 2002) to assess perceptions toward Kanji learning. Finally, semi-structured interviews were conducted with 11 students (M: 3, MI: 4, MID: 4) to gather qualitative insights. The semi-structured interviews explored three key areas: participants' learning strategies using the provided aids (illustrations, mnemonics, and drawing), the challenges they encountered, and their suggestions for improving the intervention.

Quantitative analysis was conducted on pre-test, post-test, NASA-TLX, and Attitude Questionnaire scores, with findings triangulated using data from semi-structured interviews. Kanji learning attainment was examined using Welch's ANOVA on post-test scores (total, simple, and complex items). The perceived cognitive workload was analyzed using Welch's ANOVA on weighted NASA-TLX scores, as the data met normality but not homogeneity of variance assumptions. Attitude scores were analyzed using a one-way ANOVA with Tukey's post-hoc test for dimensions meeting parametric assumptions (Total Score, Usefulness, Affective Orientation) and the Kruskal-Wallis test for non-parametric dimensions (Aptitude, Difficulty). To support and interpret the quantitative results, interview responses were analyzed using the selective coding method from content analysis (Mayring, 2015, p. 95).

5. Results

Post-test analysis showed no significant difference in Kanji learning attainment among the three groups (M, MI, and MID) for total scores ($p = .896$). Separate analyses for simple Kanji ($p = .778$) and complex Kanji ($p = .823$) also found no significant differences. Although all groups showed learning gains compared to their low pre-test scores, adding AI-generated illustrations (MI) or combining them with drawing (MID) did not improve attainment over mnemonics alone (M).

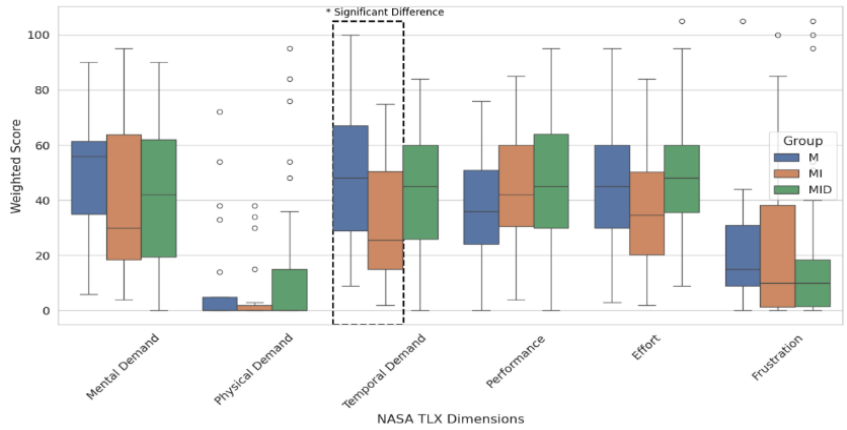


Figure 3. Weighted NASA TLX Scores Across Dimensions for Each Group

Cognitive workload analysis revealed a significant difference in total weighted NASA-TLX scores ($p = .039$). Post-hoc tests indicated that the MI group reported significantly lower workload than the MID group ($p < .05$), with interviews attributing this to the demanding nature of the drawing task. Qualitative feedback suggested that the drawing activity in the MID group was time-consuming, reducing the time available for learning. Further analysis of NASA-TLX dimensions (Figure 3) showed that the MI group also experienced significantly lower temporal demand than the M group ($p < 0.05$). Interviews supported this finding, as M group participants

reported feeling rushed, while MI participants stated that the illustrations helped them quickly connect Kanji characters to their meanings. For learning attitudes, no significant differences were observed in overall scores across the three groups ($p > 0.05$), indicating that the illustrations did not lead to a significant improvement in attitudes toward learning Kanji.

6. Discussion and Conclusion

This study examined the impact of AI-generated mnemonic illustrations on novice JFL learners to determine whether adding such illustrations to traditional mnemonics would positively influence learning outcomes. While the illustrations did not significantly improve overall Kanji attainment, the findings suggest that adding illustrations to text may not always enhance learning outcomes for logographic languages like Kanji, contrary to expectations based on theoretical frameworks. However, the results still support their use in Kanji learning because they significantly reduced perceived temporal demand. This was corroborated by the semi-structured interviews, as students in the MI group, who used illustrations, reported feeling less rushed compared to those using only mnemonics. The illustrations, however, did not lead to a significant improvement in overall learning attitudes toward Kanji.

Although direct comparisons between strategies showed no significant differences in attainment, learning gains were evident across all groups, as post-test scores improved from pre-test levels. This aligns with research by Hasanah et al. (2024), which reported notable improvements in Kanji attainment using visual image and pictograph strategies. Collectively, evidence indicates that visual aids such as mnemonic illustrations can positively influence aspects of the Kanji learning process, either by supporting direct attainment or by reducing perceived cognitive workload.

These results have practical implications for Kanji pedagogy. Students can use AI to generate personalized mnemonic illustrations tailored to factors such as age, gender, cultural background, and artistic preferences, making the learning experience more engaging and relevant. Teachers can also use this technology to design illustrations suited to specific student groups and incorporate them into classroom instruction, helping learners more effectively connect Kanji characters with their meanings. This study has certain limitations. It focused solely on identifying Kanji meanings, without addressing pronunciation or writing. The intervention period was limited to 20 minutes and covered only 40 Kanji, and long-term retention was not assessed. Future research could examine the method's effectiveness over longer periods, with more extensive Kanji sets, and explore its impact on other aspects of Kanji proficiency, including writing and pronunciation, as well as its long-term effects.

Acknowledgements

This work is partially supported by the SBI Foundation Hub for Data Science and Analytics (Grant ID SBIF002-007) and by the TIH Foundation for IoT and IoE (Grant ID RD/0123-TIH0018-003). The opinions, findings, conclusions, and recommendations presented in this paper are solely those of the author(s) and do not necessarily represent the views of the SBI Foundation, the TIH Foundation, or the Centre for Educational Technology, IIT Bombay. The study was approved by the IIT Bombay IRB (IRB approval number: IITB-IRB/2021/006).

References

Aneros, Noviyanti, Restu Pangestika, and Melia D. Judiasri. "The Use of Naritachi to Memorize Kanji." 3rd International Conference on Language, Literature, Culture, and Education (ICOLLITE 2019). Atlantis Press, 2020.

Carney, R. N., & Levin, J. R. (2002). Pictorial illustrations still improve students' learning from text. Educational psychology review, 14, 5-26.

Erlan, U. (2024). The effectiveness of mnemonics in foreign language learning among teenagers. Холодная наука, (9), 95-103.

Garg, A., & Rajendran, R. (2024). The Impact of Structured Prompt-Driven Generative AI on Learning Data Analysis in Engineering Students. In *CSEDU (2)* (pp. 270-277).

Garg, A., Soodhani, K. N., & Rajendran, R. (2025). Enhancing data analysis and programming skills through structured prompt training: The impact of generative AI in engineering education. *Computers and Education: Artificial Intelligence*, *8*, 100380.
Han, A., & Cai, Z. (2023, June). Design implications of generative AI systems for visual storytelling for young learners. In Proceedings of the 22nd annual ACM interaction design and children conference (pp. 470-474).
Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139-183). North-Holland.
Hasanah, K. N., Febrianti, D. N., Syah, T. C. Z., Simanjuntak, D. P. T., & Yusuf, M. (2024). Mastery of Kanji character meanings through visual image methods among Japanese language learners in North Sumatra. Humanities & Language: International Journal of Linguistics, Humanities, and Education, 1(4), 264-272.
Hiraga, M. K. (2006). Kanji: The visual metaphor. Style, 40(1-2), 133-147.
Japan Foundation. (2021). Survey report on Japanese-language education abroad: Present condition of overseas Japanese-language education. Japan Foundation.
Kontra, C., Goldin-Meadow, S., & Beilock, S. L. (2012). Embodied learning across the life span. Topics in Cognitive Science, 4(4), 731-739.
Levin, J. R. (1983). Pictorial strategies for school learning: Practical illustrations. In Cognitive strategy research: Educational applications (pp. 213-237). New York, NY: Springer New York.
Macedonia, M. (2014). Bringing back the body into the mind: Gestures enhance word learning in foreign language. Frontiers in Psychology, 5, 1467
Matsunaga, S. (2016). Teaching and Learning to Read Kanji as L2: Why Are They Difficult?. Indigenous Culture, Education and Globalization: Critical Perspectives from Asia, 245-262.
Mayer, R. E. (2024). The past, present, and future of the cognitive theory of multimedia learning. Educational Psychology Review, 36(1), 8.
Mayring, P. (2015). Qualitative content analysis: Theoretical background and procedures. In A. Bikner-Ahsbahs, C. Knipping, & N. Presmeg (Eds.), Approaches to qualitative research in mathematics education: Examples of methodology and methods (pp. 365-380). Springer. https://doi.org/10.1007/978-94-017-9181-6_13
Mori, Y., & Shimizu, H. (2007). Japanese language students' attitudes toward Kanji and their perceptions on Kanji learning strategies. Foreign Language Annals, 40(3), 472-490.
Paivio, A. (1986). Psychological processes in the comprehension of metaphor. Metophor And Thought, 163.
Patel, V. (2023). Investigating How Visuals Influence the Brain. Available at SSRN 5008217.
Rasiban, L. M. (2016, November). Understanding the basic of Kanji's meaning through mnemonic method learning. In Ninth International Conference on Applied Linguistics (CONAPLIN 9) (pp. 17-21). Atlantis Press
Rose, H. (2013). L2 learners' attitudes toward, and use of, mnemonic strategies when learning Japanese Kanji. The Modern Language Journal, 97(4), 981-992.
Rowley, M. (1992). Kanji pict-o-graphix: Over 1,000 Japanese Kanji and kana mnemonics. Stone Bridge Press.
Seely, C., & Henshall, K. G. (2016). Complete guide to Japanese Kanji: Remembering and understanding the 2,136 standard Japanese characters. Tuttle Publishing.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin.
Shekarabi, Z., & Tajfirooz, F. (2023). Challenges in learning Kanji: The perspective of Persian-speaking learners of Japanese. OnCUE Journal, 15(1), 24-49.
Shimizu, H., & Green, K. E. (2002). Japanese language educators' strategies for and attitudes toward teaching Kanji. The Modern Language Journal, 86(2), 227-241.
Sonjaya, I., Marcheta, N., & Pratama, M. R. (2024, February). Application of Learning Material with a Gamification Model Approach as an Innovation in Learning Japanese Kanji Letters. In International Conference on Teaching, Learning and Technology (ICTLT 2023) (pp. 243-249). Atlantis Press.
Sutedi, D., & Juangsih, J. (2024). Utilizing Cognitive Illustration as a Kanji Memorization Strategy in Kanji Learning. JAPANEDU: Jurnal Pendidikan dan Pengajaran Bahasa Jepang, 9(1), 63-75.
Sweller, J. (2023). Cognitive load theory: What we learn and how we learn. In Learning, design, and technology: An international compendium of theory, research, practice, and policy (pp. 137-152). Cham: Springer International Publishing.
Takeda-Kolb, N., & Ohsawa, H. (2024). Enhancing engagement and autonomy in advanced Japanese learning as the second language through AI-assisted video production. International Journal of Studies in Education and Science, 5(4), 432-456.

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

EmotiCon: Enhancing Public Speaking Training with Context-Aware Feedback on Emotional Delivery

Yash DESAI [a*1], Praneeth KASIRAJU[a+1], Syaamantak DAS[a], Ramkumar RAJENDRAN[a]
[a]*Centre for Educational Technology, IIT Bombay, India*
[*]desaiyash7272@gmail.com, [+]praneethkasiraju@gmail.com

Abstract: Effective public speaking relies on aligning emotional delivery with semantic content. While existing digital coaches analyze paralinguistic cues like tone and pitch, they often neglect the context of what is being said. This paper introduces EmotiCon, a novel system that offers context-aware emotional feedback by analyzing a speaker's audio, transcribed text, and preparation materials. EmotiCon segments speech, detects emotion across modalities, and compares it with relevant contextual passages to assess emotional congruence. It then generates targeted recommendations using a Large Language Model. A user study showed that context-aware feedback was better received than simple feedback. Quantitative evaluation using the extended Technology Acceptance Model (TAM) followed by semi-structured interviews revealed strong scores**.** Our findings suggest that integrating semantic context into feedback generation makes digital speech coaching tools more actionable and effective for learners.

Keywords: Public Speaking, Digital Coach, Emotion Recognition, Semantic Context, Large Language Models, Multimodal Analysis, Automatic Visualisation Feedback, Emotional Speech Training, System Evaluation

1. Introduction

In Technology-Enhanced Language Learning (TELL), developing oral proficiency, particularly public speaking, is a key objective. Public speaking, simply defined as "speaking in front of a community or group of people" (Bilgin, 2022), is central to academic, professional, and social communication. A foundational element of mastering this skill is timely, actionable feedback (Butler & Winne, 1995), which helps learners identify weaknesses and track progress.

Context plays a crucial role in shaping such feedback. It can be defined as the background, topic and setting that clarifies the meaning of the utterance (Rigotti & Rocci, 2006). Effective delivery also requires alignment between tone and the emotional weight or intention of the message (Guyer et al., 2021). In this work, we define *delivery* as the extent to which a speaker's expressed emotion aligns with the semantic content and context of their speech.

Beyond technical accuracy, speakers must convey passion, empathy, or urgency, depending on their message as delivery is influenced by the ability to form an emotional bond (Entong et al., 2024). Prior research on affect analysis of teachers' speech found that monotonous delivery is not well received (T S & Rajendran, 2022). Emotionally resonant speech has historically inspired movements and shaped discourse (Charteris-Black, 2011), underscoring that emotional expression is not merely stylistic but pedagogical. Yet most AI tools flatten emotional nuance by separating vocal expression from semantic context when generating feedback.

[1] *Authors Yash Desai and Praneeth Kasiraju made equal contributions in this paper.*

Recent advances in AI have led to the emergence of improved speech training tools. SpeakAR (Jim et al., 2025) creates immersive audience environments to simulate real-world pressure and manage anxiety. VoiceCoach (Wang et al., 2020) trains modulation using patterns learned from over 2,600 TED Talks. BETTER (Wynn & Wang, 2023) extracts audio tonality and speech emotions, and visualizes their alignment for users to view and infer. However, none fully integrate semantic context into feedback loops for improving emotional delivery.

Tools like **Yoodli**[2] and **Poised**[3] leverage LLMs to assess user delivery in real time while offering suggestions on clarity and confidence. They also factor in the conversational environment or interview setting making feedback more natural. Additionally, recent research has started to explore contextual adaptation in speech training. For example, (Zhang et al., 2025) proposed incorporating environmental context to detect speaker anxiety and offer adaptive recommendations. However, both approaches treat *setting* as context, not the semantic background of the speech. Factoring in the actual content or source material remains underexplored. This gap motivates our Research Question (RQ): **To what extent does integrating semantic context into feedback enhance learner confidence and emotional delivery in public speaking tasks?**

We present **EmotiCon**, a system that integrates semantic context into emotion-based public speaking feedback. EmotiCon accepts a speech audio file and the contextual materials the speaker used for preparation (e.g., essays, articles). It segments the speech using audio-based emotion analysis and transcribes each segment for text emotion analysis. For each segment, it retrieves the most relevant preparation material, analyzes its emotion, and compares alignment across audio, transcript, and context. A Large Language Model then generates actionable recommendations, visualized through an interactive dashboard.

We evaluated EmotiCon in a user study with eight higher-education students. Each submitted a short speech and preparation materials, then received context-aware feedback through the dashboard before taking part in semi-structured interviews. Participants found the feedback meaningful, helping them identify tone-content mismatches and boosting confidence for future presentations. Complementing the qualitative findings, participants completed a TAM questionnaire, which showed strong ratings for Perceived Usefulness (5.8/7), Attitude Toward Use (6.37/7), and a dedicated RQ dimension with questions to capture the research question, namely whether EmotiCon enhanced learners' confidence and delivery (6.37/7).

2. System Architecture

EmotiCon is designed to provide visual feedback of the user's varying emotions throughout the speech and give context-based recommendations for the same. EmotiCon implements a multimodal emotion analysis pipeline for audio data, which combines speech, text, and context modalities to produce granular visualisations of the segments.

As shown in Figure 1, EmotiCon first loads the user's speech audio and divides it into non-overlapping emotion segments. Each audio segment is processed using a pretrained "HuBERT"[4]-based transformer model ("superb/hubert-large-superb-er"[5]) for Speech Emotion Recognition (SER). The HuBERT model outputs a probability distribution over four emotion classes (happy, sad, angry, neutral) and selects the highest as the segment's dominant emotion with its prediction confidence[6]. Consecutive audio segments having the same

[2] Yoodli - AI powered speech coaching tool.
[3] Poised - Real-time communication feedback tool.
[4] HuBERT - Self-supervised speech representation model.
[5] Superb-Hubert - Benchmark suite version of Hubert for speech tasks.
[6] Confidence scores of the emotions detected by the model are referred to here as prediction confidence

dominant emotion are merged together. Following this, each audio segment is transcribed into text using Assembly AI's[7] external Automatic Speech Recognition (ASR) service. The transcribed text for each audio segment is analyzed using a transformer-based model, specifically the "j-hartmann/emotion-english-distilroberta-base"[8] (a fine-tuned DistilRoBERTa[9]) for emotion classification. The model outputs a probability distribution over emotion classes for each segment and records the highest-probability class as the segment's predicted emotion along with its prediction confidence. Labels from different sources are normalized to canonical forms.

The system extracts text from the context PDF and divides it into chunks. Each chunk is encoded into a dense vector using a pretrained sentence-transformer model. The embedding of each audio transcript is compared to these chunk embeddings, and the most relevant chunk is retrieved as contextual background. Each retrieved context chunk is analyzed with a text emotion classifier, and its dominant emotion and prediction confidence are recorded. Recommendations are generated by comparing the emotional alignment of the audio with the transcribed text and the context, and phrased by Gemini.

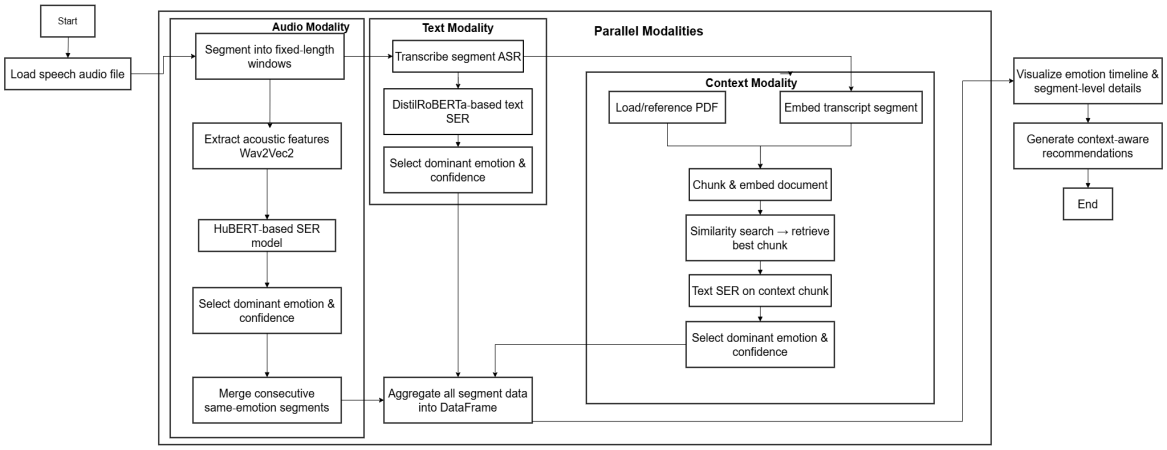


Figure 1: Emoticon's analysis and feedback generation pipeline integrating audio, text, and context modalities and generating recommendations.

3. Visual Feedback Dashboard

Figure 2 shows an interactive dashboard displaying each speech segment horizontally with the emotion concordance of the three sources (audio, text, context) revealing per-segment emotion details on hover. The color coding used for the segments in Figure 2 is:

- 🟩 **Green:** All three emotion sources match.
- 🟨 **Yellow:** Any two match.
- 🟥 **Red:** All differ.

A sidebar lists the segments by timestamp to navigate the speech. Selecting a segment updates the main panel to show a detailed analysis for that time frame, including the detected audio emotion, the transcribed text, the relevant context chunk, and the emotion scores for each modality.

[7] Assembly AI - Speech-to-text and audio intelligence API.
[8] Hartmann's DistilRoBERTa - Fine-tuned variant of DistilRoberta for NLP tasks.
[9] DistilRoBERTa - Fine-tuned variant of DistilRoberta for NLP tasks.

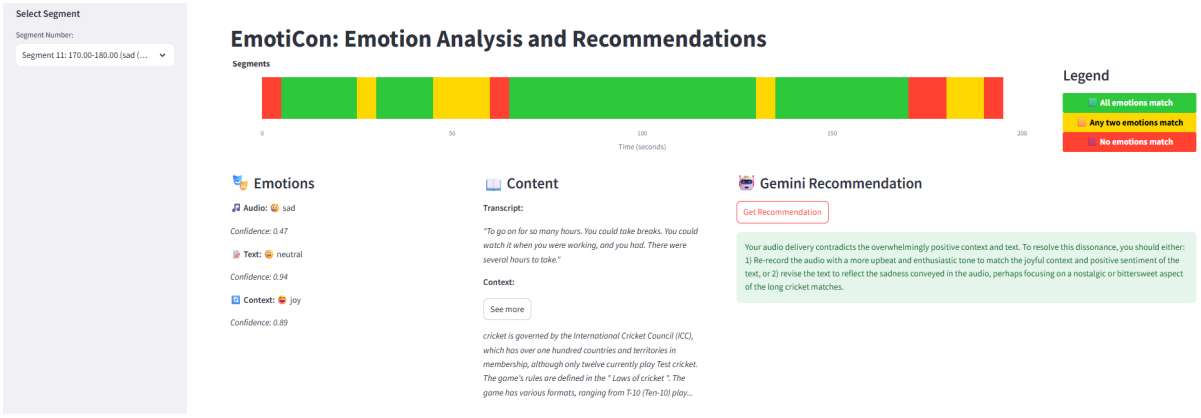


Figure 2: EmotiCon's feedback dashboard with emotion alignment visualization and recommendations.

Furthermore, the user can generate a specific recommendation for the selected segment. The system passes the segment's emotions, context, transcribed text, and neighboring transcripts into Google's Gemini 1.5 (temperature set to 0.2 to reduce hallucinations) to produce a response (Woollaston et al., 2024).

4. Methodology of Study Design



Figure 3: Study design flowchart.

A pilot study was conducted with eight higher-education students (7 male, 1 female) aged 18-35, from diverse academic and linguistic backgrounds. All had prior experience with academic presentations in English, which was chosen as the delivery language to ensure comparability. Participants selected one topic from five context-demanding options, chosen to require critical thinking, synthesis of sources, and avoidance of personal anecdotes. They were given 30 minutes to prepare a speech and were required to document all referenced textual material (URLs, sources, copied text). Audio or video sources were not permitted to maintain parity in semantic context processing, and participants were instructed to avoid unverifiable statements. Each participant then delivered a 3-6 minute speech before an audience, recorded in audio and video. The collected context material was compiled into a PDF and, along with the audio recording, served as input to EmotiCon. After this, participants explored their own speech analysis through the interactive dashboard, followed by a semi-structured interview to gather feedback on the context-based recommendations.

5. Results and Discussions:

RQ: To what extent does integrating semantic context into feedback enhance learner's confidence and emotional delivery in public speaking tasks?

To answer our RQ, we first conducted a qualitative analysis of the semi-structured interviews. Participant responses were reviewed and summarized, with illustrative quotes highlighting recurring insights or specific comments. The interviews showed that context-based feedback increased participants' awareness of their speaking style and conveyed emotions. One participant noted it enabled them to "set the narrative to pursue

their audience". The "clean, temporal segmentation of the speech" was described as informative, and using context as a modality was considered "novel for digital coaches". Several participants said the segment-level emotion feedback and alignment visualization made them "more confident, as they did not think about it before". Recommendations were perceived as engaging and insightful, aligning with prior work on the role of context in emotional delivery. Participants also noted that the UI "displays too much at once" and suggested condensing recommendations into shorter points. Another suggestion was an "audio modulator" feature to play back segments aligned with the intended emotional tone.

To quantitatively evaluate the system, participants completed a Technology Acceptance Model - TAM (Granić & Marangunić, 2019) based questionnaire rating 17 items on a 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree). We assessed 6 parameters in the questionnaire, namely - Perceived Usefulness (PU): The specificity and actionable quality of the context-aware recommendations. Perceived Ease of Use (PEOU): The clarity and usability of the dashboard. Perceived Enjoyment (PE): Engagement with interactive feedback. Attitude Toward Use (ATT) & Intention to Use (INT): Overall disposition towards the AI coach and likelihood of future use. In addition to the standard TAM constructs, we included an "RQ" dimension aligned with our research question, asking participants whether EmotiCon gave participants actionable feedback to improve their confidence and delivery.

Table 1. Results of the mean and standard deviation of the 6 parameters of the TAM questionnaire and the RQ dimension.

	PU	**PEOU**	**PE**	**ATT**	**INT**	**RQ**
Mean	5.8	5.43	5.41	6.37	5.31	6.37
Standard Deviation	0.55	0.96	1.76	0.41	1.83	0.74

The TAM and RQ results from Table 1, reinforce the qualitative findings. Perceived Usefulness scored highly (M = 5.8), indicating that participants found the recommendations actionable, consistent with their remarks about identifying tone–content mismatches. Attitude toward Use was the highest-rated dimension (M = 6.37), reflecting strong positivity and aligning with feedback that the system felt "novel" and "engaging." The added RQ dimension, measuring perceived improvements in confidence and delivery, also received a high score (M = 6.37), supporting interview claims that participants felt more confident and aware of their speaking style. By contrast, Intention to Use showed more variation (M = 5.31, SD = 1.83), suggesting the dashboard being overwhelming. These results suggest that EmotiCon was well received, with participants finding the system's feedback both actionable and useful.

6. Conclusion

We introduced EmotiCon, a novel system designed to enhance public speaking training by providing context-aware feedback on emotional delivery. The results of the pilot study validate and reaffirm the importance of integrating semantic context in feedback with multimodal analysis (audio, text and source material), which generates specific and effective recommendations for a speech. Such systems could benefit non native English speakers to practice their delivery and help them improve their communication skills. EmotiCon could potentially contribute to treating social anxiety or stage fright by providing a safe space to learn and practice emotion and tone in speech delivery.

Despite the positive outcomes, there exist limitations to be acknowledged. First, the system in its current state has no method to detect intentionally conflicting emotions, such as sarcasm, which makes it ineffective for satirical or comedic speeches. Additionally, our study only included participants over 18, so the impact on younger audiences remains unexplored.

In future work, we would like to address the shortcomings of the system and focus on simplifying the user dashboard for easier interpretation of scores and reducing the volume of data on-screen for a smoother experience. Additionally, detecting and supporting ironic styles or figures of speech and conducting broader studies across age groups and use cases is also planned. Addressing factors such as speech impairments and language nativity while expanding to facial recognition will help enhance the recommendation system making it more inclusive, reliable, and personalized.

7. Acknowledgment

The authors would like to thank all the participants and anonymous reviewers of this article. This work is supported by the SBI Foundation Hub for Data Science and Analytics under grant ID SBIF002-007.

References

Bilgin, R. (2022). A review of public speaking and its components. *Canadian Journal of Educational and Social Studies, 2*(3), 37–49. https://doi.org/10.53103/cjess.v2i3.39

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245–281. https://doi.org/10.3102/00346543065003245

Rigotti, E., & Rocci, A. (2006). Towards a definition of context. *Studies in Communication Sciences, 6/2,* 155-180. https://doi.org/10.5169/SEALS-791116

Guyer, J. J., Briñol, P., Vaughan-Johnston, T. I., Fabrigar, L. R., Moreno, L., & Petty, R. E. (2021). Paralinguistic features communicated through voice can affect appraisals of confidence and evaluative judgments. *Journal of Nonverbal Behavior*, *45*(4), 479–504. https://doi.org/10.1007/s10919-021-00374-2

Entong, M. B. M., Garil, B. A., Muarip, V. C., & Chavez, J. V. (2024). Language delivery styles in academic trainings: Analysis of speaker's emotional connection to audience for lasting learning. *Forum for Linguistic Studies*, *6*(3), 326–342. https://doi.org/10.30564/fls.v6i3.6533

T S, A., & Rajendran, R. (2022). Audio feature based monotone detection and affect analysis for teachers. *2022 IEEE Region 10 Symposium (TENSYMP)* (pp. 1–6). IEEE. https://doi.org/10.1109/TENSYMP54529.2022.9864393

Charteris-Black, J. (2011). *Politicians and rhetoric: The persuasive power of metaphor*. Springer.

Jim, M. E., Yap, J. B., Laolao, G. C., Lim, A. Z., & Deja, J. A. (2025). *Speak with confidence: Designing an augmented reality training tool for public speaking.* arXiv. https://doi.org/10.48550/arXiv.2504.11380

Wang, X., Zeng, H., Wang, Y., Wu, A., Sun, Z., Ma, X., & Qu, H. (2020). VoiceCoach: Interactive evidence-based training for voice modulation skills in public speaking. *In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–12). ACM. https://doi.org/10.1145/3313831.3376726

Wynn, A., & Wang, J. (2023). BETTER: An Automatic feedBack systEm for supporTing emoTional spEech tRaining. *Lecture Notes in Computer Science,* 746–752. https://doi.org/10.1007/978-3-031-36272-9_66

Zhang, T., Tan, J., Xiong, Z., Wu, B., & Zheng, C. (2025). Causality-guided context-aware multimodal public speaking anxiety detection for out-of-distribution generalization. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICASSP49660.2025.10888426

Woollaston, S., Flanagan, B., Ocheja, P., Dai, Y., & Ogata, H. (2024). TAMMY: Supporting EFL translation practice with an LLM-powered chatbot. In *Proceedings of the 32nd International Conference on Computers in Education (ICCE 2024)*. APSCE. https://doi.org/10.58459/icce.2024.4901

Granić, A., & Marangunić, N. (2019). Technology acceptance model in educational context: A systematic literature review. *British Journal of Educational Technology, 50*(5), 2572–2593. https://doi.org/10.1111/bjet.12864

C7
Practice–driven Research, Teacher Professional Development and Policy of ICT in Education (PTP)

Developing Teachers' Digital Competence through Expansive Learning: A Study of Teacher–Researcher Collaborative Teaching Research

Junyang LIU[a*], Bian WU[b*] & Dong LI[c]
[a]*East China Normal University, China*
[b]*East China Normal University, China*
[c]*The School Attached to Shanghai Caoyang No.2 High School, China*

Abstract: As education rapidly digitalizes and technologies like generative AI continue to evolve, strengthening teachers' digital competence has become more crucial than ever. However, many existing training programs are highly structured and pay limited attention to the diverse needs teachers encounter in their own classrooms. This study explores how expansive learning supports the development of teachers' digital competence (TDC) and transformative agency. Based on a case of collaborative curriculum design between teachers and researchers, the research analyzes how teachers identify and resolve systemic tensions in practice. Using the DigCompEdu framework and agency analysis, the findings reveal a four-phase developmental trajectory-Questioning and Analysis, Modeling and Testing, Implementation and Reflection, and Consolidation and Standardization-through which teachers evolve from passive users of digital tools to active pedagogical innovators. The study highlights the importance of collaborative, reflective, and practice-driven professional learning in advancing digital competence.

Keywords: Teacher Professional Development, Digital Competence, Expansive Learning, Collaborative Design

1. Introduction

With the rapid advancement of emerging technologies such as generative AI, teaching and learning are undergoing profound change (Wohlfart & Wagner, 2023). Strengthening teachers' digital competence has thus become urgent. In many national and regional frameworks, the role of teachers is shifting from simply using tools to transforming pedagogy through technology (Redecker, 2017). Yet, despite relevant policies, the ability to effectively integrate digital tools is still not widely recognized as core professional competence (Instefjord & Munthe, 2016). Existing training programs often follow fixed pathways and overlook individual and disciplinary needs (Gisbert Cervera & Lázaro Cantabrana, 2015; Papanikolaou et al., 2017; ElSayary, 2023). Addressing this gap, this study uses expansive learning theory to examine how teachers, in collaboration with researchers, identify systemic tensions, co-construct solutions, and enact classroom transformation, and how such processes activate transformative agency and foster digital competence.

2. Literature Review

2.1 Teachers Digital Competence

Digital competence is widely recognized as a core 21st-century skill for citizens in the digital era. In the field of teacher education, a range of conceptual models has been proposed to articulate its complexity. The widely adopted TPACK model highlights the multifaceted nature of digital integration by emphasizing the interrelationship between technological, pedagogical,

and content knowledge (Mishra & Koehler, 2006). Complementing this, the PEAT model maps the developmental trajectory of digital literacy along two dimensions, offering a dynamic view of competence growth (McDonagh et al., 2021). As digital technologies continue to evolve, the scope of teachers' digital competence has expanded beyond basic operational skills to encompass ethical awareness, digital safety, and creative thinking. This shift reflects a broader understanding of digital competence as a composite capacity to integrate technologies effectively in multi learning environments and to enhance instructional quality (Falloon, 2020).

Viewing digital competence solely as an individual attribute risks overlooking the broader organizational and contextual dynamics that shape its development. Rather than isolating digital competence at the level of single actors, it is more productively understood as a collective and institutional responsibility, influenced by structural and cultural conditions within the school system (Pettersson, 2018). This perspective has prompted a rethinking of professional development models. Increasingly, scholars advocate for collaborative and co-constructed approaches, where educators and researchers work together to define problems, co-design strategies, and engage in reflective inquiry. Such models facilitate knowledge co-creation and support context-sensitive pedagogical innovation (Shun et al., 2021).

Because digital competence is context-specific, its development requires not only technical skills but also the ability to align digital tools with pedagogical goals. This shifts teachers from passive implementers to informed decision-makers navigating complex contexts, embodying transformative agency—the capacity to break from established practices and initiate change (Brevik et al., 2019). Activating such agency is key to advancing digital competence and fostering sustainable professional identity growth.

2.2 Expansive Learning in Teacher Professional Development

Expansive Learning Theory, rooted in Cultural-Historical Activity Theory (CHAT), provides a framework for understanding how teachers identify and resolve systemic contradictions to achieve professional transformation (Engeström, 2001; Engeström & Sannino, 2010). It shifts the focus from learning as mere internalization to a process where individuals or collectives create new practices through concrete activities. Engeström's (2001) seven-stage cycle illustrates how such practices are transformed under the drive of contradictions.

Figure 1 includes the following iterative stages: Questioning current practices; Analyzing the root causes of contradictions; Constructing a new model; Examining the new model; Implementing the new practice; Reflecting on and evaluating the process; Consolidating and disseminating new practices. The process is not linear but recursive, emphasizing the generative and sustained nature of practice transformation.

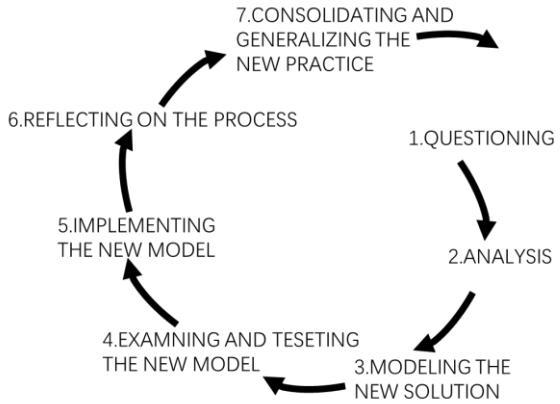


Figure 1. Expansive learning cycle

This theory has been widely applied in fields such as education and organizational change, particularly in contexts like vocational training, healthcare education, and teacher development, to understand how collectives reconstruct existing practices in response to changing external conditions (Engeström, 2010). In teacher education, expansive learning is

seen as a crucial pathway for enhancing teachers' professional agency and fostering collective transformation. Existing studies primarily focus on teachers' collaborative construction of practices in curriculum reform, co-designed pedagogy, and technology-supported teaching (Sannino, 2016; Pareto et al., 2022; Augustsson et al., 2021). Although expansive learning has been widely applied in teacher education research, its role in supporting the development of teachers' digital competence remains underexplored and warrants further investigation.

This study adopts expansive learning theory as its analytical lens to reveal how teachers navigate the transformation from digital awareness to pedagogical innovation within cross-boundary collaborations, complex tasks, and practice-driven communities. The research focuses on the following two questions:

RQ1: How does teachers' digital competence change during expansive learning, and what drives this transformation?

RQ2: How does teachers' transformative agency emerge and develop through collaboration around digital competence?

3. Research Design

3.1 The systemic context and participants

This study was conducted by a collaborative team composed of members from School C and University H (Figure 2). The participants included one vice principal from School C (C1), two teachers (C2, a primary science teacher, and C3, a STEM teacher teaching both primary and secondary school), one professor from University H (H1), and two graduate students (H2 and H3). Each member contributed based on their area of expertise, forming a cross-institutional co-design team. The object of this collaborative activity was a STEM curriculum project titled "Designing a Digitally Micro-Weather Station."

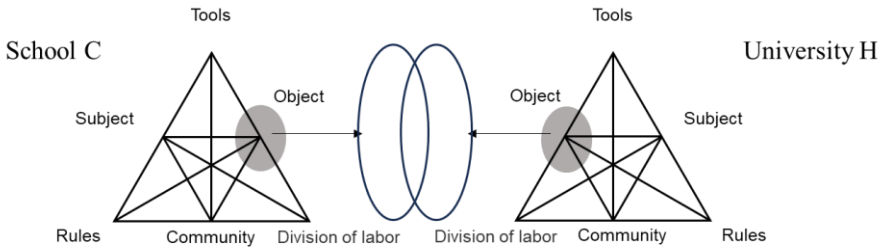


Figure 2. The activity system for collaborative team

The project fostered students' interdisciplinary thinking and problem-solving by integrating meteorology, engineering education, and ICT. Resources included lesson plans, slides, videos, simulations, open-source hardware, and interactive whiteboards. School C and University H formed a collaborative community, guided by curriculum standards, objectives, and assessment criteria. Roles were clearly divided: graduate students (H2, H3) drafted plans and developed content; teachers (C2, C3) taught; the vice principal (C1) coordinated; and the professor (H1) offered guidance. Pre- and post-lesson research discussions refined content and strategies.

3.2 Data Collection

The collaborative course was implemented from March to June 2024, with six lessons in total. These were co-designed and iteratively refined by the team. To comprehensively capture the dynamic changes during this collaboration, two main types of data were collected including audio and video recordings of teaching discussions among teachers and researchers before and after each lesson; as well as documents generated during curriculum design, including lesson plans, instructional slides, draft scripts.

3.3 Data Analysis

Guided by the theoretical framework of expansive learning, the researchers independently reviewed the full set of transcripts, observation notes, and design documents to trace the developmental trajectory of teachers' digital competence and the emergence of their transformative agency throughout the collaborative process. The aim was to identify key patterns of change and turning points in practice that reflected shifts in professional understanding and pedagogical innovation.

Multiple team discussions resolved coding disagreements. Using Engeström's (2001) seven-stage expansive learning cycle as a guide, observed actions were mapped to theory, then iteratively coded from transcripts, notes, and design documents. Overlapping actions were merged into four empirically grounded phases: (a) questioning and analysis—identifying systemic contradictions and reflecting on limits of current digital teaching; (b) modeling and testing—co-designing and trialing digitally integrated instructional approaches; (c) implementation and reflection—applying and refining solutions in classrooms; and (d) consolidation and standardization—stabilizing, sharing, and internalizing effective practices. This theory-informed yet empirically refined framework captures both the logic of expansive learning and the observed developmental trajectory, structuring the findings to show teachers' digital competence growth and transformative agency in collaboration.

To examine the development of teachers' digital competence and the emergence of their transformative agency during the collaborative curriculum design and implementation process, this study adopted a dual analytical framework. The European Framework for the Digital Competence of Educators (DigCompEdu) (Redecker, 2017) was used to analyze the trajectory of digital competence development. DigCompEdu outlines six core areas: professional engagement, digital resources, teaching and learning, assessment, empowering learners, and facilitating learners' digital competence. Each area includes specific competences that describe how educators use digital technologies to enhance teaching and learning. In this study, these areas were used to code and interpret changes in teachers' awareness, knowledge, and skills observed in teaching discussions, document writing, and classroom practice.

To analyze transformative agency, the study adopted the coding framework developed by Virkkunen (2006) and Haapasaari et al. (2016), which defines it as the capacity of individuals or groups to break away from established practices and proactively reconstruct their activities. Teachers' reflective dialogue was coded entirely according to their six expressions of agency: (a) resisting, (b) criticizing, (c) explicating, (d) envisioning, (e) committing to actions, and (f) taking actions. This framework was applied directly to capture cognitive shifts and emerging agency in response to systemic tensions.

By combining the DigCompEdu framework with the lens of transformative agency, this study mapped teachers' professional growth across multiple dimensions. This dual approach revealed how reflective critique, collective negotiation, and practical experimentation contributed to both the development of digital competence and the activation of transformative agency during the expansive learning process.

By incorporating the Expansive Learning Cycle, the study further analyzed which stage of the cycle the teachers were situated in throughout the process. This helped to reveal the interactive relationship between the development of digital competence and the emergence of transformative agency. Through this triangulated analytical approach, the study aimed to comprehensively uncover the mechanisms by which teachers move from "using technology" to "reconstructing pedagogy" in participatory practice.

4. Results

This section presents the research findings based on the four phases. Each phase is analyzed from two dimensions: (a) the developmental trajectory of teachers' digital competence (based on the DigCompEdu framework), and (b) the generation and evolution of teachers' transformative agency. This structure reveals the dynamic interplay and trajectory between

teachers' professional growth and the enhancement of their agency throughout the co-construction of the curriculum.

4.1 Questioning and Analysis Phase

Digital Competence Development

Early in curriculum preparation, teachers struggled to implement a STEM project unit using Arduino boards and sensors. Despite prior experience with multimedia tools, they were unprepared for tasks involving hardware operations and programming, revealing gaps in digital competence, especially in "Teaching and Learning" and "Digital Resources."

The lack of confidence was evident as teachers struggled to design lessons that combined subject knowledge, digital tools, and hands-on creation. As C3 noted, *"This type of class is completely new to me. I have never had experience teaching a course that integrates hardware components and digital tools like this before."* Teachers also showed hesitancy in managing technological complexity in the classroom. C2 commented, *"I think it's very difficult to ask students to both build and design the device in class. It's too much for them to process at once."* These statements reflected the perceived gap between what the curriculum expected and what the teachers felt equipped to deliver.

In addition, during interdisciplinary collaboration and co-planning sessions, teachers found it challenging to engage in professional discourse that relied heavily on technical vocabulary. This hindered their active involvement and weakened their sense of agency. C1 explained, *"When the engineers talk about components and logic, I often don't know how to respond, so I just stay quiet."* The inability to participate fully in these discussions revealed a lack of readiness not just in technical skills but in collaborative digital engagement.

At this stage, the most prominent feature of teachers' digital competence development was not technical mastery but rather the emergence of awareness — an acknowledgment of the challenges ahead and an increased sensitivity to what digital teaching truly demands. This awareness became the starting point for future learning and adaptation through sustained practice.

Transformative Agency —— Resisting and Criticizing

In response to the challenges of project-based STEM teaching, teachers began to express an emerging form of transformative agency. Rather than blindly executing instructional plans, they started to question and evaluate the pedagogical logic and practical feasibility of the tasks assigned to them. This marked an important shift: from passive recipients of externally designed curricula to active participants in reshaping them.

One key expression of this agency was resistance — teachers openly voiced doubts about the ambitious nature of the project tasks. C2 remarked, *"It's very difficult to ask students to both design and build the device within class time."* Similarly, C1 reflected, *"We don't even have enough time to prepare properly ourselves. Expecting students to manage everything in one class is too much."* These remarks reflect concern not with rejecting innovation, but with ensuring that instructional design is responsive to classroom realities. The resistance was rooted in pedagogical care: teachers were advocating for more realistic goals that matched their students' cognitive and time capacities.

Beyond task design, teachers also began to critique their roles within the interdisciplinary collaboration process. They noticed a structural imbalance in which technical voices often dominated, while pedagogical insights were sidelined. As C3 observed, *"Sometimes I just listen because I'm not sure whether my ideas align with what the researchers are saying."* C1 similarly noted, *"It feels like I'm expected to follow the structure that's already decided. There's not much space to bring in my own perspectives."* These comments reveal not only communication difficulties but also a sense of marginalization within the collaborative framework.

By both questioning unrealistic goals and expressing concerns about unequal collaboration, teachers showed a deeper kind of professional initiative. Teachers were no longer confined to the role of executors but began positioning themselves as evaluators and reformers of both the instructional content and the institutional process. While still tentative,

this marked a meaningful step toward empowered professional identity in digitally mediated, interdisciplinary teaching environments.

4.2 Modeling and Testing Phase

Digital Competence Development

In the modeling and testing phase, teachers showed marked improvement in digital competence, taking greater initiative in shaping instructional strategies. Growth was evident in "Teaching and Learning," "Digital Resources," and "Facilitating Learners' Digital Competence."

One of the key shifts was their growing ability to anticipate students' learning difficulties and respond through supportive instructional design. C3, for example, proposed: *"We could create small instruction cards for each device, including pin details and wiring guides."* This suggestion reflected a shift from problem-solving in the moment to proactive planning. Rather than seeing technical complexity as a disruption, teachers began integrating it as a component of structured learning. The use of customized aids, simplified wiring diagrams, and visual flowcharts indicated their increasing mastery in aligning technical content with pedagogical scaffolding.

Moreover, teachers became more strategic in adapting the sequencing of classroom activities to support student engagement and pacing. As they better understood both the affordances of digital tools and the cognitive demands placed on learners, they adjusted their instructional flow accordingly. This reflected a deeper competence in harmonizing technology with teaching rhythms and demonstrated their growing confidence in orchestrating complex classroom environments.

Overall, digital competence at this phase was no longer merely about technical handling; it evolved into a student-centered, anticipatory planning approach where digital tools were embedded meaningfully within pedagogical structures.

Transformative Agency —— Envisioning

In contrast to the earlier phases dominated by doubt and resistance, this stage was marked by the emergence of *envisioning*—a form of transformative agency in which teachers began to imagine and design improved instructional pathways. This capacity was not simply about reacting to problems, but about constructing alternatives based on classroom experience and pedagogical reflection.

Teachers demonstrated this agency through concrete proposals for redesign. In one planning session, both C2 and C3 suggested reordering the lesson flow: *"I think we can change the order of the activities to better match students' response rhythms."* This proposal signaled more than a surface-level rearrangement; it reflected a shift in how teachers conceptualized instruction—as a flexible, adaptive process shaped by students' needs rather than a rigid plan to be followed.

Such envisioning also revealed teachers' increasing ownership over the instructional process. They were no longer acting as implementers of an externally imposed curriculum but were becoming co-designers, capable of restructuring both content and delivery. Their confidence to intervene was supported by their expanding understanding of interdisciplinary collaboration. As they became more familiar with the technical side of the project, they were better able to articulate how design elements could support learning outcomes.

Furthermore, their ideas began to reflect a hybrid logic—combining learning theory, student psychology, and practical constraints of classroom time and materials. C1, for instance, discussed balancing task complexity with emotional engagement: *"Students feel excited when things light up or move, but if we do too much too fast, they get anxious. Maybe we need to slow down some parts."* This sensitivity to emotional pacing demonstrated not only pedagogical awareness but also a deepening of design-oriented thinking.

In sum, the envisioning observed at this stage indicated a growing pedagogical imagination and professional maturity. Teachers were not merely adapting—they were beginning to lead change by proactively shaping the structure and flow of instruction in digitally enriched environments.

4.3 Implementation and Reflection Phase

Digital Competence Development

During implementation, teachers brought the co-designed curriculum into classrooms and faced challenges such as student errors, hardware malfunctions, time pressure, and fluctuating engagement. These required them to apply digital competence more adaptively, particularly in "Teaching and Learning," "Assessment," and "Professional Engagement."

Faced with students frequently miswiring components, teachers learned to slow down and scaffold learning with more care. C2 shared, *"Students kept wiring things incorrectly, so I slowed down and started explaining and demonstrating step by step."* C1 similarly noted, *"After two or three groups failed to make their work, I realized I had to break the explanation into simpler parts."* These instructional adjustments reflected not only improved classroom management, but also a more effective use of digital tools in response to learner needs.

In addition, teachers became more engaged in post-lesson reflections, drawing upon data provided by researchers—such as student behavior logs and video excerpts—to analyze and revise their instructional strategies. This marked a shift toward more evidence-informed practice. They began asking: Where did confusion arise? How did pacing affect focus? What kinds of support worked best? Such questions signaled a growing comfort with professional inquiry and iterative improvement.

Finally, several teachers began to reflect on how their instructional design choices influenced student engagement. Some noticed that when they gave students more freedom, motivation increased—even if the results were imperfect. This recognition marked a subtle but significant evolution in digital competence: a blending of technical proficiency with pedagogical intuition, aimed at fostering deeper student involvement.

Transformative Agency —— Explicating

In this phase, teachers' transformative agency evolved from adaptation to explanation. Through reflective analysis, they began explicating—making visible the contradictions in their teaching and tracing the sources of misalignment between curriculum design, classroom flow, and student engagement. This shift toward explanation marked a higher-order form of agency rooted in reasoning, not just intuition.

Teachers became more articulate about why certain instructional structures were misfiring. C3 questioned, *"Can we let students design their own process for the operation part? Even if it gets messy, they'll be more engaged."* C2 supported this idea: *"When I tried to lead every step, students just waited passively. But when I gave them some room, they started trying things on their own."* These observations moved beyond surface behavior to pedagogical logic: teachers realized that too much control could suppress inquiry and engagement.

They also turned their attention inward, analyzing personal teaching habits. C1 admitted, *"I've realized I tend to over-structure everything. I plan for every minute, but these tasks need more room to breathe."* Teachers began to reflect on whether traditional classroom norms—tight timing, precision, control—were suitable for open-ended, hands-on STEM tasks. They no longer focused solely on whether lessons succeeded, but on the deeper reasons behind classroom tensions.

While full-scale redesigns were not yet common, the move toward explicating marked an essential cognitive shift. Teachers were engaging in structured reflection, questioning default patterns, and beginning to formulate alternatives based on student thinking, emotional pacing, and collaborative dynamics. It was a crucial step toward transformative professionalism.

4.4 Consolidation and Standardization Phase

Digital Competence Development

As the project neared completion, teachers' digital competence stabilized and extended beyond the classroom. Digital practice became part of institutional routines, with leaders converting their experiences into school-based resources, training materials, and shared

frameworks for peer learning—reflecting advanced competence in "Professional Engagement" and "Digital Resources."

C1 suggested, *"We could compile all of this into a training kit so other teachers can use it too."* C3 similarly proposed, *"What if we turn our lesson plans into a guidebook, with examples and student work samples?"* These comments indicated that teachers were no longer simply users of digital tools—they were becoming contributors to a professional knowledge base, with an eye toward scalability and sustainability.

Moreover, digital competence had become part of the school's internal professional culture. Teachers were now confident in designing digital tasks, troubleshooting in real-time, and adapting resources for varied instructional needs. This institutionalization of competence represented a critical transformation: digital skills were no longer personal or experimental—they were shared, repeatable, and capable of being transferred across classrooms and departments.

Transformative Agency —— Committing to actions

In this final phase, teachers' transformative agency matured into committing to actions—a proactive stance where teachers not only embraced digital innovation but also assumed responsibility for spreading and sustaining it. They moved from adapting to innovation to driving it, actively shaping peer learning, departmental planning, and school-wide strategies.

C2 expressed, *"Next semester, I want to lead a training for other science teachers—this design has really changed my view of tech-based instruction."* C1 echoed this commitment: *"I'm already drafting a template other teachers can use. It's a way to keep the work going after the project ends."* These statements illustrate a key evolution: teachers no longer positioned themselves as recipients of innovation but as professional actors capable of influencing broader systems.

This agency was not limited to individual enthusiasm. It manifested in organized efforts to share, standardize, and institutionalize practices. Teachers began meeting regularly to draft internal documents, revise workflows, and coordinate across subjects. They discussed how to embed successful approaches into school policy and even proposed rotating leadership structures to sustain innovation cycles. Through these actions, they demonstrated agency not just in response to their own classrooms, but in shaping collective capacity and institutional culture.

By this stage, teacher agency had extended from reflection and critique to systemic participation. The commitment to act showed that innovation had become part of their identity—not as a temporary project, but as a professional responsibility integrated into their long-term teaching practice.

5. Discussion and Conclusion

This study explored how teachers' digital competence was developed through a process of expansive learning that unfolded across four interconnected phases (Figure 3), i.e., Questioning and Analysis, Modeling and Testing, Implementation and Reflection, and Consolidation and Standardization. Within this process, teachers engaged in sustained professional learning by confronting and resolving contradictions within the activity system, such as contradictions and tensions between digital tools and instructional objectives or between institutional rules and classroom realities (Chen et al., 2024; Engeström, 2020).

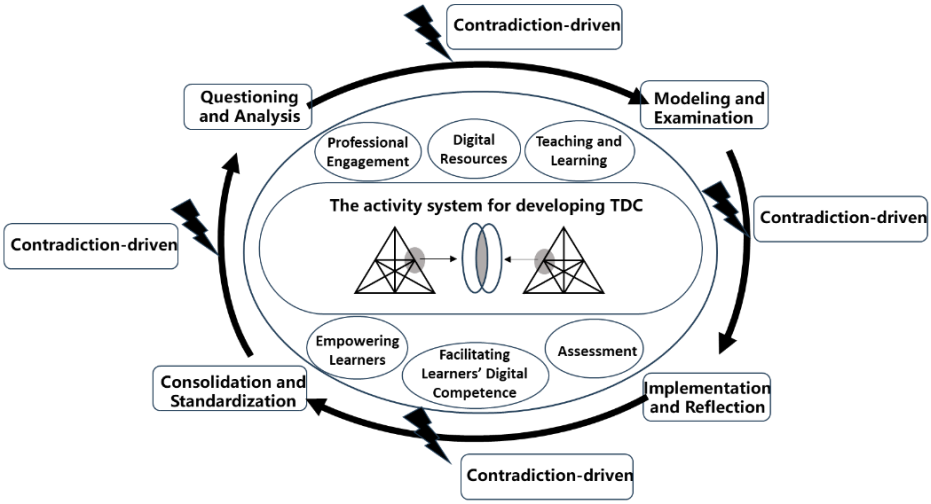


Figure 3. *A Model of Expansive Learning for Teachers' Digital Competence Development*

As they progressed through collaborative lesson planning, experimentation, and reflection, teachers came to see how digital tools could support pedagogy by aligning with student needs, enabling real-time feedback, and fostering learner agency. This shift marked a move from basic adaptation to deep pedagogical integration. Transformative agency developed alongside this change: teachers questioned ineffective routines, proposed alternatives, and increasingly committed to new practices through peer training, resource codification, and participation in organizational decision making (Wei & Sannino, 2024). This commitment signaled a shift from individual trial to institutional ownership and leadership. The process also underscored the collective, dialogical nature of teacher learning, as mixed communities of researchers and practitioners integrated local insights with theoretical guidance, creating a shared culture of digital innovation. This aligns with Bakhtin's (1982) concept of multivoicedness, where diverse dialogue fosters critical thinking and creative problem solving. Digital competence here emerged not as a fixed skill set, but as a capacity evolving through iterative cycles of identifying problems, designing and testing solutions, and refining practices based on student responses (Englund & Price, 2018). Theory and practice reinforced each other: theoretical insights informed action, while teaching experience refined conceptual understanding.

In sum, this study contributes to the theoretical articulation and empirical demonstration of how expansive learning fosters sustainable and context-sensitive development of teachers' digital competence. It shows how digital competence and transformative agency co-evolve, and why professional learning must be collaborative, practice-based, and theory-informed to respond to the challenges of digitalization.

Acknowledgements

This work was sponsored by Chinese National Social Science Fund (Educational Research), "Research on the Data-Enabled Teachers' Digital Literacy Training Model and Visual Environment Design" (Grant number BCA230274).

References

Augustsson, D. (2021). Expansive learning in a change laboratory intervention for teachers. *Journal of Educational Change, 22*(4), 475–499. https://doi.org/10.1007/s10833-020-09404-0

Bakhtin, M. M. (1982). *The dialogic imagination: Four essays by M. M.* Austin: University of Texas Press.

Brevik, L. M., Gudmundsdottir, G. B., Lund, A., & Strømme, T. A. (2019). Transformative agency in teacher education: Fostering professional digital competence. *Teaching and Teacher Education, 86*, 102875. https://doi.org/10.1016/j.tate.2019.07.005

Chen, K.-Z., Tseng, J.-Y., & Oakley, B. (2024). Transforming online teacher training through expansive learning: A case study applying cultural-historical activity theory and the Change Laboratory method. *The Asia-Pacific Education Researcher*. https://doi.org/10.1007/s40299-024-00954-y

ElSayary, A. (2023). The impact of a professional upskilling training programme on developing teachers' digital competence. *Journal of Computer Assisted Learning, 39*(4), 1154–1166. https://doi.org/10.1111/jcal.12788

Engeström, Y. (2001). Expansive learning at work: Toward an activity theoretical reconceptualization. *Journal of Education and Work, 14*(1), 133–156. https://doi.org/10.1080/13639080020028747

Engeström, Y., & Sannino, A. (2010). Studies of expansive learning: Foundations, findings and future challenges. *Educational Research Review, 5*(1), 1–24. https://doi.org/10.1016/j.edurev.2009.12.002

Englund, C., & Price, L. (2018). Facilitating agency: The change laboratory as an intervention for collaborative sustainable development in higher education. *International Journal for Academic Development, 23*(3), 192–205. https://doi.org/10.1080/1360144X.2018.1478837

Falloon, G. (2020). From digital literacy to digital competence: The teacher digital competency (TDC) framework. *Educational Technology Research and Development, 68*(5), 2449–2472. https://doi.org/10.1007/s11423-020-09767-4

Pettersson, F. (2018). On the issues of digital competence in educational contexts: A review of literature. Education and Information Technologies, *23*(3), 1005–1021. https://doi.org/10.1007/s10639-017-9649-3

Gisbert Cervera, M., & Lázaro Cantabrana, J. L. (2015). Professional development in teacher digital competence and improving school quality from the teachers' perspective: A case study. *Journal of New Approaches in Educational Research, 4*(2), 115–122. https://doi.org/10.7821/naer.2015.7.123

Haapasaari, A., Engeström, Y., & Kerosuo, H. (2016). The emergence of learners' transformative agency in a Change Laboratory intervention. *Journal of Education and Work, 29*(2), 232–262. https://doi.org/10.1080/13639080.2014.900168

Instefjord, E. J., & Munthe, E. (2017). Educating digitally competent teachers: A study of integration of professional digital competence in teacher education. *Teaching and Teacher Education, 6*7, 37–45. https://doi.org/10.1016/j.tate.2017.05.016

McDonagh, A., Camilleri, P., Engen, B. K., & McGarr, O. (2021). Introducing the PEAT model to frame professional digital competence in teacher education. *Nordic Journal of Comparative and International Education (NJCIE), 5*(4), Article 4. https://doi.org/10.7577/njcie.4226

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record, 108*(6), 1017–1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x

Papanikolaou, K., Makri, K., & Roussos, P. (2017). Learning design as a vehicle for developing TPACK in blended teacher training on technology enhanced learning. *International Journal of Educational Technology in Higher Education, 14*(1), Article 1. https://doi.org/10.1186/s41239-017-0072-z

Pareto, L., & Willermark, S. (2018). TPACK in situ: A design-based approach supporting professional development in practice. *Journal of Educational Computing Research*. https://doi.org/10.1177/0735633118783180

Redecker, C. (2017). *European Framework for the Digital Competence of Educators: DigCompEdu*. https://doi.org/10.2760/159770

Sannino, A., Engeström, Y., & Lemos, M. (2016). Formative interventions for expansive learning and transformative agency. *Journal of the Learning Sciences, 25*(4), 599–633. https://doi.org/10.1080/10508406.2016.1204547

Shum, S. B., Irgens, G. A., Moots, H., Phillips, M., Shah, M., Vega, H., & Wooldridge, A. (2021). Participatory quantitative ethnography. In *Advances in Quantitative Ethnography* (pp. 126–138). Virkkunen, J. (2006). Dilemmas in building shared transformative agency. *Activités, 03*(1), Article 1. https://doi.org/10.4000/activites.1850

Wei, G., & Sannino, A. (2024). Supporting transformative agency in initial teacher education: A Change Laboratory in China. *Learning, Culture and Social Interaction, 47*, 100838. https://doi.org/10.1016/j.lcsi.2024.100838

Wohlfart, O., & Wagner, I. (2023). Teachers' role in digitalizing education: An umbrella review. *Educational Technology Research and Development, 71*(2), 339–365. https://doi.org/10.1007/s11423-022-10166-0

Computational Thinking (CT) Proficiency among Malaysian Teacher Trainees: A Cross-Institutional Analysis in the Central Zone

Ahmad Sarji Abdul HAMED[a], Su Luan WONG[a]*, Mohd Zariat Abdul RANI[b], Mas Nida Md KHAMBARI[a], Nur Aira Abd RAHIM[a], Priscilla MOSES[c], Nur Izzati Mat ZIN[a]

[a]*Faculty of Educational Studies, Universiti Putra Malaysia, Malaysia*
[b]*Faculty of Modern Language and Communication, Universiti Putra Malaysia, Malaysia*
[c]*Universiti Tunku Abdul Rahman, Malaysia*
*suluan@upm.edu.my

Abstract: Computational Thinking (CT) has become an essential competency in 21st-century education, equipping individuals with structured problem-solving skills necessary for a technology-driven world. In Malaysia, the Ministry of Education has integrated CT into the national curriculum, however, research indicates varying extent of CT proficiency among teacher trainees, raising concerns about their preparedness to implement CT in classrooms. This study examines the extent of CT proficiency among teacher trainees in the Central Zone of Teacher Training Institutes or known locally as Institut Pendidikan Guru Malaysia (IPGM), focusing on their competency in abstraction, decomposition, pattern recognition, algorithmic thinking, logical reasoning, and evaluation. A quantitative approach was employed, using a validated multiple-choice instrument to assess CT proficiency among 448 teacher trainees. Descriptive and One-way ANOVA analyses were conducted to determine CT competency and the effect of institutional background towards CT proficiency among teacher trainees. The findings indicate that teacher trainees exhibit strong competencies in evaluation, decomposition, and algorithmic thinking, but face challenges in logical reasoning and pattern recognition. ANOVA analysis result confirmed significant differences across institutions. These disparities may reflect differences in institutional practices, instructional strategies, or student support systems. The study provides insights into CT integration in teacher education and offers recommendations for curriculum improvements to better equip teacher trainees with computational problem-solving skills.

Keywords: Computational thinking, teacher trainee, Institut Pendidikan Guru Malaysia, proficiency, development

1. Introduction

Computational Thinking (CT) has become a cornerstone of 21st-century education, gaining recognition worldwide as an essential skill for solving problems, designing systems, and understanding human behavior through the lens of computing (Wing, 2006). Rooted in computer science, CT involves a structured way of thinking that enables individuals to break down complex problems, recognize patterns, develop step-by-step solutions (algorithms), and apply logical reasoning and evaluation. Increasingly, however, CT is also framed not only as a set of discrete skills but as a mindset or epistemic stance, a way of approaching problems that includes dispositions such as persistence, adaptability, and reflective judgment (Grover & Pea, 2013; Brennan & Resnick, 2012). This broader view positions CT as both a procedural competence and a cognitive orientation toward systematic modeling, iterative refinement, and critical evaluation. As digital technologies continue to evolve rapidly and reshape our lives through the Fourth Industrial Revolution (IR 4.0), many countries have made CT a key part of

their education systems. The goal is clear: to prepare students for a future where technology and innovation play a central role (Istenic, 2020).

In Malaysia, the Ministry of Education (MOE) has taken proactive steps by embedding CT into the national educational curriculum at the primary and secondary school levels [Kurikulum Standard Sekolah Rendah (KSSR) and Kurikulum Standard Sekolah Menengah (KSSM)]. These curricula aim to cultivate problem-solving skills among students by focusing on core CT components such as abstraction, decomposition, pattern recognition, algorithmic thinking, logical reasoning, and evaluation (MOE, 2017). However, despite these forward-thinking policies, research reveals inconsistencies in how well teacher trainees grasp and apply CT. These disparities often stem from differences in curriculum design, teaching methods, and access to technological tools (Wing, 2011; Kang et al., 2022). As the bridge between policy and classroom practice, teacher trainees play a crucial role in shaping how CT is taught. Their own proficiency in CT directly influences how effectively they can integrate these concepts into various subjects and learning experiences. Unfortunately, studies show that many teacher trainees still struggle with specific CT components, particularly decomposition and pattern recognition, both vital for tackling real-world problems in educational settings (Ung et al., 2021).

Importantly, the value of CT in teacher education extends beyond technical know-how. It nurtures critical thinking, creativity, and adaptability, qualities essential for success in today's digital economy (Yadav et al., 2017). Theoretical perspectives such as constructivism and constructionism (Papert, 1980) emphasize that CT thrives when learners actively create, test, and refine solutions, while Bloom's taxonomy situates CT within higher-order thinking skills such as analysis, evaluation, and creation (Shute et al., 2017). These frameworks inform this study's research design and interpretation, as they guide how CT proficiency is operationalized and how findings will be linked to policy implications. It is also important to acknowledge that CT proficiency is multidimensional. While multiple-choice instruments can provide valid, large-scale measures of foundational CT knowledge and reasoning, they do not fully capture open-ended, design-based applications, iterative debugging, or reflective evaluation. Therefore, in this study, test scores will be interpreted as indicators of foundational proficiency, with the recognition that comprehensive CT assessment would require additional performance-based and reflective measures.

This study aims to assess how teacher trainees in the Central Zone of IPGM apply CT in educational settings. It will explore their extent of proficiency and examine how their academic backgrounds and institutions may influence their abilities. Ultimately, the findings will offer valuable insights into the effectiveness of current training models and help guide improvements. By providing evidence-based recommendations, this study hopes to support the ongoing effort to equip future educators with the tools they need to foster computational problem-solving in classrooms, an effort that aligns with both national goals and global educational priorities. The research questions guiding this investigation are: What is the extent of CT proficiency among teacher trainees in Institut Pendidikan Guru Malaysia (Central Zone)? Do the mean overall scores differ significantly across the five selected teacher training institutions (Central Zone)?

2. Literature Review

Research on CT in education has evolved from conceptual discussions to empirical investigations examining how it can be integrated into teaching and learning across diverse contexts. Globally, studies highlight the value of CT as both a set of cognitive tools and a mindset that supports analytical reasoning and structured problem-solving in multiple disciplines (Grover & Pea, 2013; Brennan & Resnick, 2012; Kafai & Burke, 2014). Countries such as the United States, the United Kingdom, and Finland have incorporated CT into national curricula using cross-disciplinary strategies, including coding, robotics, and algorithmic reasoning, to cultivate transferable problem-solving skills among students (Yadav & Berthelsen, 2021). The success of these initiatives has been linked to teacher preparation programs that deliberately embed CT in pedagogical training, ensuring educators can translate CT principles into classroom practice (Barr & Stephenson, 2011; Mon et al., 2020).

In Malaysia, CT is embedded in the Kurikulum Standard Sekolah Rendah (KSSR) and Kurikulum Standard Sekolah Menengah (KSSM) to promote core competencies such as abstraction, decomposition, pattern recognition, algorithmic thinking, logical reasoning, and evaluation (MOE, 2017). Despite these curricular commitments, empirical studies report uneven levels of CT proficiency among teacher trainees due to inconsistent training quality, varying access to digital tools, and differing pedagogical approaches (Kamaruddin et al., 2021).

Yadav et al. (2017) note that without intentional integration of CT concepts into teacher education, pre-service teachers often struggle to apply these skills effectively. These findings echo global concerns that CT integration remains uneven across subject domains, with STEM fields receiving more systematic attention than non-STEM areas (Weintrop et al., 2016; Yadav et al., 2014). CT's six core components (decomposition, pattern recognition, abstraction, algorithmic thinking, logical reasoning, and evaluation) are interdependent and each presents distinct instructional challenges. Decomposition, or breaking complex problems into smaller units, is a common area of difficulty due to cognitive overload (Grover & Pea, 2018). In Malaysia, scaffolding has been employed to improve decomposition skills (MOE, 2017), yet mastery remains inconsistent. Pattern recognition can facilitate predictive thinking and decision-making but is not always easily transferred between disciplines, especially without visual or interactive aids (Yusoff et al., 2021). Abstraction, the ability to focus on relevant details while ignoring the non-essential, is a priority in computational modelling and algorithm design, but educators require more targeted training to teach it effectively (Ung et al., 2021). Algorithmic thinking is often taught through programming languages like Scratch or Python, along with flowcharts and pseudocode (Grover & Pea, 2013), while logical reasoning and evaluation are nurtured through structured tasks and reflective activities (Denning, 2017; Shute et al., 2017). Pedagogical strategies grounded in constructivism (Papert, 1980) and Problem-Based Learning (PBL) have been shown to effectively develop CT skills by immersing learners in authentic, real-world problems requiring systematic analysis and iterative refinement (Weintrop et al., 2016).

These approaches align CT with higher-order thinking skills in Bloom's taxonomy, particularly in the domains of analysis, evaluation, and creation (Shute et al., 2017). Studies also highlight the potential of CT in non-STEM contexts, such as language and social sciences, where decomposition supports discourse analysis, pattern recognition aids in identifying thematic structures, and algorithmic thinking underpins writing strategies (Kafai & Burke, 2014; Yadav et al., 2014). However, persistent misconceptions that CT is exclusively technical hinder its adoption in these areas. Assessment of CT proficiency has relied on varied instruments, including multiple-choice tests, performance-based tasks, and self-report surveys. Multiple-choice assessments, such as the Computational Thinking Performance Test (CTPT) (Román-González et al., 2017) and the competent Computational Thinking test (cCTt) (El-Hamamsy et al., 2022), are valued for their efficiency, objectivity, and scalability (Weese & Feldhausen, 2017). When supported by strong psychometric evidence, these tools can validly measure CT's core components but may overlook design-based and reflective dimensions (Brennan & Resnick, 2012; Shute et al., 2017). In Malaysia, there is limited research on validated CT instruments for teacher trainees, and few studies have examined concurrent validity or the relationship between measured CT proficiency and teaching readiness. Taken together, the literature demonstrates both the global momentum for CT integration and the specific challenges facing Malaysian teacher education. While frameworks, competencies, and pedagogical strategies are well-documented, gaps remain in empirical evidence on how teacher trainees develop and apply CT in authentic teaching contexts. Addressing these gaps particularly through context-specific assessment tools and targeted professional development will be essential for equipping future educators to foster computational problem-solving across disciplines.

3. Methodology

This study employed a quantitative research design to examine the extent of CT among teacher trainees at IPGM in the Central Zone. The quantitative approach was chosen for its

ability to produce objective and statistically analyzable data, allowing for the identification of trends and relationships among variables. As emphasized by Creswell (2018) and Fraenkel and Wallen (2019), this method also supports broader generalization of findings and enhances the study's replicability and precision. To assess CT skills, a multiple-choice questionnaire (MCQ) was specifically developed for this research. The instrument was constructed to reflect six core components of CT: abstraction, decomposition, pattern recognition, algorithmic thinking, logical reasoning, and evaluation. The initial item pool was generated based on a comprehensive review of existing CT frameworks and assessment literature, including the works of Wing (2006) and Shute et al. (2017). Each question was designed around real-world problem-solving scenarios to encourage higher-order thinking and application of CT concepts.

The content validation of the instrument was conducted using the Delphi method, involving a panel of 3 experts with extensive experience in computational thinking, educational technology, and assessment. The validation process consisted of three iterative rounds. In the first round, experts reviewed each item for clarity, relevance, and alignment with CT components, and qualitative feedback was collected to inform revisions. In the second round, items were rated using a four-point Likert scale to assess relevance and clarity, with items scoring below the threshold Content Validity Index (CVI) of .80 being revised or removed. The final round achieved expert consensus, and the overall CVI across all items reached .93, indicating strong content validity and expert agreement. Following content validation, a pilot study was conducted with 60 teacher trainees from a different IPGM campus to test the instrument's psychometric properties. Item analyses were performed using classical test theory. Items with difficulty indices (p-values) between .30 and .90 were retained, ensuring a balanced level of challenge. Discrimination indices (D-values) were calculated, and items with values of .52 considered acceptable for distinguishing between high and low performers. Additionally, distractor analysis was conducted to evaluate the plausibility of incorrect answer choices, with non-functioning distractors being revised. The internal consistency of the instrument was measured using the Kuder-Richardson Formula 20 (KR-20), yielding a reliability coefficient of.84, which reflects high internal consistency. The final instrument comprised 30 refined items, with an even distribution across the six CT components.

The study sample consisted of 448 teacher trainees from IPGM campuses located in the Central Zone of Malaysia. The sample size was determined using Cochran's formula for large populations to ensure sufficient statistical power (Cochran, 1977). To capture the diversity of the trainee population, stratified sampling was employed based on academic programs and academic levels. This sampling strategy enhanced the representativeness of the sample and reduced sampling bias, particularly important in educational research where background and training can influence CT proficiency. Data collection was carried out through face-to-face administration of the MCQ across selected IPGM campuses. This method allowed the researchers to provide immediate clarification, reduce the likelihood of misinterpretation, and ensure the integrity of responses. Moreover, it minimized the risk of technical issues and disengagement commonly associated with online data collection. The in-person format contributed to a high response rate and improved the overall quality and completeness of the dataset. Trained enumerators facilitated the data collection process to maintain consistency and adherence to ethical standards.

The collected data were analyzed using IBM SPSS Statistics Version 28. Descriptive statistics, including means, standard deviations, and frequencies, were used to summarize overall CT performance. To determine whether significant differences existed in CT scores across groups, a One-Way Analysis of Variance (ANOVA) was conducted. The groups were defined by institution. Where significant differences were observed, post hoc test was applied to identify specific group differences while controlling for Type I error. This approach aligned with recommended procedures for robust statistical testing, as outlined by Scheffé (1999) and Pallant (2020). Despite its methodological strengths, the study acknowledges several limitations. Although stratified sampling enhanced internal validity, the geographic scope was limited to the Central Zone, potentially restricting the generalizability of the findings to other regions of Malaysia. Furthermore, while face-to-face data collection yielded high-quality data, it was time-intensive and required substantial logistical resources. Future research could address these limitations by expanding the geographic scope and incorporating qualitative

methods, such as interviews or open-ended responses. Such approaches would provide deeper insights into how teacher trainees conceptualize and apply CT within diverse educational contexts.

4. Result

This study examines the extent of CT among 448 teacher trainees in IPGM, Central Zone, analyzing gender distribution, institutional representation, and course specialization. The sample comprises 143 males (31.9%) and 305 females (68.1%), reflecting the typical gender ratio in Malaysian teacher education. Respondents are from five IPGM Kampus, with the largest group from IPGK Pendidikan Islam (35.9%), followed by IPGK Ilmu Khas (18.5%), IPGK Ipoh (18.1%), IPGK Bahasa Melayu (16.3%), and IPGK Bahasa Antarabangsa (11.2%), ensuring diverse representation. The Malay Language specialization (31%) is the most common, followed by TESL (16.1%), Physical Education (11.2%), Islamic Education (9.4%), and Home Science (9.4%). Other fields include Special Education, Mathematics, Biology, Arabic Language, Music, and Early Childhood Education, showcasing the multidisciplinary nature of teacher training. The demographic diversity provides valuable insights into the integration of CT across various disciplines. Table 1 shows demographic profile of respondents in the study.

Table 1. *Demographic Profile of Respondents in The Study.*

Items	Details	Frequencies	Percentage (%)
Gender	Male	143	31.9
	Female	305	68.1
Institution	IPGK Pendidikan Islam	161	35.9
	IPGK Ilmu Khas	83	18.5
	IPGK Bahasa Antarabangsa	50	11.2
	IPGK Bahasa Melayu	73	16.3
	IPGK Ipoh	81	18.1
Main Course of Study	Malay Language	139	31
	Teaching English as a Second Language (TESL)	72	16.1
	Islamic Education	42	9.4
	Home Science	42	9.4
	Physical Education	50	11.2
	Special Education	11	2.5
	Mathematics	15	3.3
	Biology	11	2.5
	Arabic Language	25	5.6
	Music	16	3.6
	Early Childhood Education	25	5.6

4.1 Extent of CT proficiency among teacher trainees in Institut Pendidikan Guru Malaysia (Central Zone).

Table 2 presents the average percentage of correct and wrong responses for six components of CT among teacher trainees in Institut Pendidikan Guru Malaysia (Central Zone). The results indicate varying levels of competency in CT components. Evaluation recorded the highest

correct percentage (72.32%), while Logical Reasoning had the lowest (55.13%). Pattern Recognition also showed a relatively lower performance, with 38.84% wrong responses.

Table 2. *Average Percentage of Correct and Wrong Responses.*

Components of CT	Average Percentage (%)	
	Correct	Wrong
Abstraction	65.18	34.82
Decomposition	68.08	31.92
Pattern Recognition	61.16	38.84
Algorithm	69.64	30.36
Logical Reasoning	55.13	44.87
Evaluation	72.32	27.68

4.2 Mean Overall Scores (One-way ANOVA)

One-way ANOVA was conducted to examine the differences in overall achievement scores across five educational institutions (IPGKPI, IPGKIK, IPGKBA, IPGKBM, and IPGKI). The analysis includes tests of homogeneity of variances, ANOVA, robust tests of mean equality, and post-hoc multiple comparisons using Tamhane's T2 and Games-Howell procedures, in line with Scheffe (1999) methodological rigor for post-hoc analysis.

4.2.1 Test of Homogeneity of Variance

The assumption of homogeneity of variances was evaluated using Levene's test. The results were statistically significant across all test types as shown in Table 3. Given the significance level ($p < .05$), the assumption of equal variances is violated. Therefore, robust tests and multiple comparisons suitable for unequal variances were applied.

Table 3. *Levene's Test Result.*

Test Type	Levene Statistic	df1	df2	Sig.
Based on Mean	3.065	4	443	.016
Based on Median	3.101	4	443	.016
Based on Median (adj. df)	3.101	4	436	.016
Based on Trimmed Mean	3.177	4	443	.014

4.2.2 One-Way ANOVA

The ANOVA indicated statistically significant differences in overall scores between institutions as shown in Table 4. The significant F-statistic ($p < .05$) suggests that not all group means are equal. However, since homogeneity of variance was violated, robust tests were also considered.

Table 4. *One-way ANOVA Result*

Source	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2560.585	4	640.146	5.487	.000
Within Groups	51686.931	443	116.675		
Total	54247.516	447			

4.2.3 Robust Tests of Equality of Means

Robust procedures (Welch and Brown-Forsythe) confirmed significant differences in means as shown in Table 5. Both tests yielded p-values less than .05, further validating the presence of statistically significant differences between groups, even when adjusting for heterogeneity.

Table 5*: Welch and Brown-Forsythe Result*

Source	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2560.585	4	640.146	5.487	.000
Within Groups	51686.931	443	116.675		
Total	54247.516	447			

4.2.4 Post-Hoc Multiple Comparisons

Due to unequal variances, Tamhane's T2 and Games-Howell post-hoc tests were employed. Both tests are conservative and appropriate for heteroscedastic data. Significant pairwise differences ($p < .05$) are highlighted. The results of Tamhane's T2 post-hoc analysis revealed significant differences between certain institutional pairs. Specifically, students from IPGKIK scored significantly higher than those from IPGKPI, with a mean difference of -6.637 ($p = .000$). This indicates that the mean overall achievement score for IPGKIK was substantially greater than that for IPGKPI. Additionally, the comparison between IPGKIK and IPGKBM showed a significant mean difference of 5.825 ($p = .006$), suggesting that students from IPGKBM outperformed those from IPGKIK in overall scores. The Games-Howell procedure, which similarly adjusts for unequal variances and sample sizes, corroborated these findings. The same pairwise differences were found to be statistically significant: IPGKPI versus IPGKIK (mean difference = -6.637, $p = .000$) and IPGKIK versus IPGKBM (mean difference = 5.825, $p = .006$). The consistency of results across both post-hoc methods lends strong support to the reliability of these observed differences.

5. Discussion

The findings provide critical insights into CT proficiency among IPGM teacher trainees in the Central Zone, revealing notable variations across components and institutions. Strength in the Evaluation component aligns with Shute et al. (2017), who link evaluative thinking to reflective practices such as peer feedback, self-assessment, and classroom reflections, fostering metacognitive awareness. This supports constructivist principles (Papert, 1980) and higher-order thinking in Bloom's Taxonomy, consistent with international findings where reflective pedagogies enhance self-directed learning (Mon et al., 2020). Strong performance was also observed in Algorithmic Thinking and Decomposition, echoing national (MOE, 2017) and international research (Grover & Pea, 2013) that link scaffolded activities (flowcharts, pseudocode, Scratch/Python) to structured reasoning. The integration of decomposition in KSSR and KSSM (MOE, 2017) appears effective, though possible surface-level mastery may limit transferability to abstraction (Ung et al., 2021). Weaker performance in Logical Reasoning and Pattern Recognition mirrors prior findings on their abstract demands (Weintrop et al., 2016; Denning, 2017).

These skills identifying structures across problems and supporting systematic decision-making require explicit, context-rich instruction (Yadav et al., 2017). Practical strategies include comparative case analysis, "odd one out" pattern drills, visual analogy mapping, structured debates, and discipline-specific thematic mapping. In STEM subjects, logic puzzles and conditional flowchart exercises can strengthen reasoning processes, while in non-STEM contexts, literary structure analysis and historical cause–effect mapping can enhance pattern recognition. Variations across institutions reflect systemic factors such as curriculum design, pedagogical practices, resource availability, institutional philosophy, and exposure to CT-integrated subjects echoing Kamaruddin et al. (2021) and Yusoff et al. (2021), who note that even within a centralized framework, institutional culture and educator expertise significantly shape CT outcomes. Methodologically, reliance on a multiple-choice instrument, while efficient and objective, may limit the capture of authentic CT application. Prior studies

have shown that open-ended or design-based assessments such as programming projects, computational modeling, and simulation-based problem-solving often reveal deeper reasoning processes and creativity not visible in MCQ formats (Román-González et al., 2017; Grover & Pea, 2013). Differences in outcomes between these formats are often linked to curricular emphasis, pedagogical scaffolding, available resources, and institutional priorities.

A mixed-methods approach combining MCQs with performance-based tasks and artefact analysis would provide a richer, context-sensitive profile of CT proficiency. These results reaffirm CT's multidisciplinary relevance beyond STEM (Grover & Pea, 2013; Kafai & Burke, 2014). In language education, decomposition and pattern recognition support narrative analysis, while algorithmic thinking can structure writing processes. In social sciences, logical reasoning can be fostered through structured debate, and thematic mapping can develop pattern recognition in history or sociology. Addressing persistent misconceptions that CT is STEM-exclusive (Yadav et al., 2014; Weintrop et al., 2016) requires targeted professional development, cross-disciplinary collaboration, and curriculum exemplars that integrate CT into diverse subject areas. Theoretically, these findings support a holistic conceptualization of CT as a universal problem-solving framework (Wing, 2011) grounded in constructivist learning theories (Papert, 1980). Embedding CT in teacher education demands systemic support, targeted training, and authentic assessment to move beyond surface exposure toward deep, reflective integration. The Malaysian case mirrors global challenges, underscoring the need to adapt best practices to local contexts while ensuring equitable CT skill development across institutions and disciplines.

6. Conclusion

The analysis of CT competencies indicates that trainees excel in Evaluation, Algorithmic Thinking, and Decomposition but face persistent challenges in Logical Reasoning and Pattern Recognition, reflecting strengths in structured and reflective thinking alongside a need for enhanced support in abstract and analytical skills. Institutional variations suggest that differences in teaching strategies, support systems, and program designs significantly shape CT outcomes, underscoring the importance of a standardized yet adaptable integration framework that extends beyond STEM into disciplines such as language education. Targeted professional development for teacher educators is essential to cultivate CT not merely as a set of discrete skills but as a broader cognitive stance toward problem-solving, learning, and design. While the study offers valuable comparative insights, its cross-sectional, single-method design constrains interpretive depth; incorporating interviews, classroom observations, and trainee reflections could provide richer contextual understanding of how CT is developed and applied. Future research should adopt mixed-methods designs to capture CT in authentic teaching contexts, implement longitudinal and intervention-based approaches to monitor skill progression, and rigorously validate CT assessment tools for cultural and curricular relevance within the Malaysian educational landscape.

Acknowledgements

This study is supported by the Fundamental Research Grant Scheme (FRGS) of the Malaysian Ministry of Higher Education (FRGS/1/2023/SSI07/UPM/01/1). The assistance of the Research Management Center (RMC) of Universiti Putra Malaysia in coordinating and distributing fund for this research is greatly appreciated.

References

Barr, V., & Stephenson, C. (2011). Bringing computational thinking to K-12: what is Involved and what is the role of the computer science education community? *Inroads, 2*, 48-54.

Brennan, K., & Resnick, M. (2012). New frameworks for studying and assessing the development of computational thinking. *In Proceedings of the 2012 annual meeting of the American educational research association, Vancouver, Canada*, 1-25.

Cochran, W. G. (1977). *Sampling Techniques (3rd Ed.*). John Wiley & Sons.

Creswell, J. W. (2014). *Research Design: Qualitative, Quantitative, And Mixed Methods Approaches (4th Ed.).* SAGE Publications.
Creswell, J. W., & Creswell, J. D. (2018). *Research Design: Qualitative, Quantitative, And Mixed Methods Approaches (5th Ed.).* SAGE Publications.
Denning, P. (2017). Computational Design. *Ubiquity*, 1-9.
Fraenkel, J. R., & Wallen, N. E. (2019). *How To Design And Evaluate Research In Education (10th Ed.*). McGraw-Hill Education.
Grover, S., & Pea, R. (2013). Computational thinking in K-12: A review of the state of the field. *Educational Researcher, 42*(1), 38-43.
Grover, S., & Pea, R. (2018). *Computational Thinking: A Competency Whose Time Has Come.* In S. Sentance, E. Barendsen, & C. Schulte (Eds.), Computer Science Education: Perspectives on Teaching and Learning in School (p. 20-38). Bloomsbury Publishing.
Istenic, A. (2020). *Educational Technology And The Construction Of Authentic Learning Environments.* Slovene: Univerza v Ljubljani.
Kafai, Y. B. (2016). Education from computational thinking to computational participation in K-12 education seeking to reframe computational thinking as computational participation. *Communications of the ACM, 59*(8).
Kafai, Y. B., & Burke, Q. (2014). *Connected Code: Why Children Need To Learn Programming.* MIT Press.
Kamaruddin, A.R., Low, J.H., & Sarwar, A. (2021). Enhancing the adoption of computational thinking education among the malaysians to prepare for future industrial revolution. *Proceedings of the 2021 5th International Conference on Education and E-Learning.*
Kang, C., Liu, N., Zhu, Y., Li, F., & Zeng, P. (2022). Developing college students' computational thinking multidimensional test based on life story situations. *Education and Information Technologies, 28*(3), 2661-2679.
Ministry of Education (MOE). (2017). *Kurikulum Standard Sekolah Rendah (KSSR) dan Kurikulum Standard Sekolah Menengah (KSSM): Computational Thinking Framework.* Malaysian Ministry of Education.
Mon, F. M. E., Llopis, M., & Segura, J. A. (2020). Digital competence and computational thinking of student teachers. *International Journal of Emerging Technologies in Learning (iJET), 15*(02), 29.
Pallant, J. (2020). *SPSS Survival Manual: A Step By Step Guide To Data Analysis Using IBM SPSS (7th Ed.).* Routledge.
Papert, S. 1980. *Mindstorms: Children, Computers, and Powerful Ideas.* NY: Basic Books, Inc.
Scheffe, H. (1999). *The Analysis of Variance*. New York, NY: John Wiley & Sons.
Shute, V.J., Sun, C., & Asbell-Clarke, J. (2017). Demystifying computational thinking. *Educational Research Review, 22*, 142-158.
Skulmoski, G. J., Hartman, F. T., & Krahn, J. (2007). The Delphi method for graduate research. *Journal of Information Technology Education: Research, 6*(1), 1-21.
Ung, L., Labadin, J., Mohamad, F. (2021). Computational thinking for teachers: Development of a localised e-learning system. *Computers & Education*, (177), 104379.
Weintrop, D., Beheshti, E., Horn, M. S., Orton, K., Jona, K., Trouille, L., & Wilensky, U. (2016). Defining computational thinking for mathematics and science classrooms. *Journal of Science Education and Technology, 25*(1), 127–147.
Wing, J. (2011). *Research Notebook: Computational Thinking – What And Why?* The link. Carneige Mellon.
Wing, J. M. (2006). Computational thinking. *Communications of the ACM, 49*(3), 33-35.
Yadav, A., & Berthelsen, U. (Eds.). (2021). *Computational Thinking in Education: A Pedagogical Perspective (1st ed.)*. Routledge.
Yadav, A., Hong, H., Stephenson, C. (2017). Computational thinking for all: Pedagogical approaches to embedding 21st century problem solving in k-12 classrooms. *Techtrends, 6*(60), 565-568.
Yadav, A., Mayfield, C., Zhou, N., Hambrusch, S., & Korb, J. T. (2014). Computational thinking in elementary and secondary teacher education. *ACM Transactions on Computing Education, 14*(1), 5.
Yusoff, K. M., Ashaari, N. S., Wook, T. S. M. T., Ali, N. M. (2021). Validation Of the Components And Elements Of Computational Thinking For Teaching And Learning Programming Using The Fuzzy Delphi Method. IJACSA, 1(12). https://doi.org/10.14569/ijacsa.2021.0120111

Predicting Learners' State of Mind from Cross-Contextual Educational Log Data

Junya ATAKE[a*], Chia-yu HSU[b] & Hiroaki OGATA[b]
[a] *Graduate School of Informatics, Kyoto University, Japan*
[b] *Academic Center for Computing and Media Studies, Kyoto University, Japan*
*atake.junya.86t@st.kyoto-u.ac.jp

Abstract: This study investigated the techniques for predicting learners' state of mind from cross-contextual educational log data. The digitization of the educational environment has led to the accumulation of log data, which is expected to facilitate the identification of subtle changes in learners related to G&C characteristics. However, the utilization of log data for G&C has not been examined sufficiently, nor has the change point detection (CPD) for G&C characteristics been explored. Therefore, this study predicted learners' state of mind from cross-contextual educational log data using Bayesian Online Change Point Detection with extended processes. We applied the extended steps and resampling, which have been attempted in the medical field, on real-world educational data with many missing values. As a result, we achieved a maximum AUC of 0.697, demonstrating the potential to predict changes in learners' state of mind from educational log data. Although there are still many points to consider, such as missing value handling methods and the true positive rate for detecting students with problems, this study presents important results for capturing and predicting changes in learners' state of mind, a crucial aspect of G&C.

Keywords: Guidance and counseling, Change point detection, Learning analytics, Across context

1. Introduction

Guidance and Counseling(G&C) in school education is crucial for creating a favorable learning environment and supporting learners' holistic development (Naraswari 2024; MEXT 2022). Early detection of academic, behavioral, and social-emotional challenges, followed by timely interventions, is crucial for fostering meaningful success experiences among learners (MEXT, 2022; Ito et al., 2023; Nitz et al., 2023). Teachers, particularly homeroom teachers, are considered to play key roles in detecting these subtle changes in learners early and intervening appropriately (MEXT, 2022; Green et al., 2017). In practice, some efforts rely on homeroom teachers sharing daily observational records to identify minor behavioral changes. Nevertheless, challenges such as the burden of regular documentation and reliance on subjective assessments have been highlighted (MEXT, 2022). On the other hand, the digitization of the educational environment has led to the accumulation of log data related to learners' learning and daily lives (Ogata et al., 2023). It is expected that integrating these log data with other sources will enable teachers to better understand learners and enhance G&C strategies (MEXT, 2022). However, using cross-contextual log data for G&C has not been examined sufficiently. One reason is the difficulty of interpreting granular data without considering the subject characteristics and activity contexts.

For these challenges, Atake et al. (2024) proposed a method for extracting characteristics that collects and standardizes log data to determine if the learner is in good standing, supporting teachers' understanding of learners, even for non-subject-specific teachers. Teachers evaluated these characteristics as helpful for identifying learners' potential issues and supporting their understanding. These techniques enable the identification of learner characteristics and support teachers in understanding their learners. However, the study by Atake et al. (2024) only proposed a method for capturing learners' states at a single point in time and did not explore changes over time across contexts.

This limitation reflects a broader gap in the field of learning analytics (LA), where similar approaches, such as change point detection for risk identification, have primarily focused on predicting well-defined academic outcomes like grades or dropout. In contrast, relatively few studies have explored the detection of early warning signs or dynamic behavioral and psychological changes in real-time learning environments (Li et al., 2024). Considering the above, this study proposes a method for identifying changes in learners by combining the approach proposed by Atake et al. (2024) with the Bayesian Online Change Point Detection (BOCPD) method proposed by Adams et al. (2007) and applying it to educational log data. The objective of this study is to predict learners' state of mind, a key indicator in G&C, using educational log data. This approach contributes to proposing an innovative method for G&C, the process of identifying changes in learners' state of mind, traditionally based on observable academic behaviors or performance, by offering a data-driven perspective for capturing such changes.

2. Literature Review

2.1 Capturing learner changes in Guidance and Counseling

As part of the process of understanding the learner in G&C, recognizing learners' subtle changes is essential for early support and intervention (MEXT, 2022). According to Green et al. (2017), teachers are considered to play a key role in identifying subtle changes in learners. Then, how and what changes are expected to be identified? The specific signs are academic, behavioral, and socioemotional changes such as poor academic performance, social isolation, and sudden changes in behavior (Green et al., 2017). Green et al. (2017) emphasize the importance of integrating both observational approaches, such as teachers' intuition, and objective methods, including standardized screenings, academic data, and student self-reports. To support this process, it has adopted frameworks such as Multi-Tiered Systems of Support (MTSS), which provide tier-structured, data-driven strategies for addressing diverse learner needs (Ito et al., 2023; Nitz et al., 2023; Miller et al., 2014). In addition to these, focusing on health and changes in physical aspects is crucial (Green et al., 2017). Through the improvement of individual teachers' observation skills and sensitivity, and the development of a system of multidisciplinary cooperation and data utilization as a school organization, it is expected that a comprehensive system will be established to catch small signs of children daily without overlooking them, leading to the prevention and early resolution of problems. However, it has been noted that internal symptoms and minor signs are often overlooked because they may not be apparent in students with good academic performance (Kim et al., 2022). In addition, MTSS does not focus on real-time detection of early signs, which remains underdeveloped from data.

2.2 Change Point Detection Techniques

Although mean and variance are generally used to understand changes from educational big data, they can capture only broad changes, such as those occurring over large intervals as statistical indicators (Nakamura et al., 2024). Therefore, the Change Point Detection (CPD), which refers to the technique for identifying points in time-series data where statistical properties change, is necessary to find notable changes over time (Nakamura et al., 2024). CPD methods are generally categorized into two types: offline methods, which analyze data retrospectively, and online methods, which detect changes in real-time as data is collected (Truong et al., 2020). This study focuses exclusively on online methods for real-time monitoring.

Previous studies have examined several online methods. Eubanks-Carter et al. (2012) investigated CPD techniques to detect key moments in psychological therapy, specifically "ruptures" and "resolutions" in the therapeutic alliance, evaluating four main approaches: Control chart methods, Partitioning methods, Criterion-based methods, and Regression-based methods. Control chart methods, such as Shewhart charts, Cumulative Sum (CUSUM), and Exponentially Weighted Moving Average (EWMA), detect changes sequentially with low

computational cost. In educational research, they have been applied to real-time monitoring of learner performance and engagement (Prasad, 2020; Shimada et al., 2018). For instance, Prasad (2020) used control charts to track learner grades, highlighting that Shewhart charts are more sensitive to abrupt shifts, while CUSUM is better suited for detecting small, sustained changes. Shimada et al. (2018), although not explicitly referring to "Control charts," proposed an online method that updated models per minute to detect changes in click activity during lectures. Early warning systems in learning analytics often use threshold-based triggers, which conceptually resemble simplified control charts for engagement indicators. However, few studies have addressed multivariate control charts or personalized chart designs that account for individual differences. Partitioning methods divide the entire time series into multiple segments to detect multiple change points. Representative techniques include Binary Segmentation and PELT (Pruned Exact Linear Time) (Truong et al., 2020). Nakamura et al. (2024) proposed a method using PELT to automatically detect changes in learners' engagement levels during class, based on digital textbook logs. However, these methods may involve high computational costs and statistical limitations. Most applications in education have been restricted to univariate or low-dimensional data.

As a means of addressing these challenges, there is the Bayesian online change point detection method (Adams et al., 2007). While conventional methods detect change points retrospectively based on all past data, this method uses only the data observed so far to predict what will happen next while detecting change points. Additionally, while many previous Bayesian methods were offline methods, this method enables real-time prediction based on accurate Bayesian inference, allows for sequential updates, and features a modular design that provides flexibility for easy application to various types of data (e.g., Gaussian, Poisson). However, this method is designed under the assumption of continuously observed data, where the run length is updated sequentially at each time step. As such, it does not natively handle missing observations, and additional mechanisms are required to accommodate data with random missing values. Furthermore, the proposed approach primarily targets univariate or low-dimensional data and does not readily scale to high-dimensional settings, where computational and statistical challenges arise.

An example application is found in Romero-Medrano (2022), who extended a BOCPD method to detect behavioral shifts in individuals with mental disorders using multiple sensor data from smartphones and wearable devices. The objective was to predict self-injurious and aggressive behaviors, including suicide risk, up to one week in advance, and the result showed an AUC of 0.79. The system utilized high-dimensional, heterogeneous, and frequently missing data such as step counts, GPS-based mobility, time spent at home, app usage, sleep patterns, and emotion logs, collected via the digital app. To address real-world constraints—irregular sampling, missing values, and mixed data types—the method integrated BOCPD with latent variable models, local observation models, and weighted integration mechanisms, was validated as effective.

2.3 Significance of This Study and Research Questions

The digitization of educational environments has led to the daily generation and accumulation of a large volume of log data related to learners' learning activity and daily lives. This growing body of educational log data provides an unprecedented opportunity to quantitatively evaluate learners' behavior and its changes. On the other hand, there are few examples of real-time change point detection for multiple log data in a real-world educational context, and these are the few that have been studied to improve teachers' G&C (Table 1).

Therefore, this study aims to verify the applicability of BOCPD on real-world educational log data and to predict learners' mental health conditions using these data. To this end, we propose using the Bayesian Online Change Point Detection (BOCPD) method and its extension developed by Romero-Medrano (2022) in the field of psychiatry and applying it to educational log data in this study. As an important G&C indicator, we utilized state of mind, a subjective self-report input by learners, and applied BOCPD techniques innovatively. This is expected to contribute to the early identification of changes in learners within the G&C

framework. To clarify the scope of this study, and based on a comprehensive review of the existing literature, the following research questions are posed:

- RQ1: How much change can be detected using BOCPD on real-world educational log data?
- RQ2: To what extent can learners' state of mind be predicted based on educational log data?

Table 1 Research Position of this study

	Nakamura et al. (2024)	Eubanks-Carter et al. (2012)	Romero-Medrano (2022)	This Study
Educational-Context	✓	×	×	✓
Multi-contextual log data	×	×	✓	✓
Real-time detection	×	✓	✓	✓
Method	PELT	EWMA, CUSUM etc.	BOCPD	BOCPD
Proposal	Detects multiple changes using PELT from educational big data	Detects alliance ruptures and resolutions in psychotherapy	Predict self-injurious and aggressive behaviors up to one week	Predicting changes in learner's state of mind from cross-contextual log data

3. Data Context and Methods

3.1 Data Collection and Context

This study used log data of daily activities in the Goal-Oriented Active Learner (GOAL) system to address RQs. This system serves as a tool to support learners' activities both within and beyond the school curriculum and collects and integrates personal health and learning data logs from multiple sources, such as a GARMIN device and the e-book reader (Li et al., 2021). From a G&C perspective, this is valuable as it enables comprehensive tracking of learners' activities. We used five contexts of log data: math activity time, English activity time, math score, stress level, and steps taken. The data gathered from 116 learners covers the period from June 2020, when the learners enrolled (the start was two months later than usual due to COVID-19), to March 2023. In addition, we use the state of mind (self-reported data input by learners) data collected through health observations on Moodle. Health observations are mandatory in Japanese public schools, and with the recent spread of ICT tools, data such as sleep duration, temperature, and state of mind (recorded as numerical values ranging from 0 to 100) have been collected. This study analyzes the five educational log data points and the state of mind data.

3.2 Data Processing

Figure 1 shows the procedure of data processing. The five log data sets and the state of mind data are from different activity contexts, so the original values have different meanings. Therefore, we first attempt to collect and standardize the data using the characteristics

proposed by Atake et al. (2024). The scaled values range from 0 to 10, where 10 indicates a positive state and 0 indicates a negative state in any context.

Next, the five log datasets are resampled on a daily and weekly basis because the log data is collected at different times and recorded regardless of whether a weekday or a weekend. In addition, the state of mind is generally recorded daily during the school year, but not on weekends, holidays, or during long school breaks. For these reasons, we resampled data 1 day and 7 days to attempt to detect changes in learners. This enabled us to capture both short-term fluctuations and medium- to long-term behavior patterns. During this process, activity time-related metrics are rescaled based on the total activity time for that day, while other metrics are averaged over the day or week. The resampled data is restructured according to activity time and described in five dimensions, and we applied BOCPD to these two types of data to compare which is the most effective in predicting learners' state of mind.

Furthermore, we confirmed the available data points and obtained the results shown in Figure 2. Therefore, in this study, we finally chose the period from October 2021 to March 2022, which is six months with abundant data for the five indicators.

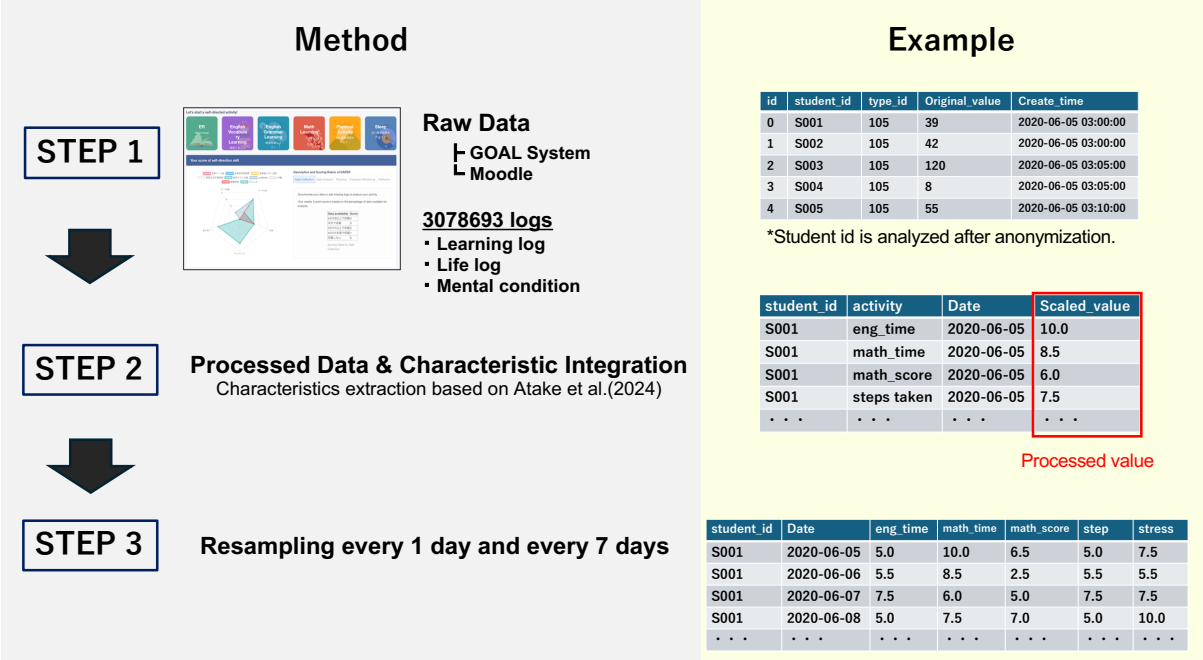


Figure 1 The Procedure of Data Processing

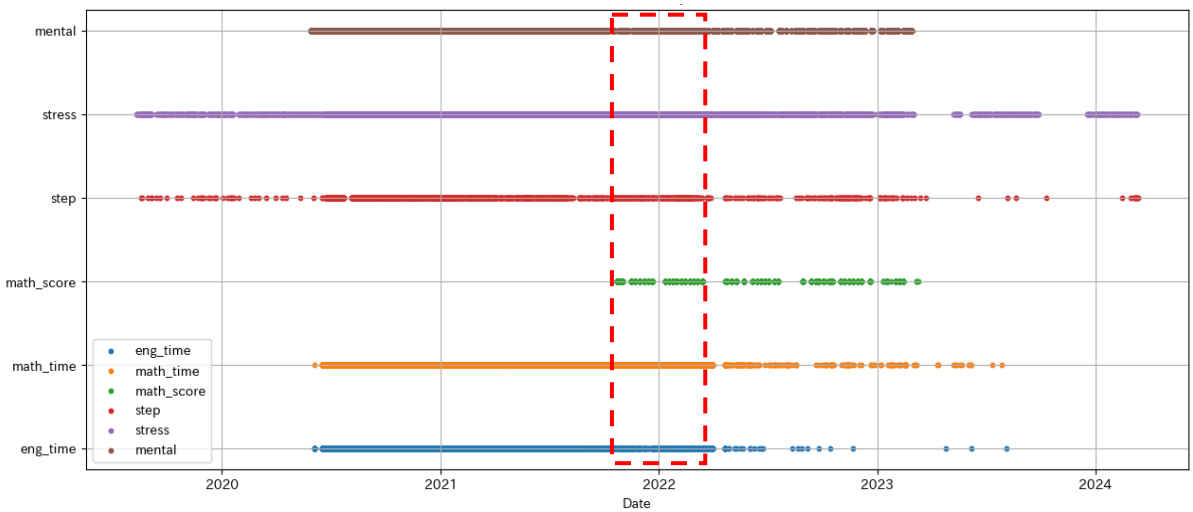


Figure 2 Available Data Periods in Each Dataset

3.3 Extension of Bayesian Online Change Point Detection

The original BOCPD method assumes continuously observed data and does not contain missing values. Therefore, we first confirmed the extent of missing values, and many missing values were identified, especially in variables such as steps and stress that were not directly linked to school activities. To address this, we compared two different strategies for handling missing values: (i) imputing missing entries with zero, treating them as if there was no activities, following the proposal by Atake et al (2024) and (ii) a proposed weighted approach in which weekly aggregated data were used while reducing the influence of weeks with high proportions of missing observations. In addition, following the proposal by Romero-Medrano (2022), some process steps were conducted. Although each variable has been rescaled to a standardized 0–10 score, the original data are derived from heterogeneous sources with different statistical types and missing patterns. Thus, despite standardization, the underlying structure remains high-dimensional and complex. To capture latent patterns of behavior, we apply a Latent Variable Model (LVM) to map each day's observation to a probabilistic mixture of classes (1).

$$p(x_t|\theta_t) = \sum_{k=1}^{k} p(x_t|z_t = k) \cdot p(z_t = k|\theta_t) \quad (1)$$

This equation shows that the observed data x_t is generated from a latent class z_t, where each class has a different probability and a different distribution.

Next, the ambiguity of the multi-class distribution is treated in multinomial Sampling, and the same method is applied in this study. Since the output of latent variable models is probability-based, it is often a flat distribution. Therefore, multinomial Sampling is effective to address the challenge. One sample is selected according to the probability distribution and treated as a pseudo-definite “observation class. This is said to increase the accuracy of the change point detection model used in the next step (2).

$$z_t^s \sim Categorical\big(p(z_t|x_t)\big) \text{ for } s = 1, \dots, S \quad (2)$$

This produces S samples for each time t, which reflect the likely latent states.

Based on the above steps, the analysis can be performed by BOCPD, which estimates the probability that a change has occurred at each time point. If the latent class series suddenly changes, it is detected as the beginning of a new behavior pattern. The joint distribution of run-length and data sequence is given by(3)

$$p(r_t|z_{1;t}) = \sum_{r_{t-1}} p(r_t|r_{t-1}) \cdot \Psi_t^{(r)} \cdot p(r_{t-1}, z_{1:t-1}) \quad (3)$$

$\Psi t(r)$ shows the predictive likelihood of the current latent class given the past, and rt is the time since the last change point. Based on the above processing, to answer RQ1, we will verify the applicability of BOCPD by examining the consistency rate between the log data applied with BOCPD and the state of mind. Then, to answer RQ2, we will investigate whether it is possible to predict state of mind using logistic regression based on the change points detected by BOCPD.

4. Result and Discussion

4.1 RQ1. Exploring the Applicability of BOCPD Through Change Point Overlap Rates

To answer RQ1, we used BOCPD to extract behavioral change points based on educational and life logs and examined the extent to which they matched learners’ self-reported “state of mind,” which was treated as a pseudo ground truth. In this study, we analyzed 116 learners, and it was difficult to retrospectively obtain teacher-provided labels for behavioral change. Therefore, even subtle changes in the self-reported mental states were regarded as valid

labels, and the overlap between these labels and the behavioral change points extracted from five log indicators was evaluated.

To capture overlaps, we compared cases where detected change points coincided perfectly with the date of the state of mind change (±0 days), as well as cases allowing for temporal tolerance (±1 day in the 1-day resampled data and ±7 days in the 7-day resampled data). For each setting, we computed precision, recall, and F1 scores between the detected change points and the state of mind change points. Table 2 summarizes the results.

Table 2 Result of Analysis

Resampling Unit	Method	Target Range	Precision	Recall	F1 Score
1 day	Zero-fill	±0 day	0.096	0.513	0.162
		±1 day	0.367	0.994	0.536
	Weighted	±0 day	0.109	0.224	0.147
		±1 day	0.197	0.411	0.266
7 days	Zero-fill	±0 day	0.094	0.179	0.123
		±7days	0.243	0.465	0.319
	Weighted	±0 day	0.109	0.115	0.112
		±7days	0.230	0.245	0.238

Under a strict matching criterion of ±0 days in the 1-day data, precision was relatively low at 0.096, while recall reached 0.513. This yielded an F1 score of 0.162, indicating that although BOCPD was able to detect a substantial portion of actual psychological shifts (moderate recall), many of the detected points were false positives, with low precision.

However, when a time tolerance of ±1 day was introduced, precision significantly improved to 0.367 and recall increased to 0.994, resulting in a notably higher F1 score of 0.536. This suggests that BOCPD is more effective when slight temporal offsets between behavioral and psychological changes are permitted. In practice, such flexibility is valuable because behavioral cues may not align perfectly with the moment learners report psychological distress.

Similarly, in the 7-day-level data, precision and recall under a ±0-day window were 0.094 and 0.179, respectively, yielding a low F1 score of 0.123. However, with a broader ±7-day range, precision improved to 0.243 and recall increased to 0.465, resulting in a higher F1 score of 0.319. Although lower than the performance observed in the 1-day data, this still implies that BOCPD can offer meaningful signals, especially in contexts where broader behavioral trends across weeks may reflect shifts in learners' mental states.

Comparing the two missing-data handling strategies, the Zero-fill method generally yielded higher recall but tended to inflate false positives, while the Weighted method, which downweighed weeks with higher missingness, achieved slightly higher precision at the cost of recall. For example, in the 1-day ±0-day setting, Zero-fill achieved a recall of 0.513 but very low precision = 0.096, whereas Weighted reduced the recall to 0.224 but improved the precision to 0.109. One likely explanation for these results is that real-world educational log data contained many missing or sparse entries, especially for steps and stress, which are not directly tied to school curriculum activities. In the Zero-fill setting, these missing values were treated as zeros, potentially exaggerating activity fluctuations and inflating false positives. It is undeniable that this missing value, which was processed to 0, may have ultimately affected the accuracy. By contrast, the Weighted method mitigated this influence but reduced overall detection sensitivity.

These results indicate that even with multiple educational log data with many missing values, BOCPD can be applied by extending it with the latent variable model proposed by Romero-Medrano (2022) and multinomial sampling. Especially, these findings demonstrate that BOCPD has potential as a practical tool for detecting early signs of psychological change in learners, particularly when used with appropriate temporal tolerances that account for real-world variability in the timing of behavioral and emotional transitions.

4.2 RQ2. Predicting Learners' State of Mind using BOCPD

To address RQ2, we predicted the changes in state of mind using the BOCPD scores and run lengths extracted from logs by BOCPD as characteristics. For prediction, the logistic regression model was constructed, and we divided the data set into a training set and a test set (9:1). The prediction model was trained and evaluated using data resampled on a daily and weekly basis, like RQ1. Prediction accuracy was evaluated using Accuracy and Area Under the Curve (AUC) (Table 3). In addition to reporting Accuracy and AUC values, we visualized the prediction performance using Receiver Operating Characteristic (ROC) curves (Figure 3). The ROC curve enables a more detailed examination of the trade-off between true positive and false positive rates across different thresholds, complementing the scalar AUC score.

Table 3 Result of prediction of state of mind using BOCPD

Resampling Unit	Method	Accuracy	AUC
1 day	Zero-fill	0.858	0.637
	Weighted	0.856	0.591
7 days	Zero-fill	0.867	0.697
	Weighted	0.832	0.611

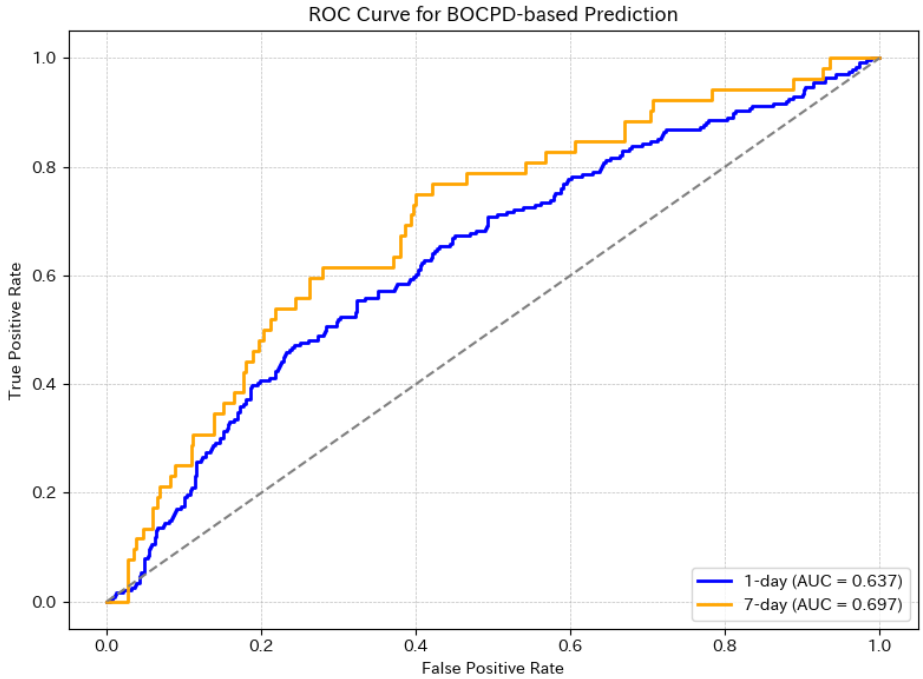


Figure 3 ROC curve for BOCPD-based prediction (Zero-fill method).

The results showed that the prediction model using change points extracted from learning and lifestyle logs as features using BOCPD achieved an accuracy of 0.858 and an AUC of 0.637 for daily data with Zero-fill, and an accuracy of 0.867 and an AUC of 0.697 for weekly data with Zero-fill. The Weighted method showed slightly lower AUC values (0.591 for daily and 0.611 for weekly), indicating that the Zero-fill method provided more stable predictive performance in this setting.

These results suggest that change point information obtained using BOCPD may have a certain validity for predicting learners' mental states. From the perspective of AUC, the prediction performance was higher for 7-day data than for 1-day data, suggesting that weekly changes in behavior and lifestyle rhythms may be more closely related to psychological states than daily fluctuations.

On the other hand, the AUC was only around 0.7 at its highest. While this result indicates that learners' state of mind can be predicted with a certain degree of accuracy from log data, it cannot be said that the predictions are highly accurate. This reaffirms that behavioral logs alone cannot fully explain psychological changes, as psychological state changes are complexly influenced by factors other than behavior (such as family environment, interpersonal relationships, sleep, and personality traits).

Nevertheless, the fact that this method can extract certain signs of internal changes in learners that are difficult for teachers to observe daily from log data suggests that it has potential for early support and preventive intervention in educational settings. In the future, combining this method with other predictive models or integrating multiple data sources (multimodal analysis) is expected to enable more accurate predictions and practical support.

In summary, applying BOCPD to educational log data has shown that it is possible to predict learners' mental health conditions to some extent. However, challenges remain in achieving higher accuracy.

5. Conclusion and Future Work

In this study, we investigated the applicability and potential of predicting learners' state of mind by applying BOCPD to cross-contextual educational log data. The results demonstrated that BOCPD should be applied to educational log data by adding extensions such as latent variable models, and that it is possible to predict learners' mental health conditions. This method has the potential to contribute to early support and preventive interventions in Guidance and Counseling settings, as it can extract signs of internal changes in learners that are difficult for teachers to observe daily from log data. The novelty of the study lies in applying BOCPD methods to real-world data with random missing values, commonly seen in the medical field, and in demonstrating their applicability to educational logs by revealing a certain level of predictability from their change points.

There are two points of discussion for future work. The first concerns the AUC accuracy, and the results may have been affected by missing data. According to Romero-Medrano (2022), the structure of missing data is complex, and non-random missing (MNAR) caused by user behavior and device operation is more likely to affect model accuracy than random missing (MCAR). In this study, there was a significant amount of missing data, particularly in steps taken and stress, and several processing steps were applied to address this. As a result, while the consistency rate of change points was higher with daily resampling, the prediction accuracy of the state of mind was higher with weekly resampling, leading to inconsistent results. Educational log data collected in the real world is prone to missing data for a variety of reasons. Therefore, a more robust justification for the imputation strategy is required. For example, future work should explore and compare alternative missing-data handling methods, including more sophisticated imputation techniques or approaches that explicitly model missingness rather than relying on simple imputation. A sensitivity analysis demonstrating the impact of different imputation strategies on the results would also significantly strengthen the findings.

The second point concerns the applicability of this approach in educational settings. To date, even data-driven frameworks such as MTSS have not been able to predict learners' states of mind from learning log data. The findings of this study could be used, for example, to alert teachers through an early warning system integrated into a dashboard, or as an additional resource for assessing school climate. Predicting changes in mental health conditions is an important process that allows teachers to consider the background and development of learners, and it also provides an important opportunity for communication with learners in G&C. Appropriate intervention can help learners overcome mental health crises, which in turn fosters restorative emotional experiences. In this regard, BOCPD is particularly promising compared to existing CPD methods. Further consideration is needed regarding the amount and context of data that can be applied, but BOCPD holds promise for interventions in G&C.

Acknowledgements

This work was partly supported by Council for Science, 3rd SIP JPJ012347 and JSPS KAKENHI Grant Number 23H00505.

References

Adams, R. P., & MacKay, D. J. (2007). Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*.

Atake, J., Hsu, C.-Y., Horikoshi, I., & Ogata, H. (2024). Extraction of important characteristics for data-informed guidance and counseling from daily usage log data. *International Conference on Computers in Education (ICCE)*. https://doi.org/10.58459/icce.2024.4941

Eubanks-Carter, C., Gorman, B. S., & Muran, J. C. (2012). Quantitative naturalistic methods for detecting change points in psychotherapy research: An illustration with alliance ruptures. *Psychotherapy Research*, *22*(6), 621-637.

Green, J. G., Keenan, J. K., Guzmán, J., Vinnes, S., Holt, M., & Comer, J. S. (2017). Teacher perspectives on indicators of adolescent social and emotional problems. *Evidence-based practice in child and adolescent mental health*, 2(2), 96–110. https://doi.org/10.1080/23794925.2017.1313099

Ito, H., Chang-Leung, C., & Poudyal, H. (2023). Inclusion of learners with developmental disabilities in Japan: *Barriers and promising practices in primary and secondary education. Asia Pacific Education Review*, 24, 415–431. https://doi.org/10.1007/s12564-022-09763-8

Kim, J., Kim, D. G., & Kamphaus, R. (2022). Early detection of mental health through universal screening at schools. *Georgia Educational Researcher*, 19(1), 62.

Li, H., Majumdar, R., Chen, M.-R. A., Yang, Y., & Ogata, H. (2021). Analysis of self-directed learning ability, reading outcomes, and personalized planning behavior for self-directed extensive reading. *Interactive Learning Environments*, 31(6), 3613–3632.

Li, K.C., Wong, B.T.M., Chan, H.T. (2024). Prediction of At-Risk Learners Using Learning Analytics: A Literature Review. In: Cheung, S.K.S., Wang, F.L., Paoprasert, N., Charnsethikul, P., Li, K.C., Phusavat, K. (eds) Technology in Education. *Innovative Practices for the New Normal. ICTE 2023. Communications in Computer and Information Science*, vol 1974. Springer, Singapore. https://doi.org/10.1007/978-981-99-8255-4_11

Miller, F. G., Patwa, S. S., & Chafouleas, S. M. (2014). Using Direct Behavior Rating–Single Item Scales to assess learner behavior within multi-tiered systems of support. *Journal of Special Education Leadership*, 27(2), 76–85.

Ministry of Education, Culture, Sports, Science and Technology. (2022). *Summary of Guidance and Counseling* [in Japanese]. Retrieved from https://www.mext.go.jp/content/20230220-mxt_jidou01-000024699201-1.pdf

Nakamura, K., Ishihara, M., Horikoshi, I., & Ogata, H. (2024). Uncovering insights from big data: change point detection of classroom engagement. *Smart Learning Environments*, 11(1), 31.

Naraswari, I. A. M. D., Dantes, N., Suarni, N. K., Gading, I. K., & Suranata, K. (2024). Solution-focused brief counseling to improve learners' social-emotional skills and psychological wellbeing. Jurnal EDUCATIO: *Jurnal Pendidikan Indonesia*, 10(1), 106-113. https://doi.org/10.29210/1202423798

Nitz, J., Brack, F., Hertel, S., Krull, J., Stephan, H., Hennemann, T., & Hanisch, C. (2023). Multi-tiered systems of support with focus on behavioral modification in elementary schools: A systematic review. *Heliyon*, 9(7), e17506. https://doi.org/10.1016/j.heliyon.2023.e17506

Ogata, H., Majumdar, R., & Flanagan, B. (2023). Learning in the digital age: power of shared learning logs to support sustainable educational practices. *IEICE TRANSACTIONS on Information and Systems*, 106(2), 101-109. https://doi.org/10.1587/transinf.2022ETI0002

Prasad, S. B. L. (2020). A review on control charts in learning performance of students. International *Journal of Innovative Research in Technology*, 7(1),1049-1051, https://ijirt.org/Article?manuscript=149905

Romero Medrano, L. (2022). *Change-point detection methods for behavioral shift recognition in mental healthcare* [Doctoral dissertation, Universidad Carlos III de Madrid]. e-Archivo. https://hdl.handle.net/10016/36642

Shimada, A., Taniguchi, Y., Okubo, F., Konomi, S. I., & Ogata, H. (2018, March). Online change detection for monitoring individual student behavior via clickstream data on E-book system. In J. Kay & R. Luckin (Eds.), *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*. 446–450. ACM. https://doi.org/10.1145/3170358.3170432

Truong, C., Oudre, L., & Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, *167*, 107299.

Empowering, Not Replacing: Using Generative AI to Coach Educators in Providing Effective Feedback

Antonette SHIBANI[a*], Jayakrishnan WARRIEM[b], Lisa-Angelique LIM[c], Gnana BHARATHY[d] & Arup CHATTOPADHYAY[e]
[a]*Transdisciplinary School, University of Technology Sydney, Australia*
[b]*NPTEL, IIT Madras, India*
[c]*Connected Intelligence Centre, University of Technology Sydney, Australia*
[d]*Faculty of Engineering & Information Technology, University of Technology Sydney, Australia*
[e]*BS Program, IIT Madras, India*
*Antonette.shibani@uts.edu.au

Abstract: As generative AI (GenAI) technologies become increasingly integrated into educational contexts, the importance of supporting rather than replacing educators with AI is gaining prominence. This study introduces Feedback Tutor, a GenAI-powered tool designed to enhance teacher feedback literacy by coaching educators in crafting effective feedback, while preserving their agency. It explores how human-AI collaboration can empower educators through a teacher professional development intervention using Feedback Tutor. Eleven educators from two institutions participated in a workshop and follow-up interviews, with qualitative data analyzed thematically. Four key themes emerged: (1) *AI-supported professional growth*, where educators valued the tool's coaching capabilities for improving clarity and feedback quality; (2) *Human-AI Negotiation and Agency*, highlighting the tool's role as a collaborative partner rather than a replacement; (3) *Socio-Technical Integration*, emphasizing the importance of aligning AI tools with existing educational workflows and infrastructure; and (4) *Ethical and Pedagogical Considerations*, underscoring the need for balanced feedback that supports student autonomy and learning. Findings suggest that GenAI tools like Feedback Tutor can significantly enhance educators' feedback practices and professional growth when designed with pedagogical intent and ethical sensitivity. However, challenges such as lack of content ingestion, and usability constraints must be addressed to optimize adoption and impact. This study contributes to the growing discourse on AI in education by demonstrating how thoughtfully designed AI tools can augment, rather than diminish, the human elements of teaching and learning.

Keywords: educator, GenAI, feedback, feedback tutor, teacher feedback literacy, empowering, artificial intelligence, teacher professional development

1. Introduction

Generative AI (GenAI) has been touted as a powerful tool to transform educational practices. Across various contexts, GenAI has demonstrated the potential to enhance student learning and assist educators in the teaching process. This includes the provision of real-time feedback for personalized support that can help improve students' writing and save teachers' time by allowing them to focus on other aspects of teaching (Kasneci et al., 2023). Indeed, feedback is an area where there is an identified need for better support so students can reflect on their academic performance and improve over time. Sufficiently detailed, high-quality comments that are usable and tailored to students work can help them improve (Dawson et al., 2019), but is time-consuming for the educator to provide. Many students report issues on the feedback they receive, which include the lack of specificity, delayed delivery, limited guidance on how to apply it in future tasks, difficulty in interpretation, and potentially affecting students' self-perception and confidence (Carless & Boud, 2018; Winstone et al., 2017).

While it is important for students to learn to make good sense of feedback provided and act upon it (Student Feedback Literacy - (Carless & Boud, 2018)), and even, develop specific skills to critically engage with automated feedback (Automated feedback literacy - (Shibani et al., 2022)), it is equally important to support *teachers' feedback literacy* due to their intertwined nature (Carless & Winstone, 2020). This is because, for students to get value out of feedback, teachers must provide effective feedback. While there is no single definition of good feedback practice, commonly agreed upon principles guide them (Nicol & Macfarlane-Dick, 2006). Novice educators are often learning to provide feedback, have limited pedagogical training, and can miss many components of such effective feedback. This can lead to concerns navigating the design, relational, and pragmatic aspects of teacher feedback literacy (Carless & Winstone, 2020). Supporting educators to learn to provide good feedback for students is an important step towards fostering students' feedback literacy.

Technology mediated feedback has seen a steady growth in literature, allowing feedback to be a continuous, real-time part of student work. Automated systems and learning analytics tools have demonstrated impact at scale due to their ability to provide feedback to large cohorts at a personalized level (Knight, Shibani, et al., 2020; Pardo et al., 2019). With advances from GenAI, it is no surprise that feedback was a viable candidate for AI to offer support. Recent works explore the use of GenAI to provide feedback on many types of tasks, including on writing (Lee et al., 2024), computer programming (Balse et al., 2023), and content creation (Pozdniakov et al., 2025) often in par with humans, with tools such as Cogniti offering AI feedback tailored to assignment rubric set up by educators in higher education (Huynh et al., 2024). However, it also raises the question of whether AI will support educators or de-skill them in the future as they turn towards AI for writing their feedback messages, without sufficient experience and skill development of their own.

Teacher Professional Development (TPD) for AI thus becomes increasingly important as educators are now required to work alongside hybrid intelligent systems (Molenaar, 2022) and the TPD should help in getting educators AI ready (Luckin et al., 2022). Long and Magerko's (2020) framework for AI literacy emphasizes that educators need opportunities to develop critical evaluation skills and collaborative approaches to working alongside AI tools. Similarly, the Ethical AI readiness framework (Luckin et al., 2022) underscores the importance of TPD that empowers teachers to make informed decisions about AI implementation rather than merely training them as end-users. These approaches recognize that educators must maintain agency in human-AI partnerships, positioning them as active designers and critical implementers rather than passive consumers of AI technologies.

The integration of GenAI thus presents a unique opportunity to enhance, rather than replace, the critical role of educators in providing feedback. While concerns about AI potentially diminishing the human element in education are valid, this technology can be harnessed to empower educators by augmenting their ability to deliver more effective, rubric-aligned feedback, ensuring that students receive the support they need to improve their work. Our study introduces a GenAI-support tool, 'Feedback Tutor', along with an intervention that helps educators improve their written feedback by providing guiding feedback on their written feedback, while also empowering their own knowledge with principles of good feedback practice and human-AI collaboration. We explore the research question: *What affordances of Feedback Tutor can empower educators to improve their feedback practices, and influence changes?*

2. Background

2.1 Feedback

Feedback has long been acknowledged as a critical factor in improving students' learning and performance. In this paper, we define feedback as a dialogical process, "whereby learners make sense of information from various sources and use it to enhance their work or learning strategies" (Carless & Boud, 2018) (p.1). Effective feedback needs to follow established principles; namely: 1) specify clear performance standards; 2) facilitate self-assessment; 3) communicate high-quality information for students to bridge learning gaps; 4)

encourage dialogue; 5) support motivation and foster self-efficacy; 6) provide opportunities to close the gap between current and expected performance (Nicol & Macfarlane-Dick, 2006). To this end, feedback needs to be timely, personalized, actionable and directed at students' self-regulated learning, not just the outcome or performance (Hattie & Timperley, 2007). The provision of such effective feedback, however, remains challenging in higher education contexts. Scaling personalized feedback to accommodate large cohorts presents a significant burden on faculty members who must simultaneously balance substantial teaching responsibilities with administrative obligations and research requirements (Henderson et al., 2019; Sembey et al., 2024). Consequentially, empirical research points to students' low engagement and satisfaction with their feedback (e.g., (Winstone et al., 2017)).

Furthermore, given the process-oriented nature of feedback, feedback literacy has become increasingly recognized as important – both teacher feedback literacy and student feedback literacy, which are intertwined (Carless & Winstone, 2020), as both teachers and students are jointly responsible for feedback to be effective. While a number of frameworks exist, student feedback literacy broadly refers to students' understandings and appreciation of feedback, their ability to evaluate the work of themselves as well as of others based on feedback, their ability to manage emotional responses to feedback, and drawing on a range of strategies to act on feedback (Carless & Boud, 2018). On the other hand, teacher feedback literacy refers to the "knowledge, expertise, and dispositions to design feedback processes in ways which enable student uptake of feedback and seed the development of student feedback literacy" (Carless & Winstone, 2020) (p. 4). Based on previous work involving interviews and focus groups with teachers about their assessment and feedback practices, Boud & Dawson (2023) further identified what feedback literate teachers actually do with respect to three distinct levels. The *macro* level relates to how feedback is built into programme design and development; the *meso* level relates to how feedback is purposely designed in specific modules or units; while the *micro* level relates distinctly to feedback practices involved in students' work. In this study, we focus on teacher feedback literacy at the micro level, with a view to developing teacher feedback literacy, because teacher feedback literacy is fundamental for fostering student feedback literacy within educational contexts.

2.2 Human-AI interaction paradigm

The humanistic approach to AI in education advocates for AI being used to facilitate higher order thinking, human interaction, and human values rather than diminishing them (Carvalho et al., 2022). This implies that the goal should be to build hybrid systems that augment rather than replace human intelligence, leveraging human strength and compensating for its weaknesses (Molenaar, 2022). Such hybrid collaborations will cover a spectrum of use-cases with varying contributions and role of AI, with the extremes of AI being the complete driver to AI being a mere assistant (Banihashem et al., 2025).

There have been numerous examples of AI systems being used for supporting teachers with data driven decisions like dashboards, providing recommendations on instructional strategies and curriculum, and assessment support. With GenAI tools maturing, there is an increase in use cases of conversational AI systems (such as chat bots) that support teaching-learning scenarios including providing direct feedback to students and assisting in teacher professional development using a variety of pedagogical techniques (Tan et al., 2025).

These technological developments directly inform the evolving needs within TPD, where empowering educators to effectively collaborate with AI has become essential. Two key concepts have emerged in this context: AI Literacy and AI readiness. Long and Magerko (2020) had defined AI literacy as a competency required for individuals to critically evaluate AI technology, communicate and collaborate with AI systems, and use AI as a tool across professional and personal spaces. Complementarily, AI readiness refers to the ability of the teacher to understand, in non-technical terms, the capabilities of AI - enabling them to make informed decisions about the use and procurement of AI in educational settings (Luckin et al., 2022). TPD will be essential to fundamentally address these competencies and ensure the effectiveness and sustainability of training benefits.

To systematically develop these competencies among educators, several frameworks have been proposed that integrate the theoretical understanding about TPD along with its practical implementation strategies. The i-TPACK framework (Dogan et al., 2025) extends the TPACK model to incorporate AI specific knowledge domains. The ETHICAL AI readiness framework (Luckin et al., 2022) provides a 7-step structure for building an AI training and apply AI readiness thinking. In the specific context of AI supported feedback, the pedagogical framework for hybrid intelligent feedback by Banihashem et al (2025) outlines a structured process for integrating human and GenAI contributions in providing feedback.

A review of the research highlights a lack of attention to the critical role of teachers and their professional development needs in the context of AI integration, especially in enhancing teachers' AI literacy (Tan et al., 2025). This gap becomes more pronounced when considering hybrid Human-AI systems, that remain in early stages of research (Molenaar, 2022). There is a clear need for pedagogical guidelines for both AI systems and educators to follow while implementing AI feedback in their teaching-learning settings. To align with the humanistic approach, it is important to explore how roles can dynamically shift between human and AI based on task complexity, what educators find beneficial in these contexts, and identify areas where human strengths can be meaningfully leveraged - forming the foundation of our work.

2.3 Contextualised interventions

Automated tools and AI technologies can help enhance teaching and learning processes by offering personalized support at scale to learners and educators. However, their adoption is often determined by factors beyond just tool efficiency. For maximum impact, the technical aspects must be well aligned with the pedagogical aspects (Knight, Gibson, et al., 2020), which include how the tool is implemented in authentic classrooms, and how it is tied to learning outcomes and assessment practices of existing curricula for student facing tools (Shibani et al., 2019). For TPD, educators require support for adopting technologies in their teaching practices (Nelson et al., 2019). Importantly, the effective use of automated feedback tools requires teachers to demonstrate teacher feedback literacy competencies, including those at the micro-level (Buckingham Shum et al., 2023). Thus, the adoption of teaching and learning technologies is not just a technical endeavour, but a socio-technical one.

A review of intervention studies found that technologies for specific instructional purposes were more frequently adopted than general purpose technologies by educators (Wu et al., 2013). Tools and practices contextualized for specific settings also provide alignment between the technical and pedagogical aims (Shibani et al., 2019) and can be extended to the affordances of GenAI. While general purpose GenAI tools such as ChatGPT can assist educators in optimising standard responsibilities (Hashem et al., 2024), interventions designed using GenAI capabilities for specific purposes can offer targeted support. Our work introduces a specialised GenAI powered tool 'Feedback Tutor' and intervention that can empower educators by supporting effective feedback practices and fostering teacher feedback literacy. Such interventions can offer professional development opportunities for educators by not merely functioning as tools but also acting as catalysts for reflective teaching and improved teaching practices. In the following section, we introduce the design of Feedback Tutor, its underlying technology, and implementation of the tool in the intervention.

3. Feedback Tutor Tool and Intervention for Teacher Feedback Literacy

3.1 Tool features

Feedback Tutor was designed to have an intuitive chat interface where educators can type in their feedback and receive suggestions for improvement. They can select an option from two types of assignment at the start: A. Writing assignment and B. Programming assignment, reflective of the most common types of feedback provided in the two institutional contexts. The tool first asks the educator to describe the context of the assignment including rubric details if available. Then educators are asked to type their draft feedback. The tool responds by suggesting specific improvements for the written feedback, also including code if

needed for programming assignments. Importantly, in addition to suggesting revisions, the tool provides coaching to the educator, empowering them to write better feedback by themselves in the future. Coaching includes the rationale for the suggestions and an explanation of why it would be helpful to students in improving their work, grounded in effective feedback principles from literature (Nicol & Macfarlane-Dick, 2006). Each chat also has features to regenerate feedback if needed, like/ dislike, and users can download a copy of the feedback as a PDF file. Users can initiate any number of chat threads as needed, which are saved in their log for later retrieval, as required. Figure 1 provides a screenshot showcasing a sample chat in Feedback Tutor interface. Additionally, the tool has functionality for admin users to view analytics and logs of the tool usage in a dashboard (these admin features are not intended for educators and hence are not the focus of the current study).

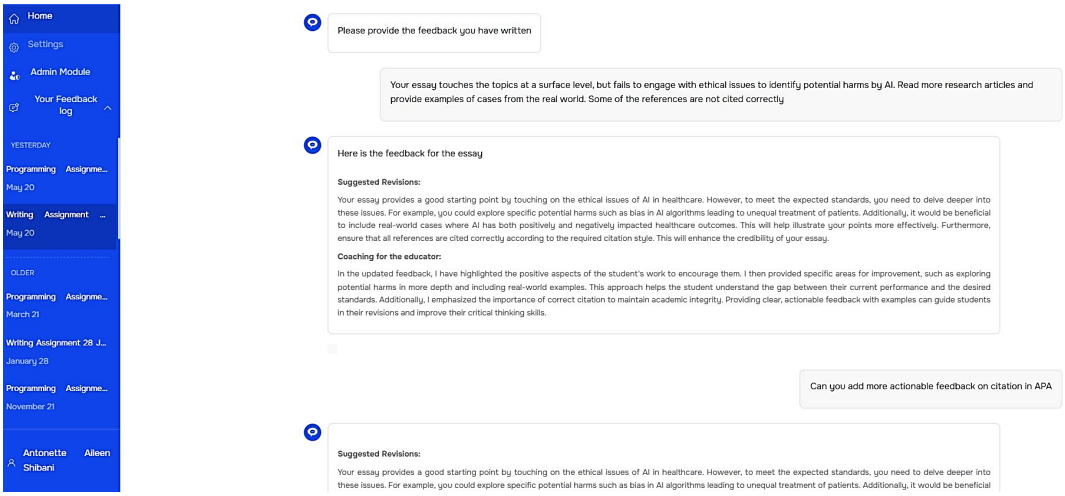


Figure 1. Screenshot of the Feedback Tutor interface for educators with a sample chat

3.2 Technical implementation

Feedback Tutor was implemented using Microsoft Azure tools and could be accessed via a URL[1]. It was developed as a React app and used the GPT-4 model for providing feedback which was cost-effective while delivering good results. The tool was deployed using Docker on Azure cloud services. Iterative testing and changes were made to the prompts provided to the large language model (LLM) to guide its feedback which specifically included instructions to preserve the original voice of the feedback creator to maintain authenticity and to provide coaching in addition to suggesting improvements to feedback. The full source code of the tool has been made available as an open-source resource on Github[2]. This includes the LLM prompts used for the writing and programming assignment feedback sections, and the detailed technical architecture of the tool.

3.3 Intervention design

In line with the socio-technical nature of AI technologies, the intervention was designed and delivered as a two-hour online professional development workshop for educators from two institutions located in different countries. The workshop consisted of introductory topics including the definitions of feedback (from academic and practical viewpoints), summative and formative feedback mechanisms, strategies to deliver them effectively such as the sandwich approach, and knowledge sharing activities among the participants using short polls designed in Mentimeter[3], facilitated by the research team in the first part. Participants were introduced

[1] https://feedbacktutorweb.azurewebsites.net
[2] https://github.com/jkmpod/feedback-tutor
[3] https://www.mentimeter.com/

to Feedback Tutor's main features, and learned how the tool could be used to improve their feedback in the second part. All participants were previously instructed to bring 2-3 samples of draft feedback they had written to try out the tool. They were provided with instructions for stimulated recall and recording as they explored the tool independently in the last part.

4. Methodology

The study employs qualitative methods to understand educators' perspectives of the Feedback Tutor tool and intervention. A total of eleven participants volunteered to participate in the online workshop and in the follow-up interview and were provided with a monetary incentive for contributing to the study. The follow-up one-on-one stimulated recall interview was conducted with each participant within 2-4 weeks after the workshop, which was guided by the recording from their trial of the tool. Each interview lasted between 45 minutes and 1 hour. Analysis for the current study uses interview data from the eleven participants - P1 to P8 from Institution 1, and P9 to P11 from Institution 2. While all of them had provided feedback and assessment in their institutional contexts, they had varied levels of teaching experience.

The transcripts were cleaned and anonymized for inductive thematic analysis. The analysis was performed by two large language model (LLM) based tools and three human coders using a hybrid approach (Barany et al., 2024). The following prompt was used in Google NotebookLM and Microsoft Co-pilot (with institutional access for secure data analysis): *"Analyse transcript.docx - do a thematic analysis of the interview with P[]. Extract the themes that emerge and also provide the supporting quotes (verbatim) for the same. Where multiple quotes exist, provide all of them".* All LLM outputs were verified for accuracy and evaluated by two researchers. They were further compared with a third researcher's own analysis to add missing quotes and re-word headings. Neither the researchers nor the LLM adopted a positionality; instead, they aimed to present all key themes that emerged from the data.

5. Findings and discussion

To answer our research question, we examined how the affordances of Human-AI collaborative tools such as Feedback Tutor can empower educators by enhancing their ability to deliver effective feedback. Our thematic analysis identified several factors, which were organized into four main themes that influence educator feedback practices when mediated through the given GenAI tool.

Theme 1: AI-supported professional growth

The theme around educators valuing professional growth using AI support emerged prominently, alongside other beneficial features of the tool. Participants consistently valued the "Coaching for Educators" aspect embedded within the tool, emphasizing its significant role in improving their feedback practices. This finding aligns with the work of Graßmann & Schermuly (2021), who emphasize that AI coaching is most effective when used in conjunction with human oversight, allowing feedback to be more precisely tailored to individual educational needs.

P3 notably highlighted this as impactful, stating, *"the coaching for educator part… helped me improve my feedback clearly and concisely"*. This capability enabled educators to explicitly visualize and refine feedback principles, thereby supporting their professional growth. P8 appreciated the tool's ability to highlight their own missteps, remarking, *"tells what I did wrong. So as an educator I also did something wrong"*. This meta-feedback aspect was consistently highlighted as a distinctive feature – setting the tool apart from general LLMs, which are often used to provide direct evaluations and feedback rather than supporting educator reflection.

Educators found elements such as language enhancement and error identification particularly beneficial. Participants appreciated controlled augmentations - for example, *"did some wording changes to make it more coherent"* (P10), *"… grammatical error .., missing something …"* (P5) - highlighting the utility of the tool in improving clarity and coherence in feedback delivery (Graßmann & Schermuly, 2021).

Educators also appreciated how the tool improved the structure and organization of their feedback, making it more accessible and digestible for students. For instance, P2 noted, *"… the flow of the feedback to make it more... understandable to the students"*. Similarly, P9 highlighted the benefit of structured formatting compared to typical paragraph style used by instructors: *"…. the feedback [was] … divided between themes …, so it's easier to read…. It streamlines your text [whereas] … I tend to be a very verbose… summarizes better what I wanted to say in the feedback*".

In addition, participants valued the tool's ability to provide clear examples, which made their feedback more concrete and actionable. As stated by P2, *"It actually gave examples in a clear manner, …"* and P5 *"… it gives more example, … giving correct code also"*.

However, some educators raised concerns about potential overreach by AI, as the tool occasionally provided overly explicit solutions in its feedback, inadvertently limiting students' opportunities for critical thinking (P4). Thus, while clearly beneficial, careful balance is necessary to ensure educators' agency and pedagogical intentions remain intact in AI-assisted feedback. Indeed, the ability to recognize and navigate these concerns is a key part of developing teacher feedback literacy (Buckingham Shum et al., 2023) and AI literacy (Long & Magerko, 2020). The intervention supported this development by enabling educators to learn how to engage effectively in hybrid human-AI settings.

Theme 2: Human-AI Negotiation and Agency

The Human-AI Negotiation and Agency theme reflected the dynamic partnership educators desired to have with AI – favoring active collaboration over passive consumption of AI-generated feedback. Educators appreciated the tool's ability to enhance their quality of feedback without diminishing their authority or personal pedagogical style, a design consideration explicitly programmed in the LLM prompt of the tool as discussed in Section 3.1. As P8 observed, *"The tool evolves my way to provide feedback… without replacing my role".* This sentiment was echoed by many participants, who viewed (and would like to view) the technology as a collaborator rather than a substitute (Graßmann & Schermuly, 2021).

However, concerns about the extent of augmentation were also expressed. For instance, P2 noted "*Sometimes … I felt that it actually kind of removed … my personal elements to the feedback",* while P11 remarked *"…sometimes adding some extra word[s] that was not expected. …".*

Several participants described engaging in an active, iterative process of evaluating and selectively adopting AI-generated suggestions. For example, P8 shared, *"I'll regenerate two three times and I'll pick the best one"* and P11 added, *"… I have to write 4 to 5 words. And, Gen AI will provide between 30, 40 words in a 3, 4 sentence that could be".* This selective engagement illustrates the educators' desire for control and intentionality in shaping AI-augmented feedback. It also highlights a new affordance offered by generative AI – the ability to engage in a dialogic process - effectively utilized by educators to co-construct meaningful feedback relevant to their contexts.

Nonetheless, challenges remain – particularly regarding the lack of AI's conversational memory, particularly the absence of session memory or iterative dialogue functionalities, which hindered continuity in interactions and forced repetitive engagement, leading to frustrations that P1 articulated as: "*Lack of conversational memory forces repeated manual interactions; it's cumbersome*". These frustrations point to a critical need for improved conversational continuity to support more seamless and effective human-AI collaboration, which was later developed and offered as a feature of the tool (explained in the next section).

Theme 3: Socio-Technical Integration

The need for Socio-Technical Integration was consistently highlighted, emphasizing that the effectiveness of GenAI tools depends significantly on their compatibility with existing educational workflows, practices and technological infrastructures. The potential for integration with Learning Management Systems (LMS) (P9) and platforms like Discourse (P1, P3, P5) was seen as a critical facilitator for adoption and usability, promising substantial time savings - an aspect also emphasized in Hashem et al (2024). The potential for improved workflow through alignment with pre-existing systems supports broader arguments by Knight

et al. (2020), who asserted that successful technology adoption in education hinges on its integration into familiar pedagogical settings.

However, educators raised concerns over limitations related to content ingestion in the form of documents or images, retention, and the tool's ability to remember previous interactions - all perceived as obstacles to usability. The inability to upload student work limited the tool's capacity to tailor feedback to specific assignments. Similarly, the absence of conversation history prevented users from continuing discussions on related topics. For instance, P1 noted, *"The main part is memory. So, we don't have to copy paste when we want any improvement in chat GPT. Here, we have to do that"*. P10 wanted to *"ask more questions"* after getting a revision and wished for *"the ability to continue the chat"* to refine irrelevant content. While at the time of this trial run, the memory feature was not available. However, to respond to multiple educators' concerns regarding the lack of conversation history, this feature is implemented in the current version of Feedback Tutor. Now, users can extend the conversation further by asking follow-up questions on the feedback, which are answered by the LLM's knowledge base.

P10 also identified a main limitation as the inability to *"continue same chat for the current feedbacks belonging to same assessment"* which meant having to *"start another chat to test the other feedback".* This process of going back to "home" for each new feedback was *"not really user friendly"*. Participants explicitly noted the inconvenience caused by interface designs requiring extensive tab-switching, underscoring that a more seamless integration would significantly enhance educators' workflow efficiency, a key aspect of effective edtech integration (Nelson et al., 2019).

Theme 4: Ethical and Pedagogical Considerations

The analysis revealed an essential dimension concerning Ethical and Pedagogical Considerations in the use of GenAI tools for educators. Educators appreciated the tool's capacity to promote structured, clear, and pedagogically sound feedback. This kind of feedback was seen as fostering student self-regulation and agency– a balanced approach to feedback aligned with effective formative feedback principles (Nicol & Macfarlane-Dick, 2006). As P4 reflected, *"The suggestions improved how I guide students without giving direct answers"*.

However, concerns were raised about the accuracy and validity of AI-generated feedback. For example, P6 questioned the reliability of outputs asking, *"How are we validating that the responses generated are always correct?"* Such concerns point to the need for transparent validation mechanisms and ongoing oversight when integrating GenAI tools into pedagogical contexts given the inconsistent nature of LLM outputs.

Participants also raised ethical concerns around the level of detail in AI-generated feedback, fearing it could undermine student autonomy and reduce their engagement with the learning material. They expressed caution about overly prescriptive responses, which may discourage independent thinking. As P4 noted, *"The complete code snippet is given…. …but we don't tell them how to do it"*, illustrating the risk of bypassing the learning process in favor of efficiency.

Additionally, a few participants discussed constraints linked to content uploads and contextual continuity – issues connected to broader socio-technical and ethical considerations of GenAI tools. While the ability to enhance conversational context by uploading assignments was widely recognized as a desirable feature, its absence was not due to technical shortcomings, but rather deliberate design decisions based on two key factors: First, ethics approval did not permit the uploading of student work, as students had not consented to having their assignments or work shared with AI systems. Secondly, the tool was intentionally designed to encourage educators to develop their capacity to write feedback independent of AI support, reinforcing the research objective of empowering their feedback practices through TPD. Timesaving was not the tool's primary objective. This underscores a fundamental tension between usability demands and ethical boundaries and highlights a key insight: efficiency using GenAI tools must not come at the cost of critical engagement or educational integrity.

6. Conclusion

In this study, we presented Feedback Tutor, a GenAI-powered tool supporting TPD intervention designed to develop teachers' feedback literacy. The tool and intervention were specifically created to empower educators with the autonomy to write independent feedback while being coached to enhance their practices, evaluated by 11 educators from two institutional contexts. Findings indicate that human-AI collaboration in feedback practices can positively transform educational pedagogy, support professional learning and improve feedback quality. However, this positive impact depends on addressing key usability issues, maintaining a balance of educator autonomy, ensuring robust socio-technical integration, and articulating the ethical implications of AI-enhanced practices. Future work will explore how teacher agency and reflections of educators evolved after the initial exposure by engaging with the tool over time. It will also unpack in more depth the tensions between efficiency and ethical design of GenAI tools, and trade-offs influencing educator perspectives on use and usefulness of tools. We hope such thoughtful integration of AI can support the development of educator expertise while augmenting, rather than replacing, their professional judgment.

Acknowledgements

We would like to acknowledge the UTS Key Technology Partner (KTP) Funding Programme 2024 for funding this research collaboration between UTS and IITM. Thanks also to the team at Jumpstart Ninja for their rapid tool development and deployment support.

References

Balse, R., Valaboju, B., Singhal, S., Warriem, J. M., & Prasad, P. (2023). Investigating the potential of gpt-3 in providing feedback for programming assessments. Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1,

Banihashem, S. K., Noroozi, O., Khosravi, H., Schunn, C. D., & Drachsler, H. (2025). Pedagogical framework for hybrid intelligent feedback, Innovations. *Education and Teaching International*, 1-17.

Barany, A., Nasiar, N., Porter, C., Zambrano, A. F., Andres, A. L., Bright, D., . . . Baker, R. S. (2024). ChatGPT for Education Research: Exploring the Potential of Large Language Models for Qualitative Codebook Development. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt, *Artificial Intelligence in Education* Cham.

Boud, D., & Dawson, P. (2023). What feedback literate teachers do: an empirically-derived competency framework. *Assessment & Evaluation in Higher Education*, *48*(2), 158-171.

Buckingham Shum, S., Lim, L.-A., Boud, D., Bearman, M., & Dawson, P. (2023). A comparative analysis of the skilled use of automated feedback tools through the lens of teacher feedback literacy. *International Journal of Educational Technology in Higher Education*, *20*(1), 40.

Carless, D., & Boud, D. (2018). The development of student feedback literacy: enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, *43*(8), 1315-1325. https://doi.org/10.1080/02602938.2018.1463354

Carless, D., & Winstone, N. (2020). Teacher feedback literacy and its interplay with student feedback literacy. *Teaching in Higher Education*, *28*(1), 150-163. https://doi.org/10.1080/13562517.2020.1782372

Carvalho, L., Martinez-Maldonaldo, R., Tsai, Y.-S., Markauskaite, L., & De Laat, M. (2022). How can we design for learning in an AI world? *Computers and Education: Artificial Intelligence*, *3*.

Dawson, P., Henderson, M., Mahoney, P., Phillips, M., Ryan, T., Boud, D., & Molloy, E. (2019). What makes for effective feedback: staff and student perspectives. *Assessment & Evaluation in Higher Education*, *44*(1), 25-36. https://doi.org/10.1080/02602938.2018.1467877

Dogan, S., Nalbantoglu, U. Y., Celik, I., & Dogan , N. A. (2025). Artificial intelligence professional development: a systematic review of TPACK, designs, and effects for teacher learning. *Professional Development in Education*, *51*(3), 519-546.

Graßmann, C., & Schermuly, C. C. (2021). Coaching with artificial intelligence: Concepts and capabilities. *Human Resource Development Review*, *20*(1), 106-126.

Hashem, R., Ali, N., El Zein, F., Fidalgo, P., & Khurma, O. A. (2024). AI to the rescue: Exploring the potential of ChatGPT as a teacher ally for workload relief and burnout prevention. *Research & Practice in Technology Enhanced Learning*, *19*.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research*, *77*(1), 81-112.

Henderson, M., Molloy, E., Ajjawi, R., & Boud, D. (2019). Designing feedback for impact. In *The impact of feedback in higher education: Improving assessment outcomes for learners* (pp. 267-285). Springer.

Huynh, M., Van Den Berg, F., Van Ogtrop, F., Pozza, L., & Kite, J. (2024). *How generative AI can make personalised feedback at scale more consistent and efficient*. The University of Sydney. Retrieved 10 May 2025 from https://educational-innovation.sydney.edu.au/teaching@sydney/how-generative-ai-can-make-personalised-feedback-at-scale-more-consistent-and-efficient/

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., . . . Hüllermeier, E. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274.

Knight, S., Gibson, A., & Shibani, A. (2020). Implementing learning analytics for learning impact: Taking tools to task. *The Internet and Higher Education*, *45*, 100729. https://doi.org/https://doi.org/10.1016/j.iheduc.2020.100729

Knight, S., Shibani, A., Abel, S., Gibson, A., Ryan, P., Sutton, N., . . . Buckingham Shum, S. (2020). AcaWriter: A Learning Analytics Tool for Formative Feedback on Academic Writing. *Journal of Writing Research*. https://doi.org/10.17239/jowr-2020.12.01.06

Lee, M., Gero, K. I., Chung, J. J. Y., Buckingham Shum, S., Raheja, V., Shen, H., . . . Siangliulue, P. (2024). *A Design Space for Intelligent and Interactive Writing Assistants* Proceedings of the CHI24 conference on human factors in computing systems, Honolulu, Hawai'i. arXiv:2403.14117.

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. Proceedings of the 2020 CHI conference on human factors in computing systems, Honolulu, Hawai'i, USA.

Luckin, R., Cukurova, M., Kent, C., & du Boulay, B. (2022). Empowering educators to be AI-ready. *Computers and Education: Artificial Intelligence*, *3*(22).

Molenaar, I. (2022). Towards hybrid human-AI learning technologies. *European Journal of Education*, *57*(4), 632-645.

Nelson, M. J., Voithofer, R., & Cheng, S.-L. (2019). Mediating factors that influence the technology integration practices of teacher educators. *Computers & Education*, *128*, 330-344.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in higher education*, *31*(2), 199-218.

Pardo, A., Jovanovic, J., Dawson, S., Gašević, D., & Mirriahi, N. (2019). Using learning analytics to scale the provision of personalised feedback. *British Journal of Educational Technology*, *50*(1), 128-138.

Pozdniakov, S., Brazil, J., Mohammadi, M., Dollinger, M., Sadiq, S., & Khosravi, H. (2025). AI-assisted co-creation: Bridging skill gaps in student-generated content. *Journal of Learning Analytics*, *12*(1), 129-151.

Sembey, R., Hoda, R., & Grundy, J. (2024). Emerging technologies in higher education assessment and feedback practices: A systematic literature review. *Journal of Systems and Software*, *211*, 111988.

Shibani, A., Knight, S., & Buckingham Shum, S. (2019). *Contextualizable Learning Analytics Design: A Generic Model, and Writing Analytics Evaluations* 9th International Conference on Learning Analytics and Knowledge, Tempe, Arizona.

Shibani, A., Knight, S., & Shum, S. B. (2022). Questioning learning analytics? Cultivating critical engagement as student automated feedback literacy. Proceedings of the 12th international conference on learning analytics & knowledge (LAK22), Online, USA. https://doi.org/10.1145/3506860.3506912.

Tan, X., Cheng, G., & Ling, M. H. (2025). Artificial intelligence in teaching and teacher professional development: A systematic review. *Computers and Education: Artificial Intelligence*, *8*.

Winstone, N. E., Nash, R. A., Rowntree, J., & Parker, M. (2017). 'It'd be useful, but I wouldn't use it': barriers to university students' feedback seeking and recipience. *Studies in higher education*, *42*(11), 2026-2041.

Wu, Y.-T., Hou, H.-T., Hwang, F.-K., Lee, M.-H., Lai, C.-H., Chiou, G.-L., . . . Chen, N.-S. (2013). A review of intervention studies on technology-assisted instruction from 2005-2010. *Journal of Educational Technology & Society*, *16*(3), 191-203.

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Pre-Digital to AI eras: Changing Role of a Teacher with Technology

Vinay YADAV[a*], Ishika ISHIKA[a] & DurgaPrasad KARNAM[a]
[a]*Indian Institute of Technology Bombay, Mumbai, India*
*24m1027@iitb.ac.in

Abstract: The role of teachers has undergone a profound transformation in response to the continuous advancement of emerging educational technologies. While innovations such as artificial intelligence, digital tools, and online learning platforms have significantly altered the educational landscape, teachers continue to occupy a central and indispensable position within the learning process. This study presents a systematic literature review aimed at analyzing the evolving roles and responsibilities, and instructional/pedagogical strategies of teachers across different technological eras. It also explores the challenges educators face in adapting to these shifts. The review identifies and categorizes these transformations into four distinct phases: (1) the Pre-Digital era, characterized by traditional didactic instruction; (2) the Broadcast era, marked by the introduction of mass media such as radio and television; (3) the Internet era, which introduced digital interactivity and online learning; and (4) the Artificial Intelligence era, where intelligent systems support adaptive, personalized learning. Findings reveal that rather than replacing teachers, technology functions as an enabler, enhancing their roles. Teachers transitioned from being *knowledge authorities* in the Pre-Digital era, to mediators and moderators in the Broadcast era, to *facilitators, designers, and co-learners* in the Internet era, and to orchestrators, ethical guides, and mentors in the AI era. The study also identifies shifts in instructional approaches across eras, emphasizing the increasing need for learner-centered, adaptive, and ethically grounded pedagogies. The paper further highlights key challenges, including digital literacy gaps, resistance to technology adoption, and ethical considerations in AI-enhanced education. This study contributes to the ongoing discourse on teacher professional roles, pedagogical adaptation, and the future of teaching in technology-mediated learning environments.

Keywords: Teacher roles, emerging technologies, digital competencies, educational technology, systematic literature review

1. Introduction

The role of the teacher has undergone a transformation in response to the rapid advancement of technology. Once regarded as the primary source of knowledge and authority within the classroom, teachers are now increasingly expected to serve as facilitators, mentors, co-learners, and designers of learning experiences (Todorova, 2022). This shift reflects a broader movement in education toward learner-centred pedagogies, which are being increasingly shaped and supported by innovations in educational technology, including online platforms, mobile learning tools, and, more recently, artificial intelligence (AI) (Demuner et al., 2024). Historically, teaching was largely characterised by direct instruction and content delivery within the physical confines of the classroom.

However, each wave of technological innovation has challenged and expanded this traditional paradigm. The integration of radio and television into classrooms in the mid-20th century, followed by the proliferation of the internet and the rise of Massive Open Online Courses (MOOCs), significantly altered how teachers engaged with content and learners (Milligan et al. 2013). Today, the advent of AI-driven systems such as intelligent tutoring systems, adaptive learning platforms, and AI chatbots marks a new era in which teachers must navigate increasingly complex educational ecosystems. While early narratives often

positioned technology as a potential replacement for teachers, the reality has proven more nuanced. Research consistently shows that technology's effectiveness depends largely on how teachers interpret, adapt, and apply it within their unique pedagogical and cultural contexts (Hazzan-Bishara et al. 2025; Amin 2016).

The novel affordances of digital tools can also reshape classroom dynamics and teacher-student relationships. With students now having instant access to vast amounts of global information, the teacher's role could shift from being a sole knowledge provider to becoming a curator, contextualizer, and guide through an overwhelming information landscape. This change also redistributes power in the classroom, as students assume more active roles in constructing knowledge. Consequently, teachers are expected to adopt more dynamic instructional strategies that support collaboration, personalization, and inquiry-based learning, while also attending to the socio-emotional needs of learners, areas in which technology remains limited. In this evolving landscape, teachers must not only facilitate learning but also act as ethical stewards and digital navigators. They are required to demonstrate proficiency in educational technologies, interpret real-time data dashboards, manage hybrid or blended learning environments, and ensure equitable access to learning for all students (Todorova, 2022). These demands place a premium on teacher adaptability, professional development, and reflective practice (Amin, 2016).

Despite growing scholarly interest in technology-enhanced education, much of the existing literature has disproportionately focused on student outcomes and experiences. Comparatively less attention has been paid to how teachers experience and negotiate their changing roles within tech-rich environments (Todorova, 2022). While some frameworks, such as TPACK (Koehler et al. 2013) and studies on changing (Dwyer, 1991) and persistent teacher beliefs and practices (Ertmer et al. 2012), and teachers' role in technology integration (Wohlfart & Wagner, 2023), have illuminated aspects of these transformations, the literature remains fragmented. In particular, there is a lack of longitudinal and synthesized perspectives that trace teacher role evolution in relation to specific technological epochs. Additionally, the emotional, relational, and identity-related challenges that teachers face in navigating these transitions remain underexplored (Amin, 2016).

Against this backdrop, the present study conducts a systematic literature review to examine how the role of the teacher has evolved across four key technological eras: the Pre-Digital era, the Broadcast Media era, the Internet era, and the Artificial Intelligence era. Specifically, it seeks to answer the question: *How has the teacher's role transformed in response to technological developments from the Pre-Digital to the Artificial Intelligence era?* By mapping these transformations, the study aims to contribute a synthesized understanding of teacher roles with evolving technologies.

2. Methodology

This study employs a structured literature review methodology to investigate the evolving role of teachers in response to technological developments in education. This methodological approach was selected due to its capacity for rigor, transparency, and replicability in synthesising existing scholarly knowledge (Massaro, Dumay, & Guthrie, 2016). The study draws from a broad corpus of academic literature to explore longitudinal and cross-contextual trends concerning teacher role transformation across four distinct technological eras: the Pre-Digital era (chalkboard/print), Broadcast Media era (radio/TV), Internet era (web-based tools), and Artificial Intelligence era (AI-driven platforms). The overarching objective of this review is to synthesize and interpret how the professional roles and responsibilities of teachers have transformed in response to successive waves of educational technologies. To gain a deeper and more nuanced understanding of this transformation, two sub-questions were formulated: (a) What pedagogical roles were dominant during each technological era? and (b) What challenges did teachers face in each era?

2.1 *Paper Selection Process*

The literature for this review was curated through a purposive selection strategy, employing the snowballing technique as its primary method. This approach facilitated the identification of highly relevant sources by tracing citations from foundational texts and incorporating key recommendations from academic mentors. Papers were selected based on their relevance to the central research questions, representation of diverse perspectives, and their capacity to provide insight into the four technological eras under investigation. In addition to expert-recommended readings, targeted searches were done on Google Scholar, ERIC, and ChatGPT. These searches were guided by a structured combination of keywords and Boolean operators to ensure breadth and depth of coverage. The search encompassed literature published between the 1970s and 2024, and spanned the domains of education, educational technology, teacher education, and instructional design. The intention was to capture both historical and contemporary perspectives on the evolving role of teachers within technology-integrated education systems.

Key search phrases included combinations such as “teacher role” AND “technology,” “educational technology” AND “pedagogical change,” “teacher identity” AND (“broadcast media” OR “television” OR “radio”), “online learning” OR “digital learning” AND “teacher facilitation,” and “AI in education” AND “teacher roles” OR “pedagogy.” A total of 26 papers were selected, representing a diverse mix of empirical studies, conceptual frameworks, theoretical analyses, systematic reviews, and policy-oriented reflections. Inclusion criteria required the paper to explicitly discuss the role or function of teachers in relation to educational technologies from one of the four technological contexts defined in the study. Selected papers spanned a variety of geographical and educational levels, with English as the publication language. Both peer-reviewed journal articles and pertinent grey literature, such as institutional reports and conference proceedings, were considered to ensure the inclusion of a wide range of scholarly and practice-based perspectives. Exclusion criteria were applied to omit papers that focused solely on student learning outcomes without reference to teacher roles, those that were primarily technical (e.g., emphasizing tool design or usability) with minimal pedagogical context, generic discussions of Educational Technology lacking specificity regarding teacher practices or transformations, non-peer-reviewed sources such as blogs and newspaper articles, and studies that did not situate their findings within a specific technological era. This rigorous selection process ensured the inclusion of literature that was both contextually grounded and analytically robust, thereby providing a strong foundation for synthesizing insights on the shifting roles, identities, and responsibilities of teachers over time.

2.2 *Categorization by Technological eras*

To systematically trace the evolution of the teacher’s role over time, the selected literature was organized into four major technological eras. This temporal categorization served as an analytical framework to compare how teachers resisted or adapted to new tools, how pedagogical practices evolved, and how broader educational paradigms shifted in response to emerging technologies. The four technological eras are:

- Pre-Digital era (Before the 1970s), characterized by traditional classroom instruction utilizing chalkboards, printed textbooks, and face-to-face learning;
- Broadcast Media era (1970s–1990s), marked by the integration of educational radio, television, audio-visual media, and early forms of computer-assisted instruction;
- Internet era (1990s–2010s), encompassing web-based tools, learning management systems, MOOCs, and e-learning environments; and
- Artificial Intelligence era (2010s–Present), defined by the rise of adaptive learning platforms, AI tutors, chatbots, intelligent tutoring systems, data analytics dashboards, and generative AI technologies.

This classification enabled a diachronic and cross-contextual analysis of how teacher roles have evolved not only in terms of functionality but also in their social, cultural, and ethical dimensions. The literature was reviewed chronologically according to these four

technological categories, allowing for a systematic comparison of pedagogical transformations over time. Each study was coded and analyzed using a structured coding framework aligned with the research questions. Key data points extracted during the analysis included: the technological context or era addressed, the conceptualization of teacher roles, the specific challenges reported by teachers, and the pedagogical changes or role shifts identified within each study. This analytical strategy facilitated a nuanced synthesis of the evolving teacher identity in tandem with the shifting technological landscape of education.

3. Findings

The findings of this structured literature review are organised under the four technological eras identified. Within each era, key patterns emerged concerning the pedagogical approaches adopted by teachers, the challenges encountered, and the nature of role transformation in response to technological change.

3.1 *Pre-Digital era*

In the pre-digital era, roughly spanning the early to mid-20th century, education was predominantly a teacher-centred endeavor. The classroom relied on basic instructional tools such as chalkboards, textbooks, printed handouts, maps, and physical models. This is a dominant mode in most of the classrooms, in developing countries until recently, where we find hints of the presence of digital infrastructure. Reviewing the origins of classrooms in India from the British era, Kumar (1988) provides a very good picture of such a classroom, calling it a textbook culture. Teachers, anchored to the textbooks, operated as the sole authority and transmitter of knowledge, controlling the pace, flow, and nature of instruction. The student's role was largely passive, limited to note-taking, repetition, and memorization. The chalkboard, in particular, served as the central visual aid and symbol of instructional delivery. It was used for problem-solving demonstrations, illustrations, writing key points, and organizing classroom dialogue. However, despite its value in visualizing content, its static nature offered no interactivity or feedback, and its use was entirely dependent on the teacher's ability to engage students through explanation and repetition (Karnam et al. 2020).

Classroom practices in this era typically involved rote learning, direct instruction, and whole-class teaching, often structured around fixed textbooks and curriculum-centred delivery. Teaching styles were hierarchical and unidirectional: the teacher lectured, students listened and wrote. This approach reinforced the teacher's identity as a subject expert, moral authority, and classroom manager. Importantly, the literature from this period reflects a consistent view that the teacher was the "centrepiece" of education, whose presence was essential not only for delivering information but also for maintaining discipline and moral order. Technology had a minimal role, and educational success was deeply intertwined with teacher authority, personal charisma, and classroom presence.

3.2 *Broadcast era (Radio and Television)*

The Broadcast Era, emerging in the 1920s and reaching widespread adoption during the 1970s–1990s, redefined the teacher's role through the integration of radio and instructional television into education. These technologies enabled mass dissemination of knowledge across geographic and socioeconomic boundaries, particularly in rural or under-resourced areas (Hawkridge 1979; Arulchelvan & Viswanathan 2008). Radio and television brought centralized, expertly produced lessons into classrooms and homes, leading many to speculate whether teachers might eventually become obsolete. Initially, radio was hailed as a revolutionary tool. Educational radio broadcasts, often aligned with school curricula, were introduced in countries like Canada and the United States with strong institutional support. In theory, these broadcasts offered standardized, high-quality content from national experts. However, in practice, teachers remained essential mediators. They were responsible for

preparing students to listen, facilitating discussions around the broadcasts, and contextualizing the information for specific classroom needs. In many cases, teachers were expected to “fill in the gaps” left by the generality of the broadcasts, thereby reinforcing their relevance rather than diminishing it.

Instructional TV (Whittington 1987) later expanded on this model by adding visual elements to the learning experience. Teachers used TV to show science demonstrations, dramatizations of historical events and language instruction videos. Additionally, resistance emerged from teachers who feared that technology could de-skill their profession or reduce their autonomy in content delivery. Despite these concerns, research consistently showed that teachers were not displaced but repositioned. Rather than being replaced by broadcasts, teachers adapted by acting as facilitators, content translators and classroom moderators, often supplementing broadcasts with discussion, written assignments and local examples. This era thus began a subtle redefinition of the teacher’s identity from the sole deliverer of content to a more dynamic role that involved integrating and contextualizing external media resources.

Some experimental practices during this time further highlighted this evolving role. For instance, “TV schools” during the Cold War attempted to replace teachers with televised instruction, especially during teacher shortages (Maccoby 1966). But these initiatives often failed due to a lack of emotional engagement and classroom management issues, both of which required a human presence. The broadcast era showed that while technology could amplify and enrich instruction, it could not replace the human dimensions of teaching. Teachers remained vital as the interpreters, moderators and emotional anchors in a system increasingly mediated by the media.

3.3 *Internet era*

The emergence of the internet in the late 20th century brought a radical transformation to education, shifting teaching practices from fixed, broadcast-based models to interactive, decentralized, and learner-driven environments (Amin, 2016). This era introduced web-based platforms, learning management systems (LMSs), MOOCs, online forums, collaborative tools (e.g., Google Docs), and virtual classrooms, each of which began to redefine the roles, responsibilities, and required competencies of teachers (Fitria & Suminah, 2020; Todorova, 2022). In contrast to the one-way nature of television and radio, the internet enabled bidirectional, multimodal communication, giving students the power to engage with content, peers, and instructors in new ways. This shift placed pressure on teachers to relinquish some degree of control and embrace more facilitative roles. They were now expected to curate digital resources, scaffold learner autonomy, foster online discussions, and provide individualized feedback within increasingly diverse virtual environments (Todorova, 2022; Amin, 2016).

Tools like MOOCs and asynchronous forums shifted the locus of learning from the classroom to the cloud. Teachers designing or facilitating MOOCs had to think at scale, focusing on clear structure, self-paced navigation and embedded feedback mechanisms (A. Triantafyllou 2021). The learning management system (LMS) became the new administrative and pedagogical centre, with teachers uploading resources, managing submissions, tracking analytics and conducting assessments online (Beaudoin 1990). In blended and flipped classrooms, teachers recorded lectures for homework and used in-class time for discussion and problem-solving. This inverted the traditional flow of classroom instruction, forcing educators to shift from performers to planners, from content deliverers to learning architects. For example, in online collaborative tools like Google Docs or Padlet, students co-constructed knowledge while teachers could monitor, guide, or give feedback often asynchronously.

The literature also reflects growing concern over the emotional and cognitive demands placed on teachers during this transition. Teachers had to adapt quickly to digital tools, often without sufficient training or institutional support (Mehboob et al., 2024). In the context of pedagogical challenges, one major thing is maintaining student engagement in online settings, where a lot of distraction weakens the learners’ focus. Moreover, managing

online engagement, learner motivation, and digital behavior introduced new pedagogical and ethical challenges (Beaudoin, 2013). They face difficulty in fair assessment and evaluation, due to concerns around academic honesty and limited control in the virtual environment. There exists a digital divide as well between teachers and students. From an ethical perspective, there also exist challenges around equity and accessibility as well as the privacy and security of student data in online platforms (Demuner et al., 2024). Teachers also report struggling to manage student behavior digitally, and the literature documents the emergence of teacher burnout in online platforms, where educators sometimes find it difficult to balance professional demands with personal life due to the blurred boundaries of online teaching.

Despite these pressures, the internet era also empowered teachers. With access to global content and communities, they could share resources, engage in professional development and adopt more inclusive, differentiated and student-centered pedagogies (Amin, 2016; Mehboob et al., 2024). They also became designers of learning experiences, using data and feedback tools embedded in platforms to tailor interventions and support struggling students (Beaudoin, 2013). The internet era redefined teachers as facilitators, co-learners, designers and data-informed mentors. While it decentralized knowledge, it simultaneously elevated the pedagogical creativity and technological fluency required of educators (Todorova, 2022; Amin, 2016)

3.4 *AI era*

The rise of artificial intelligence in education marks a new paradigm in the transformation of the teacher's role. Unlike earlier technologies that extended or mediated teacher practices, AI tools such as intelligent tutoring systems (ITS), chatbots, automated assessment platforms, and generative AI models like ChatGPT now can simulate cognitive and, to some extent, emotional functions of teaching (Gentile et al., 2023; Chan & Tsi, 2023). This includes providing personalized feedback, modeling problem-solving processes, tracking student progress in real time, and even engaging in natural language conversations (Guo et al., 2024; Chan & Tsi, 2023). These affordances have shifted the teacher's role further from being the sole knowledge source to being a learning orchestrator, someone who coordinates interactions between learners, platforms, and content (Gentile et al., 2023; Todorova, 2022). Teachers must now design blended environments where AI can automate routine instruction or diagnostics, while they focus on complex pedagogical tasks such as ethical reasoning, emotional support and creative facilitation (Chan & Tsi, 2023; Gentile et al., 2023; Zhai, 2024). Teachers working with these systems are expected to interpret performance dashboards, diagnose learning gaps based on system reports and intervene only when needed, an approach sometimes referred to as "data-driven pedagogy" (Guo et al., 2024; Zhai, 2024). As shown in studies like those by Guo et al. (2024), teachers increasingly depend on visual analytics tools to understand students' problem-solving trajectories and make responsive instructional decisions. However, literature also reveals deep tensions. Many educators express concerns about being reduced to supervisors of algorithmic systems, losing autonomy, or being deskilled by black-box technologies (Gentile et al., 2023; Zhai, 2024). Others worry about the ethical dimensions of AI use issues of surveillance, bias, data privacy and authenticity in student work (Chan & Tsi, 2023; Zhai, 2024).

Yet, AI has also brought new opportunities for reclaiming the human core of teaching. Since AI struggles with empathy, moral judgment, cultural sensitivity, and adaptive mentoring, teachers are increasingly valued for their emotional intelligence and relational presence (Chan & Tsi, 2023; Gentile et al., 2023; Zhai, 2024). In some classrooms, teachers use AI as a co-teacher or brainstorming partner, while retaining full control over how learning is scaffolded, paced, and evaluated. Thus, in the AI era, teachers emerge as critical agents who not only facilitate learning but also mediate the impact of technology on students' lives. They are required to be ethically aware, emotionally grounded, and technologically fluent—roles far beyond the traditional instructor and even beyond the facilitator of the internet age (Gentile et al., 2023; Chan & Tsi, 2023; Todorova, 2022). AI doesn't eliminate the teacher, it amplifies the complexity and value of human judgment in education.

Table 1. An *Evolution of Teacher's Role and instructional approaches Across Technological eras along with challenges*

Era	Dominant Technology	Teacher's Role	Pedagogical approaches	Challenges
Pre-Digital (Kumar,1988)	Chalkboard Print textbooks Face-to-face instruction	Knowledge Authority	Direct, rote learning, textbook-driven teacher-centred, uniform curriculum	Limited interactivity, centralized control, limited access to resources, textbook dependency, no differentiation: lack of individualized instruction
Broad cast (Hawkridge, 1979; Arulchelvan & Viswanathan, 2009; Mehboob et al., 2024; Stefanyshyn & Kendell, 2012)	Radio TV Audio-visual media	Mediator and Moderator	Passive + supplemented, guided discussion	Loss of autonomy, lack of context, infrastructural barriers, lack of teacher training, reduced content control
Internet (Todorova, 2022; Milligan et al., 2013; Amin, 2016; Triantafyllou, 2021; Li, 2023; Fitria & Suminah, 2020; Beaudoin, 1990, 2013)	Web 2.0 tools LMS MOOCs, E-learning platforms	Learning Facilitator, Digital course designer, Content curator, Online moderator	Interactive, asynchronous & blended learning, networked collaboration	Engagement, assessment fairness, digital divide, tool overload, pedagogical shift resistance
AI (Zhai, 2024; Taufikin et al., 2024; Cai, 2024; Demuner et al., 2024; Gentile et al., 2023; Chan &Tsai, 2023; Fitria, 2023; Li & Wang, 2020)	GenAI (e.g. ChatGPT), AI-based tools (eg.Intelligent Tutoring Systems, chatbots), Adaptive learning platforms, AI dashboards (data analytics), Virtual Assistants	Orchestrator of personalised learning paths, Socio-emotional and ethical guide and mentor, Data interpreter, Decision-maker, Content co-creator	Personalized and data-driven learning, Co-creation with AI, Learner-centred design, Real-time formative assessment, Blended AI-human instruction	Digital and AI literacy gaps among teachers, ethical concerns (like privacy, bias, data misuse), shifting teacher identity and role clarity, overdependence on AI systems, deskilling fears, dilemmas to the professional roles of teachers, like loss of authority/ role confusion emotional and relational limitations of AI

4. Discussion and Conclusion

The synthesis of literature across four technological eras reveals a progressive, though non-linear, transformation in the teacher's role, shaped by shifting pedagogical responsibilities and the persistent challenges of integrating technology. While educational technology has often been framed as a driver of innovation and efficiency, evidence shows it has consistently redefined rather than replaced the centrality of teachers (Rifah & Zamahsari, 2022; Carrillo, 2012; Gentile et al., 2023). Technology has acted less as a substitute and more as a catalyst, altering how teachers teach, interact, and support learners, while leaving their socio-emotional and relational capacities irreplaceable.

In the Pre-Digital era, pedagogy was largely didactic, with teachers seen as authoritative transmitters of knowledge and moral custodians who shaped learners within the physical classroom (Kumar, 1988). The teacher's professional identity was closely tied to authority, discipline, and the one-way delivery of content.

The Broadcast Media era, marked by radio, television, and audiovisual aids, repositioned teachers as mediators who contextualized centrally produced content for diverse learners (Arulchelvan & Viswanathan, 2015). The integration of broadcast media brought greater structure and standardization into classrooms, but the fixed nature of centrally produced programs often constrained teacher agency. While technology promised scalability and wider reach, the teacher's role remained crucial in adapting static materials to local contexts and ensuring learners' engagement.

With the Internet era, a significant departure from one-way communication emerged. Interactive and learner-centered approaches began to take precedence. Teachers became co-learners, digital content curators, and designers of technology-enhanced environments that encouraged collaboration and personalization (Beblavý, Baiocco, Kilhoffer, Akgüç, & Jacquot, 2019). This era underscored the teacher's capacity to orchestrate digital resources, but it also revealed new challenges: uneven infrastructure, differences in digital competence, and varied pedagogical adaptability. The Internet expanded opportunities for student autonomy, yet reinforced the indispensability of teachers in scaffolding, guiding, and contextualizing learning.

The current Artificial Intelligence era has again reshaped expectations. Teachers are now tasked with managing complex algorithmically mediated learning environments, interpreting AI-driven insights, and addressing ethical questions raised by automated systems (Chan & Tsi, 2023; Zhai, 2024). While AI tools provide powerful support for personalization and assessment, they require teachers to critically evaluate their appropriateness and reliability. The teacher's role has thus expanded to include responsibilities as an interpreter of data, a gatekeeper of ethics, and a mentor fostering students' digital and critical literacies. Despite these technological advances, empathy, adaptability, and interpersonal guidance remain outside the domain of automation and continue to define the human dimensions of teaching.

Across these eras, two imperatives for educational ecosystems become evident. First, teachers must be recognized not as passive recipients of technology but as active agents who co-shape and humanize its application. Their professional judgment determines whether technology is used meaningfully or superficially. Second, the design of educational technologies must account for the realities of teachers' work—their pedagogical commitments, institutional constraints, and contextual variations. As Ertmer and Ottenbreit-Leftwich (2010) argue, technology's potential is only realized when teachers are provided with training, autonomy, and reflective spaces to integrate innovations into practice.

While the categorization into eras provides a useful framework for tracing historical shifts, it risks flattening the diversity of teachers' experiences. Adoption trajectories differ by region, school type, and subject domain, meaning that the teacher's role is better understood as a culturally and temporally situated construct, shaped by the interplay of pedagogy, policy, and technology. Rather than a linear progression, role evolution is a process of dynamic negotiation, with teachers continually reinterpreting their professional identity in response to new tools, systemic pressures, and pedagogical ideologies.

This review contributes a synthesized, diachronic understanding of how teacher roles have evolved across four major eras—Pre-Digital, Broadcast Media, Internet, and AI. It highlights continuities and shifts, showing that despite recurring claims about teacher replacement, technology has consistently augmented the importance of educators. Teachers emerge not as obsolete but as adaptive, relational, and ethically grounded professionals. Socio-emotional competencies, critical thinking, and pedagogical judgment remain beyond automation, reaffirming the centrality of human educators in technology-mediated contexts.

The implications for policy and practice are significant. Educational leaders and policymakers must invest in sustained teacher training that extends beyond technical skills to include ethical reasoning, emotional intelligence, and pedagogical adaptability. Technology developers should involve teachers directly in design and implementation processes to ensure tools are responsive to classroom realities. Teacher education programs must prepare future educators for fluid roles requiring technological fluency alongside human-centered pedagogy.

Nonetheless, limitations remain. This review draws primarily on English-language and purposively selected literature, constraining global generalizability and possibly introducing selection bias. The division into four eras, while analytically helpful, simplifies the uneven and overlapping nature of technology adoption across contexts. Future research should move toward more nuanced analyses through empirical, longitudinal, and comparative studies, with particular attention to how teachers negotiate identity, ethics, and agency in AI-mediated classrooms. Action research and co-design involving teachers and developers could ground innovation in pedagogical reality, while cross-disciplinary and cross-level studies would illuminate how technological shifts differentially reshape teaching roles.

References

Amin, J. N. (2016). Redefining the role of teachers in the digital era. The International Journal of Indian Psychology, 3(3), 40–45.

Arulchelvan, S., & Viswanathan, D. (2009). Radio, television and the Internet providing the right to education in India. Asian Journal of Distance Education, 6(1), 39–52.

Beaudoin, M. (1990). The instructor's changing role in distance education. American Journal of Distance Education, 4(2), 21–29.

Beaudoin, M. F. (2013). The evolving role of the instructor in the digital age. In Learning management systems and instructional design: Best practices in online education (pp. 233–247). IGI Global.

Beblavý, M., Baiocco, S., Kilhoffer, Z., Akgüç, M., & Jacquot, M. (2019). *Index of readiness for digital lifelong learning: Changing how Europeans upgrade their skills (Final report).* Centre for European Policy Studies. https://www.ceps.eu/ceps-publications/index-of-readiness-for-digital-lifelong-learning/

Bingimlas, K. A. (2009). Barriers to the successful integration of ICT in teaching and learning environments: A review of the literature. Eurasia Journal of Mathematics, Science and Technology Education, 5(3), 235–245.

Cai, Z. (2024). Research on teachers' role orientation and strategies in the age of artificial intelligence. Advances in Educational Technology and Psychology, 8(4), 64–72.

Carrillo, F. A. G. (2012). Can technology replace the teacher in the pedagogical relationship with the student? Procedia – Social and Behavioral Sciences, 46, 5646–5655.

Chan, C. K. Y., & Tsi, L. H. (2023). The AI revolution in education: Will AI replace or assist teachers in higher education? arXiv preprint arXiv:2305.01185.

Demuner, J. A., Gomes, J. de S., Kreuzberg da Silva, C., da Cruz, N. A., dos Reis, D. A., & de Souza, M. de F. (2024). The role of the teacher in the age of emerging technologies: Challenges and opportunities to reconfigure teaching and learning. Revista Aracê, 6(2), 2536–2551.

Dwyer, D. (1991). Changes in teachers' beliefs and practices in technology-rich classrooms. Educational Leadership, 48, 45–52. https://www.academia.edu/download/8018790/acotmay91.pdf

Ertmer, P. A., Ottenbreit-Leftwich, A. T., Sadik, O., Sendurur, E., & Sendurur, P. (2012). Teacher beliefs and technology integration practices: A critical relationship. Computers & Education, 59(2), 423–435.

Fitria, H., & Suminah. (2020). Role of teachers in the digital instructional era. Journal of Social Work and Science Education, 1(1), 70–77.
Fitria, T. N. (2023). The use of artificial intelligence in education (AIED): Can AI replace the teacher's role? Epigram, 20(2), 165–187.
Guo, G., Kumar, A. M. S., Gupta, A., Coscia, A., Maclellan, C., & Endert, A. (2024, June). Visualizing intelligent tutor interactions for responsive pedagogy. In Proceedings of the 2024 International Conference on Advanced Visual Interfaces (pp. 1–9).
Gentile, M., Città, G., Perna, S., & Allegra, M. (2023, March). Do we still need teachers? Navigating the paradigm shift of the teacher's role in the AI era. In Frontiers in Education (Vol. 8, Article 1161777). Frontiers Media SA.
Hawkridge, D. G. (1979). Print, television, and radio in education: Principles and examples. The International Conference on Television, Radio and Print, Italy.
Hazzan-Bishara, A., Kol, O., & Levy, S. (2025). The factors affecting teachers' adoption of AI technologies: A unified model of external and internal determinants. Education and Information Technologies, 1–27. https://doi.org/10.1007/s10639-025-13393-z
Karnam, D. P., Mashood, K. K., & Sule, A. (2020). Do student difficulties with vectors emerge partly from the limitations of static textbook media? European Journal of Physics, 41(3), 035703. https://doi.org/10.1088/1361-6404/ab782e
Koehler, M. J., Mishra, P., & Cain, W. (2013). What is technological pedagogical content knowledge (TPACK)? Journal of Education, 193(3), 13–19.
Kumar, K. (1988). Origins of India's "textbook culture." Comparative Education Review, 32(4), 452–464.
Li, Y. (2023). Exploration of practical teaching: Impact of Internet era on teachers and students and corresponding strategies. Educational Innovations and Emerging Technologies, 3(2), 7–10. https://doi.org/10.35745/eiet2023v03.02.0002
Liu, J., & Wang, S. (2020, June). The change of teachers' role in teaching under the environment of artificial intelligence. In 2020 International Conference on Artificial Intelligence and Education (ICAIE) (pp. 98–102). IEEE.
Maccoby, G. C. A. (1966). The Peace Corps Educational Television (ETV) Project in Colombia – Two years of research. https://files.eric.ed.gov/fulltext/ED020663.pdf
Massaro, M., Dumay, J., & Guthrie, J. (2016). On the shoulders of giants: Undertaking a structured literature review. *Accounting, Auditing & Accountability Journal, 29*(5), 767–801. https://doi.org/10.1108/AAAJ-01-2015-1939
Mehboob, K., Kamran, F., & Ain, Q. U. (2024). From chalkboards to digital platforms: The impact of media on modern education. Gomal University Journal of Research, 40(4), 411–424.
Milligan, C., Littlejohn, A., & Margaryan, A. (2013). Patterns of engagement in connectivist MOOCs. Journal of Online Learning and Teaching, 9(2), 149–159.
Meccawy, M. (2023). Teachers' prospective attitudes towards the adoption of extended reality technologies in the classroom: Interests and concerns. Smart Learning Environments, 10(1), 36.
Rifah, L., & Zamahsari, G. K. (2022, November). Can technology replace the teachers' role in higher education settings? A systematic literature review. In Proceedings of the 7th International Conference on Sustainable Information Engineering and Technology (pp. 217–221).
Stefanyshyn, D., & Kendell, J. (2012). The influence of radio and television on culture, literacy, and education. Retrieved from https://blogs.ubc.ca/etec540sept12/2012/10/28/1687/
Triantafyllou, S. A. (2021). MOOCs and EdTech tools for distance learning. In *Proceedings of the 4th World Conference on Future of Education.* Diamond Scientific Publishing. https://www.dpublication.com/proceeding/4th-wcfeducation/
Taufikin, M. S. I., Azifah, N., Nikmah, F., & Kuanr, J. (2024, April). The impact of AI on teacher roles and pedagogy in the 21st century classroom. In 2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS) (Vol. 1, pp. 1–5). IEEE.
Todorova, T. (2022). The teacher in the digital world – New roles, representations, perceptions. In Proceedings of the International Conference on Future of Teaching and Education (Vol. 1, pp. 70–80). https://doi.org/10.33422/icfte.v1i1.5
Whittington, N. (1987). Is instructional television educationally effective? A research review. American Journal of Distance Education, 1(1), 47–57. https://doi.org/10.1080/08923648709526572
Wohlfart, O., & Wagner, I. (2023). Teachers' role in digitalizing education: An umbrella review. Educational Technology Research and Development, 71(2), 339–365.
Zhai, X. (2024). Transforming teachers' roles and agencies in the era of generative AI: Perceptions, acceptance, knowledge, and practices. Journal of Science Education and Technology, 1–11. https://doi.org/10.1007/s10956-024-10174-0

An Exploratory Study on the Influence of Anonymity, Personality, and Social Factors on Participation in Online College Communities

Eleana Louise HERNANDEZ[a], Vanessa Claire MA[a], Marco Lorenzo MITRA[a], Marco Gabriel TARROSA[a] & Ryan EBARDO[a,*]
[a]*De La Salle University, Philippines*
*ryan.ebardo@dlsu.edu.ph

Abstract: Online communities based on social media have become extensions of university life, where students interact, seek information, engage in academic discourse, and build social relationships. In these spaces, anonymity, personality traits, and social factors influence the type of information students share and consume. Guided by the Theory of Reasoned Action (TRA), this study examined how perceived anonymity, neuroticism, extraversion, and fear of missing out (FoMO) affect the behavioral intention of 212 students to participate in online college communities. Using Partial Least Squares – Structural Equation Modeling (PLS-SEM), the analysis revealed that FoMO does not have a positive effect on participation intention. However, perceived anonymity, neuroticism, and extraversion were found to positively influence students' attitudes toward participation, thereby affecting their intention to engage in anonymous college communities. The study recommends further investigation into all Big Five personality traits, exploration of specific anonymous platforms such as subreddits or Facebook pages, and a deeper analysis of the relationship between FoMO and anonymity. Understanding these factors can help shape academic policies that leverage social media platforms as alternative spaces for academic discourse.

Keywords: Anonymous College Communities, Theory of Reasoned Action (TRA), Perceived Anonymity, Personality Traits, Fear of Missing Out (FoMO)

1. Introduction

Using anonymous online communities is becoming more prevalent, as it allows users to freely discuss thoughts, opinions, and experiences while keeping identities hidden. Among these communities, pages specific to colleges have risen. These come in the form of Facebook freedom walls, university subreddits, and confession pages. With the unique anonymity feature of such online platforms, college students share challenges and experiences, seek emotional support, and discuss coping strategies (Rodriguez et al. 2023; Sacks et al., 2021). Furthermore, these communities play a crucial role in providing students a safe space to express themselves and find support, while also offering valuable insights for educators and administrators to better understand and address student needs. Existing research on these communities is scant, and it is crucial for studies to find how various personality traits and social factors influence attitudes, intentions, and participation in these online spaces. While personality traits and social factors have been widely studied in social media, minimal studies explore their effect in anonymous contexts. Using the Theory of Reasoned Action (TRA) as a theoretical framework, the study analyzed the effect of anonymity, extraversion, and neuroticism on attitude; and subjective norms and fear of missing out (FoMO) on Filipino college students' intention to participate in anonymous college communities.

2. Literature Review

Sacks et al. (2021) found that confession pages feature a wide range of concerns such as parking, food, loneliness, university life, and academic concerns. In a Philippine context, the findings of Rodriguez et al. (2023) suggest that these pages provide a supportive and encouraging environment for students to share their experiences and discuss different topics.

One of its features is anonymity, which affects online behavior, as individuals seek anonymity to express themselves more freely or behave toxically (Nitschinsk et al., 2023). Higher levels of perceived anonymity also foster a stronger sense of moral courage (Pan et al., 2023). Meanwhile, Carpenter and Wilet (2021) concluded that anonymity can encourage self-disclosure, leading to more open discussions. Personality traits can also shape one's online behavior. Among the Big Five Personality Traits, extraversion is one of the highest predictors of social media use, as it is easier for people who are more sociable to express themselves in both real-life and online scenarios. (Lampropolous et al., 2022; Bunz, 2021). For neuroticism, those experiencing unstable emotions tend to seek social contact (Chen and Peng, 2022). Additionally, a motivation to use social media for those with high neuroticism is to socially interact, despite having small online networks (Bowden-Green et al., 2021).

Subjective norms can also shape social media use intentions. In students, subjective norms may come in the form of peer influence. Trivedi et al. (2021) found that peer influence significantly mediates the relationship between performance expectancy, communication functionality, and self on intention to use social media in higher education. This mediation implies that students may experience influences from their peers which impacts intentions to use social media. Additionally, peer influence has been found to affect compulsive social media use (Wang et al., 2023). Lastly, Fear of Missing Out (FoMO) has been found to affect social media use. The meta-analysis by Fioravanti et al. (2021) revealed a strong positive correlation between FoMO and both normal and problematic social media use. In a longitudinal study, Groenestein et al. (2024) found that FoMO and social media use were positively correlated, and they found part of a reciprocal trajectory.

This study will extend the TRA to integrate perceived anonymity, extraversion, neuroticism, and FoMO in order to add to the current literature surrounding anonymous college communities.

3. Theoretical Foundation

TRA posits that an individual's intention to perform a behavior is the strongest predictor of that behavior (Ajzen & Fishbein, 1980; Ajzen, 1975). Behavioral intention is shaped by two primary constructs: attitude toward the behavior and subjective norms. TRA has been widely applied to explain various digital behaviors, including participation in online communities, making it a relevant lens to understand students' intention to participate in anonymous college communities. Perceived anonymity refers to the degree to which users believe their identity is protected in a given online space. Although not originally part of TRA, this construct is relevant in digital contexts where identity protection is a key factor in online participation. Thereby positively affecting attitude towards anonymous college communities. Individual personality traits are included to account for personal predispositions that may influence students' attitudes. Extraversion is the tendency to be outgoing and sociable, while neuroticism is the tendency to experience negative emotions such as anxiety and sadness. FoMO is introduced as a motivational construct that may amplify behavioral intention. FoMO is defined as the apprehension that others are having rewarding experiences from which one is absent. In anonymous communities, where trending confessions or viral posts circulate rapidly, students may feel compelled to stay engaged to avoid social exclusion.

The study hypotheses are summarized in Figure 1 and are framed into the following statements:

H1: Perceived anonymity positively affects attitude towards participation in anonymous college communities.

H2: Neuroticism leads to a positive attitude toward anonymous college communities.

H3: Extraversion leads to a positive attitude toward anonymous college communities.
H4: Subjective Norms lead to positive intention to participate in anonymous college communities.
H5: Fear of Missing Out leads to positive intention to participate in anonymous college communities.
H6: Positive attitude leads to positive intention to participate in anonymous college communities.
H7: Positive intention leads to actual participation in anonymous college communities.

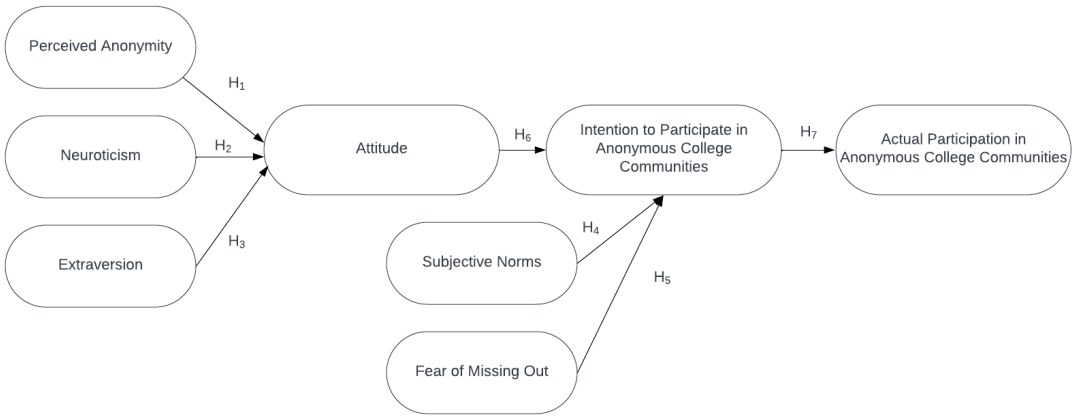


Figure 1. Theoretical Framework and Hypotheses

4. Methodology

A quantitative research design based on the TRA will be used to investigate how different constructs influence Filipino college students' participation in anonymous online communities. To validate the hypotheses and examine the applicability of the TRA in the context of anonymous college communities, the researchers adapted questions from previous validated studies to construct a comprehensive survey. Perceived Anonymity was measured using Hite et al.'s (2014) Perceived Anonymity Measurement Instrument. Neuroticism and Extraversion were measured using items from the Big Five Inventory (John et al., 1991; John & Srivastava, 1999). Subjective Norms were adapted from the study of Yu and Song (2021), while the FoMO construct used the scale developed by Przybylski et al. (2013). Questions for Attitude and Intention were drawn from Peslak et al. (2012), based on the TRA framework, and items related to Actual Participation in confession pages were adapted from Le et al. (2021). The survey was distributed to college students from nine universities across the Philippines, targeting those aged 18 and above. A pilot survey, consisting of 36 participants, was administered. A total of 212 responses were collected before proceeding to the data analysis.

Pilot test results from a sample of 36 students, shown in Table 1, were examined using the Partial Least Squares regression method with SmartPLS software. This method was chosen because it is suitable for predictive research based on our conceptual framework with multiple hypotheses. Bootstrapping was done in SmartPLS to correctly assess path significance with the sample size. This validation analysis demonstrates that the instrument meets the reliability and validity standards set by Hair et al. (2017), with Cronbach's Alpha, Average Variance Extracted (AVE), and Composite Reliability all exceeding the recommended minimum thresholds of 0.70, 0.50, and 0.70, respectively.

Table 1. *Instrument Validation*

Construct	Cronbach's Alpha	Composite Reliability (CR)	Average Variance Extracted (AVE)
Perceived	0.911	0.967	0.716

Anonymity			
Neuroticism	0.748	0.977	0.528
Extraversion	0.826	0.930	0.635
Subjective Norms	0.871	0.869	0.733
FoMO	0.862	0.894	0.506
Attitude	0.849	0.868	0.621
Intention to Participate	0.877	0.879	0.804
Actual Participation	0.907	0.924	0.781

Bootstrapping involves repeatedly drawing random samples with replacement from the original dataset to generate a large number of simulated samples. Hair et al. (2017) states that bootstrapping is especially useful in small-sample research contexts, where traditional parametric assumptions may not hold, and the statistical power is limited. A t-statistic of 1.96 or higher is required to establish a significant relationship at a significance level of 5%, which corresponds to p-values of 0.05 or lower (Hair et al., 2017). Table 2 shows the path coefficient, t-statistics, and p-values extracted for each hypothesis.

Table 2. *Structural Model Test Results: Factors Affecting Attitude, Intention, and Actual Participation in Anonymous College Communities*

Hypotheses	Path Coefficient	T-statistics	P-values	Decision
H1: Perceived Anonymity -> Attitude	0.255	3.539	0.000	Supported
H2: Neuroticism -> Attitude	0.214	3.163	0.002	Supported
H3: Extraversion -> Attitude	0.225	3.406	0.001	Supported
H4: Subjective Norms -> Intention	0.399	5.862	0.000	Supported
H5: FoMO -> Intention	0.098	1.597	0.110	Not Supported
H6: Attitude -> Intention	0.306	4.486	0.000	Supported
H7: Intention -> Actual Participation	0.621	10.388	0.000	Supported

5. Results and Discussion

Based on the statistical analysis of H1, it has been highlighted that perceived anonymity positively affects attitude towards participation in anonymous college communities. This finding is consistent with prior studies that revealed individuals with higher levels of perceived anonymity are more likely to participate actively and honestly in digital spaces (Carpenter & Wilet, 2021; Pan et al., 2023). The statistical analysis of the relationship between neuroticism and attitude (H2) indicates a significant relationship between the two. This aligns with previous research which states that those experiencing unstable emotions tend to seek social contact (Lampropolous et al., 2022; Chen & Peng, 2022). The result, explained by Bowden-Green et al. (2021), stated that neurotic people are not motivated to grow their online networks, however, social interaction is a primary motivation for them to use social media.

The significant positive relationship between extraversion and aligns with previous studies where extraversion is the most significant predictor of social media use and intensity (Lampropolous et al., 2022; Bunz, 2021). The study showed a significant relationship between subjective norms and intention to participate (H4). Wang et al. (2023) supports this idea by exploring how individual and peer-related factors lead to compulsive use of social media. Peer influence is a strong mediator between the key determinants in users' intention to use social media (Trivedi et al., 2021). This supports the notion that subjective norms such as perceived expectations from peers play an important role in shaping behavioral intentions. Similarly,

Leijse et al. (2023) conducted research on peer influence showing that while it may lead to negative effects like anxiety, it can also regulate behavior when social norms are emphasized. This could mean that subjective norms encourage open sharing and participation due to perceived peer activeness.

In contrast, there is a statistically insignificant relationship between FoMO and intention to participate. While previous literature suggests that FoMO can drive addictive or habitual use of social media (Brailovskaia & Margraf, 2024; Fioravanti et al., 2021), our results suggest that FoMO alone may not be sufficient to influence deliberate intentions to participate in anonymous spaces. One explanation could be in the nature of the communities themselves: anonymity makes it difficult to form social connections and strong relationships with others. Need to belong is defined as the drive to form and maintain lasting, positive, and significant interpersonal relationships (Baumeister & Leary, 1995, p. 497). It is possible that the lack of significant relationship is due to the underlying motivations, such as need to belong, which may drive them towards more identity-based platforms.

The study revealed that attitude positively influences intention to participate (H6), and intention to participate positively influences actual participation among students in anonymous college communities (H7). These results confirm the original proposition of TRA and other studies in social media (Yuan et al., 2021; Abbas Naqvi et al., 2019), wherein attitude leads to intention, which leads to participation.

6. Conclusion and Recommendations

This study investigates the determinants of students' engagement in anonymous college communities, affirming the relevance of TRA for understanding digital participation. The analysis supports six of seven hypotheses, revealing that perceived anonymity, neuroticism, and extraversion each positively shape students' attitudes toward these platforms, while subjective norms strengthen intention to participate. Attitude emerged as a strong predictor of intention, which in turn significantly drives actual participation, confirming the applicability of TRA in the study's context. Additionally, Fear of Missing Out does not influence intention, suggesting that the motivators for engaging anonymously differ from those in fully identifiable social media contexts.

Future research could broaden the scope by increasing their sample size and incorporating diverse educational institutions to enhance generalizability. Examining additional moderating factors, such as gender and age, may clarify under which conditions personality traits and anonymity exert stronger effects. Additionally, future research could distinguish between specific platforms like Reddit or Facebook, as they may differ in interaction styles and levels of anonymity. Lastly, our research uncovered the insignificant relationship between FoMO and intention to participate, likely due to the anonymous nature of these communities. Further research can explore FoMO in relation to anonymity and how this affects participation in public and private communities. The findings of this study should be considered when designing, moderating, or integrating such platforms into campus life. Understanding what drives students to participate can help promote safer and more purposeful use of anonymity in academic settings.

References

Abbas Naqvi, M. H., Jiang, Y., Miao, M., & Naqvi, M. H. (2020). The effect of social influence, trust, and entertainment value on social media use: Evidence from Pakistan. Cogent Business & Management, 7(1). https://doi.org/10.1080/23311975.2020.1723825

Ajzen, I. (1975). Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research. Contemporary Sociology, 6(2), 244. https://doi.org/10.2307/2065853

Ajzen, I., & Fishbein, M. (1980). Understanding attitudes and predicting social behavior. Englewood Cliffs, NJ: Prentice-Hall.

Alabri, A. (2021). Fear of Missing Out (FOMO): The effects of the need to belong, perceived centrality, and fear of social exclusion. Human Behavior and Emerging Technologies 2022.

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. Psychological Bulletin, 117(3), 497–529.

Bowden-Green, T., Hinds, J., and Joinson, A. (2021). Understanding neuroticism and social media: A systematic review. Personality and Individual Differences, 168.

Brailovskaia, J., & Margraf, J. (2023). From fear of missing out (FoMO) to addictive social media use: The role of social media flow and mindfulness. Computers in Human Behavior, 150, 107984.
Bunz, U. (2021). Investigating the relationship between social media use, big five personality, and well-being. Journal of Communication Technology, 4(3), pp. 25-52.
Carpenter, J. P., and Willet, K. B. S. (2021). The teachers' lounge and the debate hall: Anonymous self-directed learning in two teaching-related subreddits. Teaching and Teacher Education, 104.
Chen, M., & Peng, A. Y. (2022). Why Do People Choose Different Social Media Platforms? Linking Use Motives With Social Media Affordances and Personalities. Social Science Computer Review, 41(2), pp. 330-352. https://doi-org.dlsu.idm.oclc.org/10.1177/08944393211049120
Fioravanti, G., Casale, S., Benucci, S. B., Prostamo, A., Falone, A., Ricca, V., & Rotella, F. (2021). Fear of missing out and social networking sites use and abuse: A meta-analysis. Computers in Human Behavior, 122, 106839. https://doi.org/10.1016/j.chb.2021.106839
Groenestein, E., Willemsen, L., van Koningsbruggen, G. M., & Kerkhof, P. (2024). Fear of missing out and social media use: A three-wave longitudinal study on the interplay with psychological need satisfaction and psychological well-being. New Media & Society.
Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2017). A primer on partial least squares structural equation modeling (PLS-SEM) (2nd ed.). SAGE Publications.
Hite, D.M., Voelker, T., Robertson, A. (2014). Measuring Perceived Anonymity: The Development of a Context Independent Instrument. Journal of Methods and Measurement in the Social Sciences, 5(1), 22-39. https://doi.org/10.2458/azu_jmmss.v5i1.18305
John, O. P. (1991). The Big Five Inventory-Versions 4a and 54. University of California, Berkeley, Institute of Personality and Social Research.
John O.P., Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. Handbook of personality: Theory and research, 102-138
Lampropoulos, G., Anastasiadis, T., Siakas, K., Siakas, E. (2022). The impact of personality traits on social media use and engagement: An overview. International Journal on Social and Education Sciences, 4(1), pp. 34-51. https://eric.ed.gov/?id=EJ1332326
Le, T. D., Le, L., Phan, Q., Tran, K. T., Nguyen, P. & Liu, G. (2021). Participating anonymous online student communities and university brand relationship outcomes. Cogent Business & Management, 8(1). https://doi.org/10.1080/23311975.2021.1947558
Leijse, M. M. L., Koning, I. M. and van den Eijnden, R. J. J. M. (2023). The influence of parents and peers on adolescents' problematic social media use revealed. Computers in Human Behavior, 143, 107705. https://doi.org/10.1016/j.chb.2023.107705
Nitschinsk, L., Tobin, S. J., Varley, D., and Vanman, E. J. (2023). Why do people sometimes wear an anonymous mask? Motivations for seeking anonymity online. Personality and Social Psychology Bulletin. https://doi.org/10.1177/01461672231210465
Pan, X., Hou, Y., and Wang, Q. (2023). Are we braver in cyberspace? Social media anonymity enhances moral courage. Computers in Human Behavior, 148, 107880.
Peslak, A., Ceccucci, W., Sendall, P. (2012). An empirical study of social networking behavior using Theory of Reasoned Action. Journal of Information Systems Applied Research (JISAR), 5(3), pp. 12-23. https://scholarworks.merrimack.edu/mgt_facpub/19
Przybylski, A.K., Mruayama, K., DeHaan, C., Gladwell, V. (2013). Motivational, emotional, and behavioral correlates of fear of missing out. Computers in Human Behavior, 29(4), 1841-1848.
Rodriguez, R. L., Padilla, R.J., Montefalcon, M.D., Abisado, M. and Raga, R. (2023). The post behind anonymity: A thematic discourse analysis of Facebook posts from confession pages in different universities in the Philippines. In Proceedings of the 2023 11th International Conference on Information and Education Technology (ICIET), pp. 529–533, Fujisawa, Japan.
Sacks, B., Gressier, C. and Maldon, J. (2021). #REALTALK: Facebook confessions pages as a data resource for academic and student support services at universities. Learning, Media and Technology, 46(4), pp. 550–563. https://doi.org/10.1080/17439884.2021.1946559
Trivedi, S. K., Patra, P., and Singh, S. (2022). A study on intention to use social media in higher education: the mediating effect of peer influence. Global Knowledge, Memory and Communication, 71(1/2), 52–69. https://doi.org/10.1108/GKMC-11-2020-0169
Wang, W., Chen, R. R., and Yang, X. (2023). Examining compulsive use of social media: the dual effects of individual needs and peer influence. Industrial Management + Data Systems, 123(12), pp. 3109–3136. https://doi.org/10.1108/IMDS-10-2022-0631
Yu, Z., Song, X. (2021). User intention of anonymous social application "Soul" in China: Analysis based on an Extended Technology Acceptance Model. Journal of Theoretical and Applied Electronic Commerce Research, 16(7), 2898-2921. https://doi.org/10.3390/jtaer16070159
Yuan, D., Rahman, M. K., Issa Gazi, Md. A., Rahaman, Md. A., Hossain, M. M., and Akter, S. (2021). Analyzing of user attitudes toward intention to use social media for learning. SAGE Open, 11(4).

Confronting the Impact and Challenges of the AI Era: Teachers' Creative Use of Digital Tools to Enhance Digital Teaching Competence

Lin LI[a], Yan LIU[a*]
[a]*School of Education, Shanghai Jiao Tong University, China*
*flyliuyan0707@sjtu.edu.cn

Abstract: Teachers' digital teaching competency is key to navigating the educational transformation in the AI era. This study explores the factors influencing teachers' digital teaching competence by constructing multiple linear regression equations. The findings show that human-computer collaboration ability can significantly and positively explain 86.2% of the changes in digital teaching competence, making it the primary factor influencing digital teaching competence. In addition, this study considered the effects of teachers' gender, teaching experience, professional seniority, and pedagogical content knowledge and critical thinking on digital teaching competence. Combined with the results of the study, it is suggested that teachers should emphasize improving digital literacy and human-computer collaboration ability in the AI era, innovatively use digital tools during the teaching process, and deepen the understanding of pedagogical content knowledge and the cultivation of critical thinking, in order to better navigate the opportunities and challenges of the AI era.

Keywords: AI Era, Digital Tools, Digital Teaching Competence, multiple regression linear equations

1. Introduction

With the development of digital technology and the deep integration of artificial intelligence into education and teaching, we have entered the AI era of education (Ahrofi, 2025). In this context, the sustainable development of educational digitization and the transformation of education in the AI era drive teachers to master more solid digital teaching competence to address the impact and challenges of the AI era on the role of teachers (Rahmadi et al., 2021). Research has indicated that the integration of technology in teaching is a key factor in improving learning quality (McKnight et al., 2016). This requires teachers not only to possess a certain level of digital technology knowledge and skills but also to have the ability to seamlessly integrate technology into the classroom, employing innovative teaching strategies to achieve efficient human-computer collaboration in the classroom (Yang, 2024). Researchers have noted that there is a certain correlation between teachers' digital teaching competences and their ability to collaborate with digital technologies such as artificial intelligence (Kim, 2024). This means that when teachers can better collaborate with technologies like artificial intelligence, their digital teaching competences will also improve accordingly. In addition to focusing on teachers' human-computer collaboration capabilities, teachers' pedagogical content knowledge is equally important. According to the TPACK theory, when teachers use technology for teaching, their teaching effectiveness is also significantly influenced by their pedagogical content knowledge (Schmid et al., 2024). The use of digital technologies like AI can greatly facilitate teaching, but teachers who blindly use digital technologies in teaching may have negative teaching effects (Zhang et al., 2024). Therefore, critical thinking is a key ability for both teachers and students to adapt to digital transformation (Butler, 2024). The UNESCO-published “AI Competency Framework for Teachers” also

emphasizes that teachers should be able to critically evaluate the application of AI technology in teaching (Cukurova & Miao, 2024). Additionally, from the perspective of teachers' instructional skills, research by Fu Zhuqin and Cai Li (2022) indicates a significant positive correlation between teachers' critical thinking and effective classroom instruction, suggesting that teachers' instructional capabilities are influenced by their critical thinking skills. Other studies also indicate that critical thinking helps teachers flexibly address complex and dynamic teaching situations and resolve various new challenges encountered (Zhang Yin et al., 2018). Thus, teachers' critical thinking may also have a certain influence on their digital teaching capabilities.

Although the aforementioned studies have conducted a comprehensive and in-depth analysis of teachers' digital teaching competences from a theoretical perspective, there is a lack of corresponding empirical research to further validate the validity and applicability of these theoretical perspectives. To address this research gap, this study aims to construct a multiple linear regression model to explore the impact of teachers' human-computer collaboration abilities on their digital teaching competences, while also considering the roles of pedagogical content knowledge and critical thinking as two important factors.

2. Method

2.1 Participants

This study randomly selected primary and secondary school teachers in a certain district of S City, China, as the survey subjects. A total of 165 questionnaires were collected, of which 150 were valid, resulting in a validity rate of 90.91%. This sample is dominated by female teachers, a total of 120, accounting for 80%. In this sample, there were 27 (18.0%) teachers with less than 5 years of teaching experience, 31 (20.7%) with 6-10 years of teaching experience, 27 (18.0%) with 11-15 years of teaching experience, 15 (10.0%) with 16-20 years of teaching experience, and 50 (33.3%) with more than 20 years of teaching experience. In addition, in terms of the distribution of professional seniority, 49 (32.7%) teachers were elementary school level 1 teachers, 33 (22.0%) were elementary school level 2 teachers, 9 (6.0%) were senior elementary school teachers, 32 (21.3%) were secondary school level 1 teachers, 4 (2.7%) were secondary school level 2 teachers, 13 (8.7%) were senior secondary school teachers, and 10 (6.7%) had no titles.

2.2 Research Tools

In terms of scale design, the reference indicators for teachers' digital teaching competence (DTC), human-computer collaboration (HC), and critical thinking (CT) were borrowed from Zheng Xudong's (2019) digital competence model for primary and middle school teachers. The DTC dimension includes five items, with an example question being, “I can select appropriate digital resources for instructional design based on the development of students' core competencies”; the HCC dimension includes five items, with an example question being, “I can proficiently use AI tools to assist me in instructional activities”; and the CT dimension includes five items, with an example question being, “Before applying digital technology in instruction, I will thoroughly assess its compatibility with instructional activities.” The reference indicators for teachers' pedagogical content knowledge (PCK) primarily draw from the PCK-related items in the TPACK scale designed by Schmidt (2009) et al., totaling 5 items. An example question is: “I can select effective teaching methods to enhance students' core competencies”. The questionnaire was in the form of a five-point Likert scale, with scores from 1 to 5 representing “strongly agree”, “agree”, “generally”, “Disagree” and “Strongly Disagree”. After the samples were collected, the questionnaire was analyzed for reliability in this study, and its Cronbach's alpha coefficients for each dimension were greater than 0.8, with good reliability(Thanasegaran, 2009). Then, the questionnaire items were subjected to exploratory factor analysis using SPSS 26.0, and the resulting KMO statistic was 0.949, and the *p*-value of the Bartlett's test of sphericity was less than 0.05, which indicated good validity.

2.3 Data analysis

In this study, the correlation analysis was conducted through SPSS 26.0 by first using Pearson's coefficient to test whether there is a significant correlation between different genders (Gender), years of teaching experience (TE), professional seniority (PS), and teachers' human-computer collaborative ability, pedagogical content knowledge, and teachers' critical thinking on their ability to teach digitally at the level of 0.05, and then constructed a multiple linear regression model for analysis.

3. Results

3.1 Correlation analysis

From the results of the correlation analysis, the teachers' digital teaching competence in this study showed a significant negative correlation with the teachers' teaching age and title, indicating that the lower the teachers' teaching age and teaching title, the higher the teachers' digital teaching competence. There was a significant positive correlation between teachers' digital teaching competence and teachers' human-computer collaborative ability, subject-matter pedagogy knowledge and critical thinking, indicating that the higher the human-computer collaborative ability, subject-matter pedagogy knowledge, and critical thinking, the higher the teachers' digital teaching competence. In addition, there was no correlation between teachers' titles and their pedagogical content knowledge and critical thinking. There was no correlation between teachers' gender and all variables.

Table 1. Analysis of correlation results

	DTC	HCC	PCK	CT	TE	PS	Gender
DTC	1.000	.929**	0.808**	0.848**	-0.267**	-0.200*	-0.107
HCC	0.929**	1.000	0.720**	0.797**	-0.309**	-0.209*	-0.060
PCK	0.808**	0.720**	1.000	0.692**	-0.139	-0.150	-0.133
CT	0.848**	0.797**	0.692**	1.000	-.0229**	-0.122	-0.108
TE	-.267**	-.309**	-0.139	-.229**	1.000	0.151	-0.088
PS	-.200*	-.209*	-0.150	-0.122	0.151	1.000	-0.082
Gender	-0.107	-0.060	-0.133	-0.108	-0.088	-0.082	1.000

**p<0.01, *p<0.05

3.2 Multiple linear regression analysis

The results of the above analysis show that human-computer collaborative competence, subject pedagogy knowledge and critical thinking are all significantly related to teachers' digital teaching competence. In this study, teachers' digital teaching competence was taken as the dependent variable, human-computer collaborative competence was taken as the independent variable, and considering that teachers' pedagogical content knowledge and critical thinking as well as teachers' years of teaching experience and professional title may have some interfering effects on the model, they were also taken as the control variables, and the multivariate linear regression model was constructed as follows:

$$DTC = a + b_1HCC + b_2PCK + b_3CT + b_4TE + b_5PS$$

Where a is the constant, b_1, b_2, b_3, b_4, b_5 are the regression coefficients. After testing, the residual sequence of the regression model passed the autocorrelation and normality tests: the Durbin-Watson statistic was 2.083, indicating that the residuals did not exhibit significant first-order autocorrelation; the Shapiro-Wilk normality test confirmed that the residuals were normally distributed (p > 0.05). Additionally, multicollinearity diagnostics show that the variance inflation factor (VIF) for all variables is less than 5, indicating that the model does not have serious multicollinearity issues. The regression results are shown in Table 2.

Table 2. *Analysis of MLR results*

Variables	Model 1	Model 2	Model 3
（constant）	0.564**	-0.048	-0.015
HCC	0.848**	0.534**	0.529**
PCK		0.207**	0.209**
CT		0.257**	0.257**
TE			-0.001
PS			-0.006
R2	0.862	0.919	0.919
△R2	0.861	0.917	0.917
F	927.163**	549.472**	326.321**

**p<0.01, *p<0.05

Based on the regression results, we found that all three regression models passed the test at the 0.01 significance level. In Model 1, which contains only the independent variable human-computer collaborative ability, the value of R^2 is 0.862, indicating that teachers' human-computer collaborative ability explains 86.2% of the variation in digital teaching competence and is the main factor influencing digital teaching competence. With the addition of teachers' pedagogical content knowledge and critical thinking, the model overall explains 91.9% of teachers' digital teaching competence, indicating that teachers' human-computer collaborative ability, pedagogical content knowledge, and critical thinking are able to cover most of the reasons for the changes in teachers' digital teaching competence. However, we found that after introducing teachers' pedagogical content knowledge and critical thinking, the coefficient of explanation for human-computer collaboration ability decreased from 84.8% to 53.4%. Combined with the results of the correlation analysis, this indicates that teachers' pedagogical content knowledge and critical thinking explain part of the impact of human-computer collaboration ability on digital teaching competence, thereby correcting our previous overestimation of the explanatory power of human-computer collaboration ability. This also indicates that teachers' pedagogical content knowledge and critical thinking skills will also affect their human-computer collaboration abilities. In Model 3, we further introduced teachers' teaching experience and professional title as control variables based on the results of the correlation analysis, but the overall explanatory degree of the model did not change significantly, and there was no significant linear relationship between these two variables on teachers' digital teaching competence. Combined with the correlation results, teaching experience and professional title may explain a small portion of the effect of the above three variables on the dependent variable, but this effect is not significant.

4. Discussions and implication

Combining the results of the above data analysis, we conclude that teachers' human-computer collaboration ability is the main factor affecting their digital teaching ability, and that when teachers' human-computer collaboration ability is stronger, their pedagogical content knowledge is richer, and their critical thinking is stronger, their digital teaching ability is correspondingly higher. This finding suggests that if we want to improve teachers' digital teaching ability, we should start from strengthening teachers' human-computer collaboration ability and provide them with targeted training, and at the same time pay attention to strengthening teachers' pedagogical content knowledge and critical thinking.

In addition, we found that although teachers' teaching experience and professional title have some correlation on teachers' digital teaching ability, they do not have a significant linear relationship on the dependent variable, teachers' digital teaching ability, in the multiple linear regression model. This indicates that although teachers' years of teaching experience and professional title have some effect on their digital teaching competence, this effect is not significant. It proves that teachers' digital teaching competence is mainly related to their own

competence, knowledge accumulation and personal attributes. Further, combining the results of the correlation analysis, we find that teachers' knowledge accumulation is not directly related to their traditional teacher characteristic factors such as gender, teaching experience and professional title, which breaks certain stereotypes or assumptions that may have existed in the past, and emphasizes the plurality and independence of teachers' knowledge accumulation. We also found a significant negative correlation between teachers' human-computer collaboration skills and critical thinking and teaching experience, suggesting that younger teachers may be more active and adept at using digital technology. This may be due to the fact that young teachers grew up in an era of rapid development of digital technology, and they are more likely to accept and master new technologies, thus showing an advantage in human-computer collaborative skills. In addition, younger teachers may also be more accustomed to questioning and reflecting, thus showing greater competence in critical thinking as well. However, this advantage requires a more dialectical perspective: on the one hand, technical proficiency is indeed often associated with youth, but truly effective “human-computer collaboration” lies in the educational wisdom of deeply integrating technology with pedagogy, which often requires accumulated experience; on the other hand, critical thinking is not merely about questioning habits but also involves a profound understanding of the educational value, potential risks, and ethical implications of technology application, which may precisely be the unique perspective that experienced teachers can offer based on their rich practical experience. Therefore, the vitality of young teachers and the experience of older teachers are not mutually exclusive; intergenerational collaboration and complementarity are essential to fully unlock the educational value of human-machine collaboration.

Therefore, we make the following recommendations: in order to enhance teachers' digital teaching ability in the digital era, in addition to focusing on their teaching ability, teachers' personal development training should be emphasized. The technical advantages and innovative thinking of young teachers should be fully utilized, and they should be encouraged to actively use digital technology in their teaching practice, promote the innovation of teaching methods, and explore the teaching mode of human-computer collaboration. Meanwhile, for senior teachers, it is also necessary to continuously enhance their abilities in digital technology and critical thinking through continuous professional training and learning to meet the needs of educational development. Teachers' knowledge accumulation is a complex and multifaceted process that may be affected by a variety of factors, including individual learning habits, teaching environment, and professional development opportunities. This finding has important implications for education policy makers, school administrators, and teachers themselves, suggesting that we should pay more attention to teachers' individual differences and professional development needs, and provide them with diversified learning and development opportunities to promote their knowledge accumulation and professional growth.

Acknowledgements

This research is supported by National Social Science Fund of China 2023 National General Project in Education Science Planning: "Accompanied Data-Driven Diagnosis and Cross-Contextual Enhancement of Teachers' Digital Teaching Capabilities" (BCA230281).

References

Ahrofi, S. (2025). Optimization Of Teachers' digital Competence In Facing Artificial Intelligence (AI) Based Learning Transformation In The Era Of Society 5.0. In *The Fourth International Conference on Government Education Management and Tourism* (Vol. 4, pp. 104-104).

Butler, H. A. (2024). Predicting everyday critical thinking: A review of critical thinking assessments. *Journal of Intelligence, 12*(2), 16.

Cukurova, M., & Miao, F. (2024). *AI competency framework for teachers*. UNESCO Publishing.

Kim, **hee. "Leading teachers' perspective on teacher-AI collaboration in education." *Education and Information Technologies 29.7* (2024): 8693-8724.

McKnight, K., O'Malley, K., Ruzic, R., Horsley, M. K., Franey, J. J., & Bassett, K. (2016). Teaching in a Digital Age: How Educators Use Technology to Improve Student Learning. J*ournal of Research on Technology in Education, 48*(3), 194–211. https://doi.org/10.1080/15391523.2016.1175856

Rahmadi, F., Munisa, M., Rozana, S., Rangkuti, C., Ependi, R., & Harianto, E. (2021). Development of integrated school management based on information systems in North Sumatra. Fitrah: *Journal of Islamic Education, 2*(2), 96–109. https://doi.org/10.53802/fitrah.v2i2.64

Schmid, M., Brianza, E., Mok, S. Y., & Petko, D. (2024). Running in circles: A systematic review of reviews on technological pedagogical content knowledge (TPACK). *Computers & education, 214*, 105024.

Schmidt, D. A., Baran, E., Thompson, A. D., Mishra, P., Koehler, M. J., & Shin, T. S. (2009). Technological pedagogical content knowledge (TPACK) the development and validation of an assessment instrument for preservice teachers. *Journal of Research on Technology in Education, 42*(2), 123-149.

Thanasegaran, G. (2009). Reliability and Validity Issues in Research. I*ntegration & Dissemination, 4,* 35–40.

Yang, H. (2024). Research on the Diversification and Adaptability of College English Teachers' Roles under the Trend of Human-Computer Collaborative Education. *Contemporary Education Frontiers, 2*(1), 108-115.

Zhang, S., Jaldi, C. D., Schroeder, N. L., López, A. A., Gladstone, J. R., & Heidig, S. (2024). Pedagogical agent design for K-12 education: A systematic review. *Computers & Education, 223,* 105165.

Zhang, Yin, Luo, Xingkai & Zhang, Hongxia. (2018). A Study of Science Teachers' Critical Thinking Dispositions and Classroom Behaviors. *Global Education Perspectives* (08), 59-68.

Zheng, X.-D. (2019). *Research on the construction and application of digital competency model for primary and secondary school teachers in China* (Doctoral dissertation, East China Normal University). PhD.

Teaching Machine Learning to Senior Primary Students: Evaluating a Course for AI Literacy in a Hong Kong Primary School

Siu Cheung KONG[a,b] *, Yin YANG[a] & So Sum CHOW[b]
[a]*Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong SAR*
[b]*Artificial Intelligence and Digital Competency Education Centre, The Education University of Hong Kong, Hong Kong SAR*
**sckong@eduhk.hk*

Abstract: This study examines the effectiveness of a machine-learning course designed to enhance artificial intelligence (AI) literacy among 145 sixth-grade students in a Hong Kong primary school. The six-hour course introduced foundational AI concepts, including supervised learning (artificial neural networks and K-nearest neighbours) and reinforcement learning, with hands-on activities featuring AlphAI robots. Quantitative analysis, supported by Wilcoxon signed-rank tests, revealed significant improvements in students' conceptual understanding of machine learning and deep learning. Qualitative feedback demonstrated the course's effectiveness, with hands-on activities using AlphAI robots proving effective in deepen comprehension and enabling practical application of machine learning methods. The findings show how the course demystify AI for young learners, encouraging critical engagement with it's "black box" nature. These insights emphasise the value of interactive, hands-on learning strategies in AI literacy education in preparing young students for an AI-driven future.

Keywords: AI literacy education; deep learning; machine learning; primary school, robotics

1. Introduction

The rapid rise of generative AI—a subset of machine learning capable of creating novel content—has reshaped how humans interact with technology. A prominent example is Generative Pre-trained Transformer (GPT), a class of large language models built on the Transformer architecture and attention mechanisms, enabling breakthroughs in natural language processing. Unlike traditional rule-based AI systems, generative AI models like GPT can use deep learning (multi-layered artificial neural networks) to produce human-like text, images, and other media, demonstrating machine learning's transformative potential. This progress allows AI to augment human expertise and solve complex problems across industries. Yet, despite its growing influence, the public often perceives machine learning—especially generative AI—as a mysterious "black box" (Tedre et al., 2021). The process, from user prompts to internal computations and final outputs, remains opaque to most people, leading to an unspoken misplaced fear for artificial intelligence and automation. This knowledge gap is especially concerning in education, where students should develop a foundational understanding of "how AI functions" to thrive in an increasingly AI-driven future (Royal Society, 2017).

2. Literature Review

2.1 Current Approaches of AI Literacy Education for Young Students

Recent advancements in AI literacy education recognise AI literacy as a vital skill for future generations, leading to the development of innovative programmes across primary and secondary education levels (Kong & Yang, 2023; Kong & Yang, 2024; Martin et al., 2023; Martin et al., 2024; Maspul, 2024).

Customised learning design focuses on developmental appropriateness and simplicity, ensuring young students can grasp complex AI concepts. Research indicates that interactive methods are effective in introducing machine learning concepts to younger audiences (Gibellini et al., 2023). Hands-on learning approaches have proven to be particularly effective across various age groups. Robot-assisted programmes are especially promising for primary learners (Hong & Kim, 2025).

2.2 Gaps in Current Research

Despite these advances, significant research gaps remain in the field of AI literacy education for young students. First, educational robots remain under-researched despite their potential to make AI concepts more transparent than conventional "black box" tools (Yue et al., 2022). The intuitive interface and visualisation capabilities of AlphAI learning robots with its software used in this study aim to address this gap, offering an innovative approach to demystifying AI processes for young learners (Absalon & Deneux, 2025; Yang & Kong, 2025).

Second, most AI literacy framework targets secondary and other students, with limited attention given to senior primary students (Kim et al., 2021). By focusing on sixth-grade students, this study seeks to address this oversight, providing a teaching practice that progresses from foundational artificial neural networks (ANN) to reinforcement learning while maintaining conceptual rigor. This focus on younger learners is crucial for developing early interest and understanding in AI, setting the groundwork for advanced study in later educational stages.

Despite AI's growing societal prominence, current educational practices often fail to provide effective instruction for developing AI literacy, especially among younger students. This gap highlights the urgent need for engaging curricula that can introduce core AI concepts while fostering hands-on experience. Building upon prior research that examined school teachers' perceptions of a machine learning course's effectiveness (Yang & Kong, 2025), the current study investigates the impact of the machine learning course—enhanced by AlphAI learning robots—on students' conceptual understanding and their learning experiences. Specifically, the study addresses the following research questions:

(1) To what extent did the machine learning course, incorporating AlphAI robots, enhance sixth-grade students' conceptual understanding of machine learning?
(2) How did students perceive the impact of the course on their conceptual understanding of machine learning?

3. Research Design

3.1 Course Structure and Implementation

The six-hour curriculum introduced sixth-grade students to fundamental machine learning concepts through structured sessions combining theoretical instruction and practical activities. Professionally trained teachers guided students through progressive learning activities using course worksheets and AlphAI robots to ensure comprehensive understanding (Yang & Kong, 2025).

3.1.1 Instructional Approach

The curriculum covered core machine learning and deep learning concepts, with particular emphasis on hands-on learning experience through robot training exercises to reinforce conceptual understanding while developing practical skills. A key instructional focus was clarifying both the relationship between AI, machine learning, and deep learning, and the distinct roles of fundamental concepts within this framework, as visualised in Figure 1.

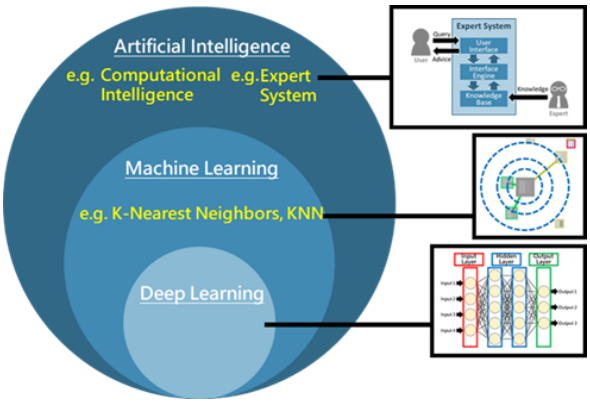
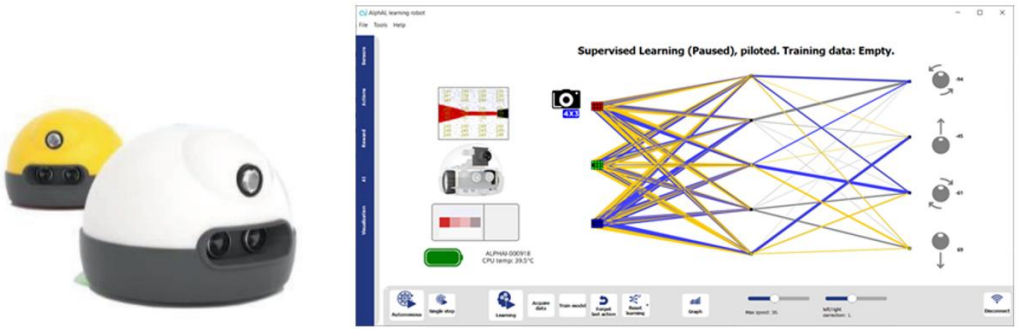


Figure 1. The relationship among commonly used terms of artificial intelligence, machine learning, and deep learning.

In particular, the course introduced foundational concepts beginning with ANNs—the basis of deep learning architectures—and supervised learning, a core machine learning method utilising labelled datasets to train predictive models. Subsequent units gradually progressed to advanced topics, including the role of weights and backpropagation in ANNs, reinforcement learning, and KNN, a machine learning algorithm distinct from ANN-based approaches. The course culminated in exploring the connections between core machine learning concepts—including the methods covered—and real-world applications, such as Google's Teachable Machine, which leverages GPT-like models.

3.1.2 Hands-On Learning Experience with AlphAI Robots

Hands-on activities were integrated into each session using AlphAI learning robots (Martin et al., 2023), providing students with experiential learning opportunities. These robots (Fig. 2(a)), connected to AlphAI software (Fig. 2(b)), offered a graphical interface that visualised AI algorithms like ANNs, Reinforcement Learning, and KNN. Students engaged in experiments such as training robots to race, modifying pixel inputs and movement outputs, and comparing learning algorithms, which helped reinforce abstract concepts.



(a). The AlphAI learning robots (b). Visualisations of algorithm in the form of an artificial neural network on screen (i.e. supervised learning)

Figure 2. The AlphAI Learning Robots and Software Interface.

3.2 Participants

The study utilised a convenience sample of 145 sixth-grade students (94 boys, 51 girls; ages 10-13, *M*=10.94, *SD*=0.530) from five classes at a primary school. Class sizes ranged from 23 to 37 students. All participants had approximately two years of prior programming experience. Following ethical guidelines, we obtained written informed consent from both the school administration and parents.

3.3 Data Collection and Analysis

The study used a mixed-methods approach, collecting both quantitative and qualitative data: (1) pre- and post-tests on machine learning and deep learning concepts and (2) focus group interviews with 15 students. The 13-item concept test assessed understanding of machine learning and deep learning, showing moderate reliability (Cronbach's alpha = 0.656). Focus group interviews captured detailed student experiences with the machine learning course. Five focus groups of three students each, totalling 15 participants, were facilitated by a trained researcher for approximately 30 minutes. To address the first research question, a Wilcoxon signed-rank test was conducted to analyse pre- and post-test score differences in conceptual understanding. For the second question, thematic analysis explored qualitative feedback from focus group interviews.

4. Results

4.1 Analysis of Students' Conceptual Understanding of AI Literacy

The Wilcoxon signed-rank test revealed a significant improvement in students' AI literacy, with an increase from pre-test total scores (M = 5.03, SD = 2.072) to post-test total scores (M = 7.07, SD = 2.790), Z = 5.860, p < 0.001. Table 1 shows the descriptive data of each concept item. These results underscore a substantial enhancement in students' AI conceptual understanding following course completion. Gender-based analysis using the Mann-Whitney U test found no significant differences in score improvements between boys and girls (U = 1769.50, p = .365).

Table 1. *The Descriptive Data of Machine Learning and Deep Learning Concepts*

	Pre		Post		
	M	*SD*	*M*	*SD*	*Z*
Item 1	.31	.46	.56	.50	-4.185***
Item 2	.25	.44	.59	.49	-5.742***
Item 3	.69	.46	.76	.43	-1.050
Item 4	.48	.50	.63	.49	-2.121*
Item 5	.22	.41	.53	.50	-4.672***
Item 6	.43	.50	.54	.50	-2.018*
Item 7	.40	.49	.49	.50	-1.638
Item 8	.40	.49	.71	.46	-4.965***
Item 9	.29	.46	.35	.48	-.802
Item 10	.45	.50	.60	.49	-2.357*
Item 11	.48	.50	.52	.50	-.762
Item 12	.17	.38	.28	.45	-1.938
Item 13	.36	.48	.51	.50	-2.393*

*p < .05, **p < .01, ***p < .001

Significant improvements were observed in eight assessed items, particularly in understanding machine learning fundamentals. The results revealed enhanced comprehension of general machine learning processes, supervised learning, and specific algorithms including KNN and ANNs (including layer structures and neurons). Students also developed better awareness of backpropagation as a weight-adjustment mechanism and recognized overfitting as a common training challenge. The most notable improvement appeared in understanding backpropagation, with significantly more students correctly identifying it as a process where output errors are fed back to adjust neuron weights in

hidden and input layers to optimize performance. These findings confirm the course's effectiveness in teaching core machine learning concepts.

4.2 Student Perceptions of the Course's Impact on Their Conceptual Understanding

Thematic analysis of five focus group interviews revealed three key themes about students' learning experiences: (1) enhanced conceptual understanding of machine learning, (2) valuable hands-on experience with AlphAI learning robots, and (3) recognition of real-world machine learning implementation challenges.

All 15 students reported benefiting from the course's structured approach. Many used analogies to explain concepts, like one student (ID9) comparing methods to classroom scenarios: "*Supervised learning is like a teacher guiding a student with labels, reinforcement learning uses rewards like in school, and unsupervised learning is when the robot explores alone—no labels, no teacher.*" Others demonstrated sophisticated insights, with a student (ID8) noting machine learning's data dependence: "*Machine operate by directly processing correct and incorrect knowledge (well- and poor-quality data) to achieve their purpose, but it needs to be given data to analyse (by human).*"

Students particularly valued AlphAI robots' physical interactivity. A student (ID15) contrasted passive lectures with active experimentation: "*Listening to the teacher explaining the concepts would be very abstract but actually doing it and testing it out would give a deeper impression.*" The iterative trial-and-error learning process was also a recurring theme in the interviews, with four students emphasising its educational value.

Practical challenges emerged during activities, particularly regarding data quality. A student (ID10) emphasised the critical role of high-quality training data: "*When we are making a model of ourselves with Teachable Machine, we need to take pictures of rock, paper, and scissors (gestures) from different angles (as inputs) to train the model to recognise them. However, the problem is that we may not be shooting from the right angle, and the model will not recognise my gestures accurately when it is trained and tested.*"

Most students expressed strong satisfaction with the course. A student (ID13) summarised: "*Training models and competing was fun and memorable—I even thought about how to optimise them at home."* These findings demonstrate the course's success in making machine learning accessible while fostering critical understanding of its real-world constraints.

5. Conclusion and Implication

This study demonstrated the effectiveness of a machine learning course incorporating AlphAI robots in enhancing AI literacy among sixth-grade students. The quantitative analysis, notably the Wilcoxon signed-rank test, revealed significant improvements in students' understanding of machine learning and deep learning concepts. Additionally, the qualitative insights from focus group discussions underscored the importance of hands-on learning experience in demystifying complex AI concepts for young learners.

These findings have important implications for AI literacy education in the K-12 setting, particularly for young students. The results indicate that a well-designed curriculum that balances theoretical knowledge with practical application can enable students aged 10 to 13 to grasp fundamental machine learning concepts and critically evaluate the training process. To uncover the "black box" nature of machine learning, educational frameworks should emphasise scaffolding complexity and fostering active reflection among students.

However, the generalisability of this study is limited by its small sample size. Future research should consider expanding the sample size and including more diverse educational settings to validate and extend these findings. Additionally, extending instructional time and refining educational materials could further improve comprehension and engagement. By overcoming these limitations, future studies can further improve AI literacy courses for young learners, equipping them with the skills and critical insight necessary to thrive in an increasingly AI-driven world.

Acknowledgements

The authors would like to acknowledge the funding from the Central Reserve Allocation Committee, The Education University of Hong Kong of the Hong Kong Special Administrative Region, China (Uncovering the "Black Box" of Machine Learning: Promoting Artificial Intelligence Literacy with AlphaAI robots in Senior Primary/Junior Secondary Schools across Hong Kong and France). (Project No. EdUHK 04A55).

References

Absalon, M., & Deneux, T. (2025, April). “AlphAI”: Teaching AI algorithms to K12 by training learning robots and visualizing how it works. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39, No. 28, pp. 29096-29103). https://doi.org/10.1609/aaai.v39i28.35181

Gibellini, G., Fabretti, V., & Schiavo, G. (2023, April). AI education from the educator’s perspective: Best practices for an inclusive AI curriculum for middle school. In *Extended abstracts of the 2023 CHI conference on human factors in computing systems* (pp. 1-6). https://doi.org/10.1145/3544549.358574

Hong, J., & Kim, K. (2025). Impact of AIoT education program on digital and AI literacy of elementary school students. *Education and Information Technologies, 30*(1), 107-130. https://doi.org/10.1007/s10639-024-12758-0

Kim, S., Jang, Y., Kim, W., Choi, S., Jung, H., Kim, S., & Kim, H. (2021, May). Why and what to teach: AI curriculum for elementary school. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 17, pp. 15569-15576). https://doi.org/10.1609/aaai.v35i17.17833

Kong, S. C, & Yang, Y. (2023, December). Designing and evaluating an Attention-Engagement-Error-Reflection (AEER) approach to enhance primary school students’ artificial intelligence literacy and learning-to-learn skills: A pilot study. In W.Q. Chen, et al. (Eds.), *Workshop Proceedings of the 31st International Conference on Computers in Education ICCE 2023* (pp. 394-399). Matsue, Shimane, Japan: Asia-Pacific Society for Computers in Education. https://doi.org/10.58459/icce.2023.1417

Kong, S. C., & Yang, Y. (2024, May). Using the robot-assisted attention-engagement-error-feedback-reflection (AEER) pedagogical design to develop machine learning concepts and facilitate reflection on learning-to-learn skills: Evaluation of an empirical study in Hong Kong primary schools. In *Proceedings of the 16th International Conference on Computer Supported Education - Volume 2* (pp. 155-162). Angers, France. https://doi.org/10.5220/0012505700003693

Martin, M., Chevalier, M., Burton, S., Bonvin, G., Besançon, M., & Deneux, T. (2023, April). Effects of introducing a learning robot on the metacognitive knowledge of students aged 8–11. In *International Conference on Robotics in Education (RiE)* (pp. 169-183). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-38454-7_15

Martin, M., Deneux, T., & Chevalier, M. (2024, April). From automaton to AI robot: The added value for learning. In *International Conference on Robotics in Education (RiE)* (pp. 403-410). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-67059-6_35

Maspul, K. A. (2024). Using interactive language development tools to enhance cognitive and literacy skills in K-12 education. *J-SHMIC: Journal of English for Academic, 11*(1), 20-35. https://doi.org/10.25299/jshmic.2024.vol11(1).15835

Royal Society. (2017). *Machine Learning: The Power and Promise of Computers That Learn by Example: An Introduction.* Royal Society.

Tedre, M., Toivonen, T., Kahila, J., Vartiainen, H., Valtonen, T., Jormanainen, I., & Pears, A. (2021). Teaching machine learning in K–12 classroom: Pedagogical and technological trajectories for artificial intelligence education. *IEEE access, 9*, 110558-110572. https://doi.org/10.1109/ACCESS.2021.3097962

Yang, Y., & Kong, S. C. (2025, April). Teaching machine learning for senior primary and junior secondary students: A case study of teacher professional development in Hong Kong. In *Proceedings of the 17th International Conference on Computer Supported Education* - Volume 2: CSEDU, (pp. 35-42). Porto, Portugal. https://doi.org/10.5220/0013200300003932

Yue, M., Jong, M. S. Y., & Dai, Y. (2022). Pedagogical design of K-12 artificial intelligence education: A systematic review. *Sustainability, 14*(23), 15620. https://doi.org/10.3390/su142315620

Co-Designing AI and Multimodal Analytics Integrated Reflective Learning in Japanese High School Context

Rwitajit MAJUMDAR[a*], Shin-Ichiro KUBOTA[a] & Huiyong LI[b]
[a]*Research and Educational Institute for Semiconductor and Informatics, Kumamoto University, Japan*
[b] *Research Institute for Information Technology, Kyushu University, Japan*
*majumdar@kumamoto-u.ac.jp

Abstract: This research focuses on a codesign effort to integrate AI and Multimodal Learning Analytics technology and develop pedagogy for reflective learning practices. Collaborating with stakeholders from a Japanese educational con-text, researchers, students, teachers, and educational board policymakers focused on integrating Generative AI policies placed by the Ministry of Education in Japan into the activity design and the learning platform functions. The study analyzed the artifacts from the 6 participatory design sessions. These artifacts include learning activities designed by teachers and their iterations with researchers, articulation of user needs and design ideation from a co-design session conducted with students and teachers, and policy alignment of the technology and activity designed during the pilot phase. While the study demonstrates co-design steps for integrating frontier learning support tools in the Japanese context, the findings are relevant globally, and the study discusses how to balance instructional goals and technology updates in the current educational realities.

Keywords: Co-design, Generative AI, Multimodal Learning Analytics, K-12

1. Introduction

With the rise of the artificial intelligence (AI) paradigms, numerous policies and strategies for institutional adoption have been actively discussed (Ifenthaler, Majumdar et al, 2024). In Japan, the Minis-try of Education has recently also introduced policies to guide the integration of Gen-erative AI into K-12 education, prompting schools to align their technologies and pedagogies accordingly (MEXT, 2024). Many schools, already equipped with ICT infrastructure through initiatives like Google School, are well-positioned to adopt various online learning practices. However, teachers still face challenges in redesigning existing learning activities to effectively incorporate AI technologies, particularly within the time constraints of their daily schedules. Despite these challenges, there is a growing focus on digital transformation within schools, where teachers show a willingness to adopt ICT solutions that streamline tasks, gather evidence of good practice and simplify reporting and sharing. Equally important is addressing the needs of learners in this evolving educational landscape. However, to fully harness benefits of AI and multimodal learning analytics, learning activities and systems must be designed with a deep understanding of students' cognitive, emotional, and social needs. This ensures that AI-enhanced education not only supports academic achievement but also fosters critical thinking, collaboration, and digital literacy skills essential for future-ready learners.

Currently eLearning platforms have also introduced newer workflows to integrate generative AI-driven features (Arghir, 2024). While many such AI are large language model based, other forms of AI models for instance computer vision-based are still less explored. It also remains challenging to design a balanced instructional support system for reflective learning activities (Oviatt, 2018), ensuring compliance with existing policies and carefully considering how and when such AI systems can be implemented in practice. This underscores the need for strong collaboration between end-users, researchers and developers to create effective and policy-aligned solutions.

Against this backdrop, this research aims to explore the following objective: How can we design learning activities and systems that effectively leverage AI-driven features while meeting the needs of both educators and policy frameworks?

2. Literature Review

2.1 Co-designing Learning Analytics

Learning analytics is an interdisciplinary field that seeks to leverage data, system design, and educational practices to enhance teaching and learning experiences. While some efforts stem from a computational perspective, there are existing studies which deploy co-design approaches to design and develop learning analytics solutions (Sarmiento & Wise, 2022, Ogata et al. 2024). While much of the early focus has been on online settings, there has been a growing interest in integrating learning analytics into face-to-face classrooms at both higher education and school levels. These initiatives often adopt a human-centered approach, prioritizing the needs and practices of educators and learners over purely technology-driven solutions. Large-scale projects in Japan has also employed co-design approach-es (Ogata et al. 2024), fostering collaboration between educators, students, industrial partners and policy maker to ensure that learning analytics tools align with pedagogical goals and classroom realities.

2.2 Co-designing with Novice Users

Participatory design with novice users is also a challenge as their latent needs often are difficult to elicit. Lepe-Salazar et al. (2025), considered students, beginners, and enthusiasts as novices of computer science and software design who often struggle with the complexities of software design, particularly during the early stages of gathering and analyzing requirements. Mirror, is a participatory design approach that actively involves such novice stakeholders in shaping software elements. By providing structured yet flexible tools (Mirrors), the method enables participants to document their thoughts, generate ideas, and collaboratively refine design concepts. Stakeholders express their opinions, needs, and expectations through structured prompts presented as three mirrors. The First Mirror captures initial thoughts and real-world challenges, ensuring the design process starts with user realities. Participants actively contribute to generating new design components in Second Mirror. Stakeholders engage in structured self-reflection and group discussions to refine ideas in Third Mirror.

The method ensures that proposed elements align with user goals and system usability. Participants can modify, adjust, and re-evaluate their contributions in a dynamic, iterative manner. Designers act as facilitators rather than sole creators, empowering users to take an active role in shaping their digital experiences.

3. Research Context and Learning Environment

The High School. The research context involves stakeholders from a high school located in Southwest Japan. To facilitate the coordination of this project, a smaller working group was established through collaboration between the school principal and the lead teacher. The group included both native Japanese teachers and foreign English language instructors, such that it represents the diverse population that coordinates the Japanese public educational context. The coordinating teachers play a central role in aligning the project with the school's educational objectives and daily practices.

The school comprises three grade levels, each consisting of approximately nine classes with around 40 students per class. In addition to their regular academic activities, students participate in various extracurricular clubs. The teachers approached the chemistry club members to participate in the pilot activity.

Representatives from the prefectural educational board, particularly those involved in AI policy-making, also participated in the project. Their involvement ensured that the research aligned with broader educational policies and strategies concerning the integration of AI technologies in schools.

The Learning Environment. LA-ReflecT, a microlearning platform (Majumdar et al. 2023) with analytics enabled infrastructure was adopted for this study. LA-ReflecT enables teachers and learners engage in data-informed reflection on their practices. Teachers can use LA-ReflecT as an external tool integrating it with their own LMS through the LTI proto-col. They can author multimedia micro learning activities and publish it in their course. The learner can attempt the tasks in the activity and the learning design navigates them to a learning analytics dashboard immediately after attempting a task for reflecting on the learning episode. A recently enabled AI and the multimodal analytics feature supported integrated ChatGPT API and a web-based eye tracking. Figure 1 shows an overview of the teacher and learner workflow in the LA-ReflecT platform.

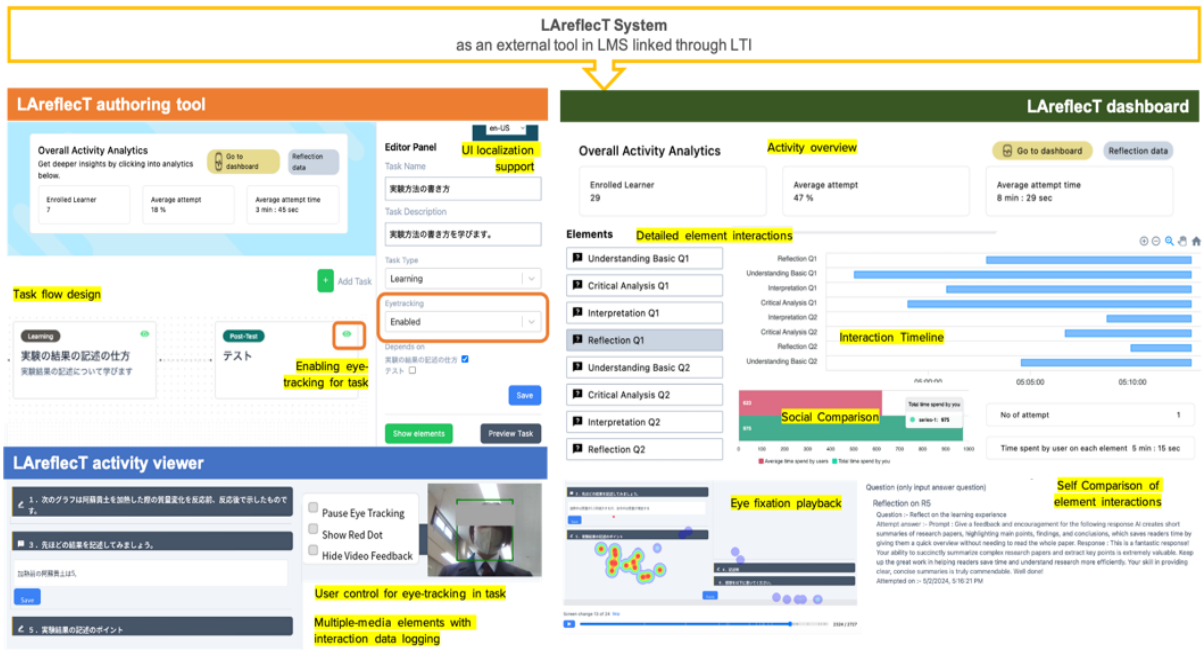


Figure 1. LA-ReflecT platform's features used in this study

4. User Study

4.1 Core Agenda of the Co-design Sessions

This research follows the participatory design iterations proposed in and for phase 1 of motivating and defining the seed design, adopted the use case regarding integrating generative AI (Majumdar, Singh & Chen, 2024) and distributed learning analytics (Majumdar, Singh, Rajendran, Narayan & Gatare, 2024) that was already studied in the higher education context. Based on that we approached the school and conduct-ed 6 co-design sessions to initiate the pilot (phase 2). The various artefacts generated during these sessions are analyzed to articulate the design process that was followed.

The six sessions involved researchers, teachers, students, and the educational board representatives. The first session (22 Sep, 2024 – in person with researchers, school management) introduced the project's objectives and the learning tools, while also reviewing the school's IT infrastructure and the consent procedures for student participation. In the second session (12 Dec, 2024 – online with researchers and school management), researchers discussed the ethics board's approval and the implementation of consent processes, including offline parent consent forms and an opt-out option for data logging. A

Moodle site was shared, featuring the learning tool and a demonstration video. An initial activity integrating ChatGPT into LA-ReflecT was introduced, and teachers were encouraged to create their own activities. In the third session (8 Jan, 2025 – online), three teachers presented activities on information security, chemical experimental methods, and friction. The research team reviewed these and suggested adding feedback prompts through ChatGPT. Eye-tracking and dash-board features were also demonstrated. Five chemistry club students were recruited for a pilot study. The fourth session (23 Jan, 2025 – hybrid) had students complete one activity in-session, with two others done independently. In the fifth session (5 Feb, 2025 – online), the research team reviewed the pilot with school staff and educational board members, discussing policy alignment and feedback strategies. The final session (14 Feb, 2025 – in person) was a design workshop using the Mirrors tool (Lepe-Salazar et al., 2025) to explore reflective learning and AI integration.

4.2 Insights from the Co-design Sessions

Elicit needs of AI and multimodal analytics integration with value proposition.

Researchers and teachers interacted to iterate the three original activities and discussed the inclusion of AI and multimodal analytics support. We present the core value proposition that emerged in reference to two of those activity redesign.

For the activity on the topic of friction, there was a learning and a posttest task. The learning task had multiple resources such as interactive physics simulations available on PhET (phet.colorado.edu), content reflection prompts that was added by the teacher as elements in LA-ReflecT. The teacher included a particular roller skating track design as a problem to then discuss various concept of potential and kinetic energy. In the post test, a similar problem with a new track design was given. Here the teachers included AI-based hints after the first round of discussion with the researchers. They wanted to investigate the differences when the hint is presented as a feedback versus as a chatbot (as presented in Figure 2).

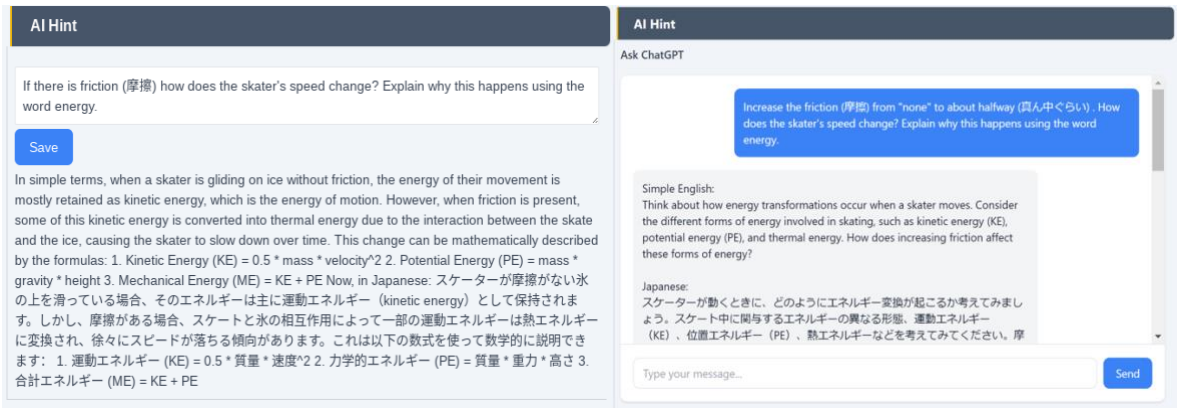


Figure 2. Implementation of AI hint as a hint and chatbot

Value proposition: To enhance existing instructional design, incorporating teachers' interests in reflecting on their practices creates opportunities for integrating innovative tools and services. In this case, the use of generative AI as a hint-providing tool allowed the teacher to explore various interaction modalities and examine how different prompting strategies influence AI-generated responses. This approach fosters deeper engagement and supports continuous instructional improvement.

Another activity on chemical experiment methods included a learning task and a posttest, completed by students during the first demonstration with eye-tracking enabled. The tasks involved graph reading and interpretation. The teacher aimed to highlight differences in students' attention patterns (Figure 3) and their focus shifts be-tween the learning and assessment tasks. As a formative assessment, students were encouraged to review their performance on the dashboard for reflection.

Value proposition: Integrating multimodal analytics, such as web-based eye tracking, offers valuable opportunities for social comparison of learning behaviors over time and across groups, benefiting both teachers and students. This approach not only supports deeper insights into engagement patterns but also sparks curiosity about the technology's accuracy and the validity of its computed indicators, fostering a more reflective and data-informed learning environment.

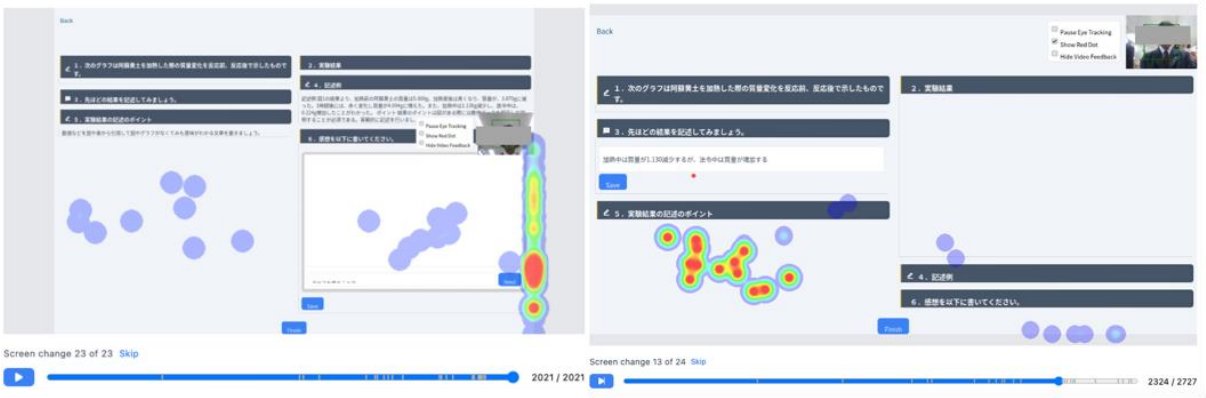


Figure 1. Eye tracking for different students attempting the same task.

Elicited user needs and ideas from mirror artefacts. 3 students and 2 teachers expressed their needs and ideas with the mirror task. In mirror one (related to issue), one theme emerged as the difficulty in reflection on ones mistakes, study methods and data needs (core issues, S1: "Reflect on the mistakes I made and get the reasons for them", S3: "I can't remember the study methods I used in the past.", T2: "I want to know how other people study, the difference between good and bad students"). Another core issue was related to motivation to act and monitor (S2: "I don't like to study, but I like when it is gamified", T1: "I can't maintain my motivation for studying English or taking qualification exams").

Peer comments in mirror two were shared for each issues and approaches. While AI-driven analysis was valued students also desired insights from specialized teachers and peer opinions, fostering a sense of camaraderie, especially when common mis-takes are identified. Participants noted that different individuals are motivated by varying factors, suggesting the integration of gamification elements tailored to user preferences. Further an AI for finding and categorizing different study methods was thought useful, but users should have the control whether their method should be shared with others. In Mirror three some specific ideas were shared incorporating the discussions and feedback. For instance, S3 mentioned about a platform where you can share the study methods you have used "Like" and save the study method → Try it out, and use AI analysis to see if it's effective → Brush up on one's study method. See Figure 4 shares artefacts created in Mirror 3 form by other participants (S1, S2, T1). They have shared various functions for gamification of the daily learning process which might motivate learners to continue in a self-directed way.

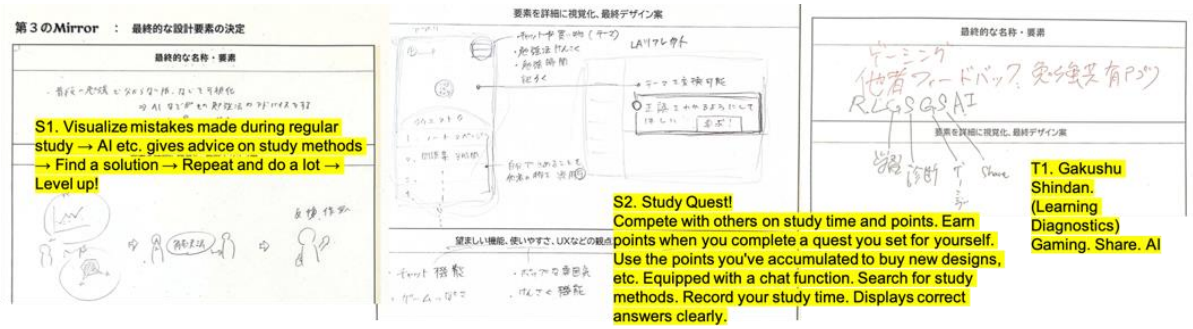


Figure 4. Design ideas that evolved amongst students and teachers in Mirror 3

5. Discussion and Future Work

This study elaborates a systematic co-design effort while piloting an eLearning tool that incorporates AI and multimodal learning analytics functions. The Japanese educational policy, with regards to the use of Generative AI, also recommends protecting learners specially at K12 level by ensuring a teacher guided use. Such a support is possible in tools like LA-ReflecT that enables authoring of activities and tracking of use based on users role. It further empowers learners to actively use data for self-improvement and goal attainment.

The described short yet frequent co-design session is what was effective in gathering the needs and refining the idea. After this initial phase, a teacher and researcher group co-developed and implemented a learning activity in one class with 40 students as pilot to group learners based on learning behaviors captured with multimodal data (Majumdar et al. 2025). We repeated the mirrors activities with those students and currently analysing to elicit further needs and ideas. A new round of revision of the functions, workflow and activity design will start after that. While the study is in the Japanese context, cross cultural studies are necessary as different countries are coming up with different policies related to AI in education (Bhutoria, 2022).

Acknowledgements

This project is partially funded for the Promotion of Joint International Research (International Collaborative Research) JP24KK0051.

References

Arghir, D. C. (2024). Implementation of Learning Management Systems with Generative Artificial Intelligence Functions in the Post-Pandemic Environment. *Information Technologies and Learning Tools*, 100(2), 217.

Bhutoria, A. (2022). Personalized education and artificial intelligence in the United States, China, and India: A systematic review using a human-in-the-loop model. *Computers and Education: Artificial Intelligence*, *3*, 100068.

Ifenthaler, D., Majumdar, R., Gorissen, P., Judge, M., Mishra, S., Raffaghelli, J., & Shimada, A. (2024). Artificial intelligence in education: implications for policymakers, researchers, and practitioners. *Technology, Knowledge and Learning*. https://link.springer.com/article/10.1007/s10758-024-09747-0

Lepe-Salazar, F., Escobedo, L., & Nakajima, T. (2025). Novice-friendly probes for the gathering and analysis of requirements and subsequent design of software. *International Journal of Human-Computer Studies*, 195, 103405.

Majumdar, R., Liang, C., Ocheja, P., & Li, H. (2025, May). A Cooperative Learning Framework with Joint Attention and Interaction Data in the LA-ReflecT Platform. In *Proceedings of the 2025 Symposium on Eye Tracking Research and Applications* (pp. 1-3).

Majumdar, R., Prasad, P., Kadam, K., Gatare, K., & Warriem, J. M. (2023). LA-ReflecT: A Platform Facilitating Micro-learning and Its Multimodal Learning Analytics. In *European Conference on Technology Enhanced Learning* (pp. 731-735). Cham: Springer Nature Switzerland.

Majumdar, R., Singh, D., Chen, M.R.A. (2024) Designing Structured Reflections For Guiding Learners' Interactions With Generative AI. In CELDA 2024.

Majumdar, R., Singh, D., Rajendran, R., Narayanan, S., Gatare, K. (2024) LA-ReflecT platform's affordances for distributed multimodal learning analytics. In CollabTech 2024.

Ogata H., Liang C., Toyokawa Y., Hsu C., Nakamura, K., Yamauchi, T., Flanagan, B., Dai, Y., Takami, K., Horikoshi I., Majumdar R. (2024) Co-designing Data-driven Educational Technology and Practice Reflections from the Japanese Context. *Technology, Knowledge and Learning*. https://doi.org/10.1007/s10758-024-09759-w

Oviatt, S. (2018, October). Ten opportunities and challenges for advancing student-centered multimodal learning analytics. In Proceedings of the *20th ACM International Conference on Multimodal Interaction* (pp. 87-94).

Sarmiento, J. P., & Wise, A. F. (2022, March). Participatory and co-design of learning analytics: An initial review of the literature. In *LAK22: 12th international learning analytics and knowledge conference* (pp. 535-541).

Use of Generative Artificial Intelligence, MEXT Policy (2024) https://www.mext.go.jp/a_menu/other/mext_02412.html

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Visualizing Knowledge-Based Learning Logs to Support Multi-Perspective K-12 Math Instruction

Yudai OKAYAMA[a*], Junya ATAKE[a], Kensuke TAKII[b], Changhao LIANG[b], & Hiroaki OGATA[b]
[a]*Graduate School of Informatics, Kyoto University, Japan*
[b]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
*okayama.yudai.28z@st.kyoto-u.ac.jp

Abstract: Conventional learning analytics (LA) systems often rely on test scores and behavioral logs, providing limited support for understanding learners from multiple pedagogical perspectives, which are essential for personalized instruction. Recognizing the need for more accessible and beneficial LA tools for K-12 teachers, we developed a visualization dashboard based on the Open Knowledge and Learner Model (OKLM) that integrates diverse learning data into a unified structure to enable flexible, multi-perspective interpretation of students' learning states. Through interviews with three junior high school mathematics teachers, we explored how the dashboard supports teachers in interpreting learners' states from various pedagogical perspectives. The results suggest that perspectives such as self-awareness and relative positioning were especially helpful for understanding students, and that the system allows flexible interpretation aligned with instructional goals. These findings help realize the potential of OKLM to offer richer insights for personalized instruction in real-world educational contexts.

Keywords: Knowledge-aware learning analytics, Dashboard, ICT-supported teaching, K12

1. Introduction

Learning Management Systems (LMS) are increasingly used in educational settings to support online learning and have produced large volumes of learner activity logs. These logs form the basis of Learning Analytics (LA), which aims to improve teaching and learning through data-informed insights. Yet, in both educational practice and learning sciences, there is growing consensus that personalized instruction requires a deeper understanding of each learner's state and characteristics (Tomlinson, 2014). Test scores and basic behavioral logs from LMS alone are insufficient. In particular, recent discussions have emphasized the importance of knowledge-aware learning analytics that can make sense of learner actions in the context of their knowledge states—an aspect still underexplored in current LA systems (Chen, 2019).

Takii et al. (2024) proposed the Open Knowledge and Learner Model (OKLM), a framework that maps learning logs to a domain knowledge structure and stores learners' knowledge states in a graph database. As an initial application, they developed an OKLM-based visualization tool for higher education. However, the system primarily allowed instructors to monitor how long each student spent viewing content, which limited its ability to support the kind of multidimensional understanding of learners that is increasingly emphasized in personalized instruction (Tomlinson, 2014). The OKLM's potential to capture various dimensions of learning through structured knowledge mapping had not yet been fully realized in its initial application. To address this limitation and enhance teachers' multifaceted understanding of learners, we developed an advanced OKLM Visualization Dashboard. This development aims to fulfill teachers' needs for more accessible and beneficial LA tools in K-12 settings, as the widespread integration of LA into elementary and secondary school

classrooms remains relatively under-researched (Valtonen et al., 2025). The effectiveness was evaluated in a Japanese junior high math setting, specifically aiming to support teachers in interpreting student information from diverse perspectives as a pilot study.

2. Theoretical Background and OKLM Visualization Dashboard

2.1 OKLM infrastructure

The OKLM stores learning logs from multiple learning support systems, structured by textbook-based knowledge units. Each learner's activity is recorded in terms of when, to which unit, in what aspect, and to what extent it occurred. These logs are organized as a network of interconnected units based on textbook sequencing, aligned with the Japanese national curriculum guidelines (MEXT, 2017).

A key feature of OKLM is its openness to integration with diverse systems. In this study, we particularly focus on this interoperability, which enables OKLM to flexibly incorporate xAPI-format learning logs from various platforms. This design allows the model to adapt to a wide range of educational settings, from elementary education to advanced learning analytics.

Two systems currently connected to OKLM are the GOAL (Goal-Oriented Active Learner) system and the Peer Help System. GOAL enhances students' self-directed learning (SDL) by encouraging regular reflections on academic activities (Wang et al., 2024). The Peer Help System supports peer-assisted learning by allowing students (help-takers) to post questions. Based on the knowledge states modeled in OKLM, the system recommends appropriate help-givers (Jiang et al., 2024). Help-takers then evaluate the responses they receive. These logs, particularly those related to self-awareness, learning strategies, and collaborativeness, are formatted in xAPI and integrated into the OKLM database.

2.2 Theoretical groundings of the dashboard

To support personalized instruction, recent work has highlighted the value of interpreting learners from multiple pedagogical perspectives. Among these perspectives, self-awareness and metacognition are seen as essential components of self-regulated learning, guiding teachers in feedback design and instructional decision-making (Zimmerman, 2002; Schunk, 2012). Collaborative learning that fosters deeper thinking through peer interaction and learning strategies such as reviewing and self-assessment are also recognized as key factors in improving learning quality (Johnson & Johnson, 1987). From a practical standpoint, teachers also rely on indicators such as learning outcomes to determine instructional priorities (Black & Wiliam, 1998), inter-unit balance to identify disparities in understanding across topics (Squires, 2013), and relative positioning to assess individual performance within the class (Becker & Rosen, 1992).

Based on these perspectives, this study hypothesizes that the OKLM visualization dashboard can support teacher understanding through multiple pedagogical viewpoints. While the dashboard is not explicitly designed to present all perspectives in a structured manner, it is intentionally designed to allow teachers to flexibly interpret students' learning states by leveraging the integrated structure of OKLM along with data from systems such as GOAL and the peer help system. The specific perspectives hypothesized in this study, along with their corresponding data sources and pedagogical implications, are as follows:

- *Learning outcomes*: Teachers can assess student achievement in each unit using objective data such as review tests and periodic test scores.
- *Self-awareness*: Subjective difficulty ratings entered via GOAL allow teachers to identify units students found personally challenging.
- *Metacognition*: By comparing self-assessments (e.g., perceived difficulty) with actual results, the dashboard helps estimate the accuracy of students' self-perceptions.
- *Collaborativeness*: Feedback from help-takers in the Peer Help System provides insights into each student's engagement in peer-supported learning.

- *Relative positioning*: Individual performance is shown alongside class averages, helping teachers gauge each student's standing within the class.
- *Inter-unit balance*: By presenting student data across multiple units, teachers can detect imbalances in students' understanding.
- *Learning strategies*: Combining self-reflection and difficulty ratings enables teachers to infer whether students adopt strategic behaviors, such as reviewing difficult content.

2.3 Interface of OKLM Visualization Dashboard

The following section describes the OKLM Visualization Dashboard proposed in this study. Figure 1 presents an overview of the system interface.

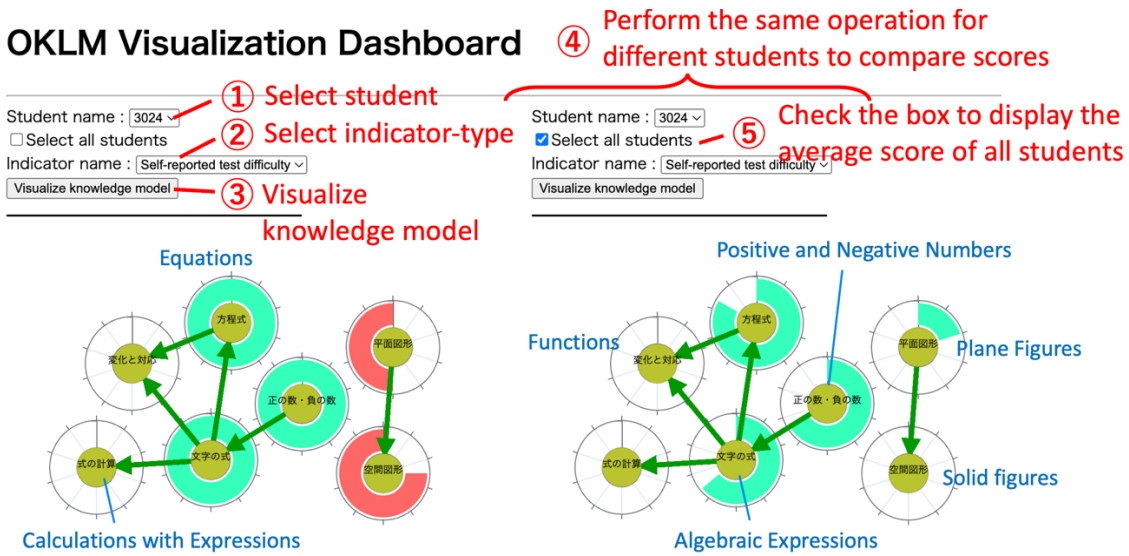


Figure 1. User Interface of OKLM Visualization Dashboard

Teachers begin by selecting a learner and an indicator type from the control panel in the upper-left area. The indicator type refers to a node in the OKLM database representing the "to what aspect" dimension of learning activity. Selecting an indicator determines which type of data will be visualized. When the "Visualize Knowledge Model" button is clicked, a network-based visualization appears in the lower panel, reflecting both the knowledge structure and the learner's indicator scores. This view helps teachers interpret the learner's current state and characteristics. The interface also includes a mirrored module on the right, allowing side-by-side comparisons of different indicators. Selecting "all students" displays average scores for the chosen indicator type across the class. By selecting various indicator types and comparing them with the class average, instructors can view each student's knowledge state from multiple perspectives.

3. Methodology

To evaluate the utility of the OKLM Visualization Dashboard in a K-12 context, we conducted a semi-structured interview with three mathematics teachers (Teachers A, B, and C) at a Japanese junior high school (School X). All had prior experience using both the GOAL and Peer Help systems and were familiar with the integrated data. Teachers A and B were active subject teachers with 10 and 8 years of experience, respectively; Teacher C, with 25 years of experience, served as vice principal. This study formulates the following research questions:

- RQ1: Can the proposed OKLM Visualization Dashboard help teachers to understand students from proposed perspectives?
- RQ2: Are the insights obtained from proposed perspectives useful for teachers?

3.1 Data Source

We used log data from 120 first-year students at School X during the 2024 academic year, derived from three sources and stored in the mathematics OKLM database: (1) GOAL, (2) Peer Help System, and (3) periodic test results. From GOAL, we collected four types of data reflecting different aspects of students' engagement with weekly review tests: test scores, perceived difficulty, performance expectations, and the presence of written reflections. From the Peer Help System, we collected ratings given by help-takers to help-givers after receiving assistance on posted problems. These ratings were automatically stored as distinct indicator types in OKLM, allowing teachers to interpret not only learning outcomes but also aspects such as self-awareness, metacognition, and collaborativeness. Periodic test scores, though manually entered, were included as indicators of learning outcomes, anticipating future integration with school management systems. Some indicators, such as perceived difficulty and performance expectations, include negative values. Visual elements like color gradients are employed to support intuitive interpretation by teachers.

3.2 Interview Design

At the beginning of the interview, a brief explanation of the OKLM Visualization Dashboard was given using real data from four randomly selected first-year students from the 2024 academic year. Teachers then explored the system to familiarize themselves with its features. Afterward, they participated in a semi-structured interview guided by predefined questions.

Following the research questions, the interview consisted of two main sections, each applied to the seven perspectives for interpreting learners' states: Section 1 (RQ1) asked whether each type of information could be interpreted from the dashboard, and Section 2 (RQ2) asked whether it was useful for understanding learners' states or characteristics.

In Section 1 (Q1-Q7), teachers were first asked their agreement with the statement: "Can this dashboard help you interpret information from the seven perspectives?", using a 5-point Likert scale (1 = Strongly Disagree to 5 = Strongly Agree). In Section 2 (Q8 - Q14), they were asked the usefulness of the seven perspectives with the statement: "Is the following information useful in understanding students' states or characteristics?" using the same scale. After each response, teachers were encouraged to explain their reasoning, suggest improvements, and share general impressions. The interviews were recorded and transcribed.

For qualitative analysis, we applied a constant comparative approach (Glaser, 1965), combining deductive coding based on the predefined pedagogical perspectives with inductive coding for additional themes that emerged during analysis.

4. Result & Discussion

Figure 2 shows teachers' evaluations of the dashboard: Section 1 results are on the left, and Section 2 on the right.

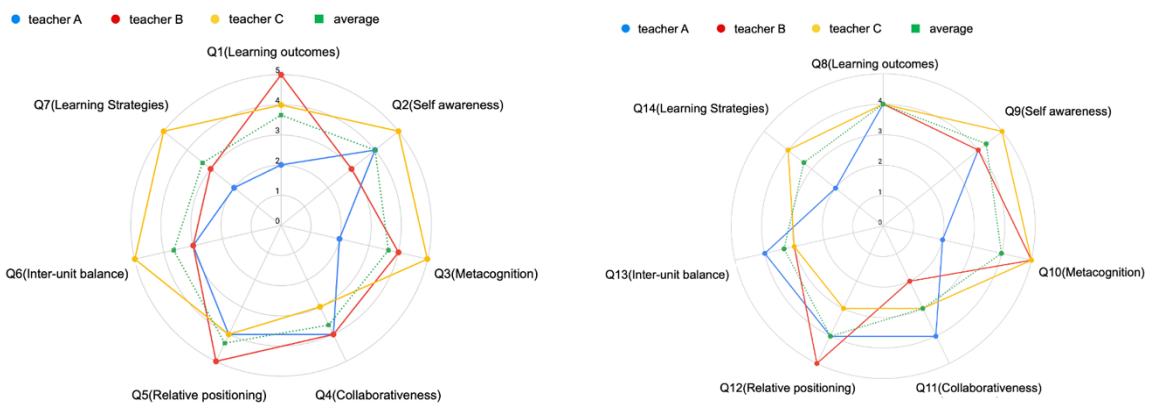


Figure 2. Likert scale results of Section 1 (Left) and Section 2 (Right)

In Section 1, only self-awareness and relative positioning received average scores above 4. However, all teachers rated at least three perspectives 4 or higher, indicating that they perceived the dashboard as supporting multi-perspective learner understanding. Interview comments reflected each teacher's instructional stance. Teacher A noted, "It's easy to see which unit students are good at or struggle with," and "I can get a general sense of where each student stands in the class," demonstrating that the dashboard facilitated understanding from the perspectives of self-awareness and relative positioning.

Other perspectives showed more variability. On learning outcomes, Teacher B stated, "I strongly agree with how the learning outcomes for each unit are shown," while Teacher A commented, "Honestly, it's hard to say at this point." Regarding metacognition, Teacher B appreciated "the gap between actual scores and perceived performance," and Teacher C noted, "The way the GOAL scores and questionnaire results are quantified seems potentially useful." In contrast, Teacher A observed, "The gap between students' self-evaluations and their actual performance is hard to grasp—it varies a lot by student." These differences may stem from varying levels of familiarity with the system and individual pedagogical beliefs. Teacher A, who rated several aspects lower, remarked, "It's possible that I haven't fully figured out how to read this graph yet," suggesting that prior experience with similar tools influenced interpretation. This aligns with earlier research showing that teachers' digital literacy and pedagogical beliefs shape how they adopt educational technologies (Ertmer, 2005).

Section 2 showed that learning outcomes, self-awareness, metacognition, and relative positioning again received average scores above 4.0. These perspectives were perceived as relatively useful in understanding students' states. Meanwhile, collaborativeness and inter-unit balance were rated lower. Although collaborative learning is often emphasized in pedagogy (Johnson & Johnson, 1987), the perceived usefulness of collaborativeness varied. Teacher C remarked, "I'm not sure how well this captures collaboration," suggesting a gap between pedagogical ideals and system implementation. Regarding inter-unit balance, Teacher B commented, "I don't think inter-unit balance is that important," showing a lesser focus on cross-unit comparisons. The learning strategies perspective also yielded mixed responses. While teachers acknowledged its value, they pointed out usability concerns. Teacher B said, "I do think it's useful to look at, but it's a lot of work to go through each student individually," implying a need for more intuitive or aggregated visualizations to make such data actionable in classroom contexts.

In summary, for RQ1, the interviews suggested that perspectives such as self-awareness and relative positioning were generally effective in helping teachers interpret students' conditions and characteristics. However, for the other five perspectives, there was considerable variation in perceived interpretability and usefulness. These differences likely reflect individual differences in ICT proficiency and instructional beliefs. Regarding RQ2, the two perspectives found to be informative for RQ1 (self-awareness and relative positioning) were also rated as relatively useful. Learning outcomes and metacognition were additionally seen as potentially helpful, but perceptions varied among teachers in both interpretability and usefulness.

One limitation of this study is the small dataset used: only four first-year junior high school students were included. This restricts the generalizability of the findings. However, teacher A remarked, "I'd like to see cross-grade connections, such as moving from equations to simultaneous equations," highlighting the potential value of more longitudinal or cross-grade data for deepening insight.

5. Conclusion & Future Work

This study conducted a pilot investigation into the potential of a knowledge-oriented, cross-activity educational dashboard for supporting teacher decision-making in K-12 contexts. By visualizing learner states through the OKLM, the dashboard enabled teachers to interpret students' conditions from multiple perspectives. The system also demonstrated interoperability with external learning support tools via xAPI, allowing the integration of log

data at the level of individual knowledge units. This flexibility suggests feasibility for continued expansion as educational environments evolve.

To further enhance its practical value, the dashboard should expand its capacity to visualize behavioral and non-cognitive information that is typically difficult for teachers to access. For example, data on learning retention, motivation shifts, or post-assessment behavior may enrich teacher understanding beyond traditional performance indicators. As teacher B remarked, “It’s helpful to be able to see indicators that we don’t usually look at,” highlighting the potential of the system to reveal previously unseen learner characteristics. This kind of adaptability is essential for differentiated instruction, as teachers must interpret diverse student profiles to offer personalized support (Tomlinson, 2014)

In terms of future improvement, refining the user interface for more intuitive interpretation remains a priority. Additionally, we aim to introduce automated features such as system-generated task recommendations, especially to support learners who encounter specific conceptual difficulties. These enhancements would help connect dashboard insights more directly to actionable classroom practices.

Acknowledgements

This work was supported in part by JST Council for Science, 3rd SIP JPJ012347, the establishment of university fellowships towards the creation of science technology innovation JPMJFS2123, and JSPS KAKENHI 23H00505 and 25K21357.

References

Becker, W. E., & Rosen, S. (1992). The learning effect of assessment and evaluation in high school. *Economics of Education Review, 11*(2), 107-118. https://doi.org/10.1016/0272-7757(92)90002-K

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *80*(2), 139–148.

Chen, W. Knowledge-aware learning analytics for smart learning. (2019). *Procedia Computer Science*, vol. 159, pp. 1957-1965. https://doi.org/10.1016/j.procs.2019.09.368

Ertmer, P. A. (2005). Teacher pedagogical beliefs: The final frontier in our quest for technology integration?. *ETR&D*, 53, 25-39. https://doi.org/10.1007/BF02504683

Glaser, B. G. (1965). The constant comparative method of qualitative analysis. *Social problems*, *12*(4), 436-445. https://doi.org/10.2307/798843

Jiang, P., Liang, C., & Ogata, H. (2024). Data-driven Peer Recommendation and Its Applications in Extracurricular Learning. *Proceedings of the 32nd International Conference on Computers in Education.* https://doi.org/10.58459/icce.2024.4888

Johnson, D. W., & Johnson, R. T. (1987). *Learning together and alone: Cooperative, competitive, and individualistic learning* (2nd ed.). Prentice-Hall, Inc.

Ministry of Education, Culture, Sports, Science and Technology-JAPAN. (2017). *Curriculum guidelines for junior high schools* (Notification No. 29). https://www.mext.go.jp/content/20230120-mxt_kyoiku02-100002604_02.pdf

Schunk, D. H. (2012). Learning theories: An educational perspective (6th ed.). Pearson Education.

Squires, D. A. (2013). The Balanced Curriculum Model: Description and Results. *SAGE Open, 3*(1).

Takii, K., Koike, K., Horikoshi, I., Flanagan, B., & Ogata, H. (2024). Oklm: A universal learner model integrating everyday learning activities with knowledge maps. *In Companion Proceedings of the 14th International Learning Analytics and Knowledge Conference*, 191-193.

Tomlinson, C. A. (2014). *The differentiated classroom: Responding to the needs of all learners* (2nd ed.). ASCD.

Valtonen, T., Paavilainen, T., López-Pernas, S., Saqr, M., & Hirsto, L. (2025). Elementary and Secondary School Teachers’ Perceptions of Learning Analytics: A Qualitative Approarch. *Technology, Knowledge and Learning, 30,* 601-619.

Wang, Z., Hsu, C., Horikoshi, I., Li, H., Majumdar, R., & Ogata, H. (2024). Classifying Self-Reflection Notes: Automation Approaches for GOAL system. *Proceedings of the 32nd International Conference on Computers in Education*. https://doi.org/10.58459/icce.2024.4883

Zimmerman, B. J. (2002). Becoming a Self-Regulated Learner: An Overview. *Theory Into Practice*, *41*(2), 64–70.

Conflow: An Online Concept-Mapping Tool to Measure Conceptual Fluency

Shivin AGARWAL[a +], **Anveshna SRIVASTAVA**[a] **& Nisheeth SRIVASTAVA**[a]
[a] *Indian Institute of Technology Kanpur, Kanpur, India*
+ shivin.agarwal15@gmail.com; shivin23@iitk.ac.in

Abstract: Assessment at scale operates under constraints of reproducibility, interpretability, and comparability. In practice, these constraints confine pedagogical evaluations to recitations of fact or reproductions of formulaic calculations. Efforts to design novel, more holistic methods of assessment tend to violate one of the three contracts listed above and thus find limited uptake in practice. In this paper, we describe *ConFlow* - an online assessment instrument that builds on the strengths of concept mapping and aims to satisfy these constraints while enabling holistic, concept-level assessment. We describe the tool and a validation study we undertook to evaluate its performance vis-à-vis personalized ratings provided by college undergraduates. ConFlow lets learners build concept maps, analyze them, and present a 'quadrant graph' depicting the mapper's fluency with the mapped concepts. The quadrant graph categorizes conceptual fluency as 'Well-Understood', 'Partially Understood', 'Fuzzy', and 'Unknown' concepts. ConFlow generates this graph based on the number of conceptual moves and the accuracy of the conceptual links and phrases made during the online mapping process. In this pilot work, we present our attempt at validating the tool for wider usage by the learning community. Our process-based assessment method has implications for capturing learning trajectories, and it offers a more profound understanding of concept-level uncertainties and misconceptions.

Keywords: ConFlow; Conceptual fluency; Concept-mapping; Quadrant-graph; Assessment

1. Introduction

Learning is as diverse as the people who engage in it. For one student, something that comes naturally may prove difficult for another. Despite this diversity, educational assessments use standardized tests to assess everyone using the same scale, providing structure but overlooking the complexity of unique learners. Although evaluation approaches have advanced significantly, techniques for assessment at scale struggle to satisfy three fundamental criteria: interpretability, comparability, and reproducibility.

Interpretability suggests that instructional methods and assessment outcomes must be transparent and understandable, aiding in comprehending the underlying reasoning behind the use of educational interventions (Baird, et al. 2019). Comprehension is essential since it helps instructors recognize which teaching strategies work or do not work in different pedagogical contexts. The 'comparability' criterion is crucial for systematically assessing the efficacy of various interventions across different contexts. It calls for diverse educational practices to be measured against uniform standards and metrics (Evans & Lyons, 2017). Without comparability, differences in implementation or contextual factors can obscure the actual impact of a teaching method, making it difficult to generalize findings or conduct meta-analyses that inform policy and practice (American Educational Research Association [AERA], APA, & NCME, 2014). The last criterion of 'reproducibility' ensures that a technique or intervention will produce consistent results when used in similar circumstances (Schuwirth & Van der Vleuten, 2018). Consistency is essential for scaling effective practices across diverse educational environments as well as for validating prior knowledge and the success of the educational outcome. Reproducibility in education is especially difficult because

classrooms are dynamic and eclectic, with differences in student backgrounds, teacher skills, curriculum design, and even daily classroom conditions impacting results.

In this study, we explore the online concept-mapping process to tackle the above assessment constraints of interpretability, comparability, and reproducibility, along with developing a holistic assessment of a learner's fluency with individual concepts, based on both their learning process and outcome. Almost half a century ago, David Ausubel suggested that the most important factor influencing learning is what the learner already knows and that teachers should ascertain their prior knowledge and teach accordingly (Ausubel, 2012). Since the construction of concept maps involves the invocation of prior knowledge, along with the use of different modalities of text and visuals in establishing links across concepts, concept maps could potentially be used to test Ausubel's hypothesis. Hence, in this study, we set out to introduce and validate 'ConFlow', an online concept-mapping tool to assess learners' fluency with different concepts.

1.1 Concept Maps

Concept mapping is a graphical representation of a person's knowledge of a domain, where nodes (concepts), linking arrows, and phrases are used to provide insights into their mental organization of concepts (Novak, 2012). Nodes, connected by linking arrows and phrases, constitute 'propositions,' which are evaluated, along with the overall map structure, to assess conceptual understanding (Novak & Gowin, 1984). They can effectively address the challenges of interpretability, comparability, and reproducibility in education. They enhance interpretability as they present information visually in a structured node-link format, enabling teachers and students to quickly identify gaps in understanding and integrate new information with existing knowledge (Novak, 2012). They support comparability as their standardized formats, when evaluated with established rubrics, allow educators to objectively assess and benchmark the depth and clarity of students' cognitive structures (Katagall, Dadde, Goudar, & Rao, 2015). Lastly, as they require learners to externalize their internal knowledge frameworks, concept maps offer a reproducible artifact that can be consistently generated and reviewed over time.

Concept maps face several limitations, such as building simplistic maps that do not capture the complexity of the domain, a lack of metacognitive engagement that impedes learners' ability to structure and internalize new information (Mayer & Moreno, 2002), a lack of objectivity in assessment that impedes comparability across contexts (Novak, 2012), the static nature of maps limiting the fostering of deep understanding (Machado & Carvalho, 2020), resistance from students when it is not meaningfully integrated with the classroom routine (Maker & Zimmerman, 2020), and difficulty with revision and reorganization when using the pen-paper mode. As education systems become increasingly data-driven and digital, the opportunity for creating a dynamic concept mapping tool that performs a holistic assessment by letting learners metacognitively engage with their learning process is ripe. Hence, here, we propose an approach that leverages advances in adapting concept mapping toward pedagogical evaluations (Srivastava et al., 2021). In an earlier work, we had shown how procedural analysis could be used to measure learners' conceptual fluency (Srivastava et al., 2014). In the present work, we have moved to an online interface, designing *ConFlow*, that lets learners build concept maps, analyze them, and present a quadrant graph that depicts the mapper's fluency with the different mapped concepts.

2. Method

2.1 The Tool: ConFlow

The 'ConFlow' application allows users to register and choose their preferred domain. Currently, it hosts concepts from basic statistics and biology. For the present validation task, our focus was the 3-D structure of the DNA (Deoxyribonucleic Acid) biomolecule.

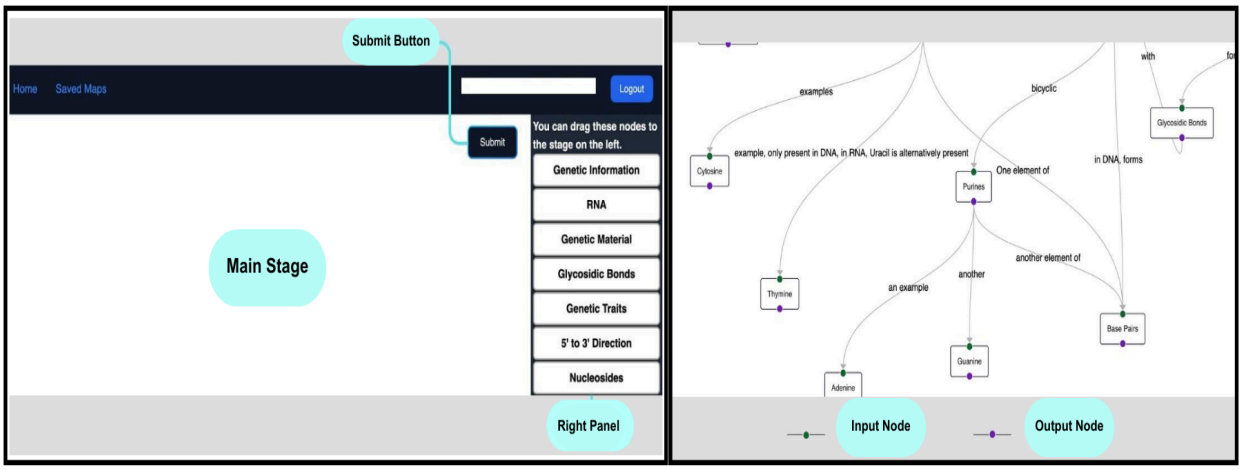


Figure 1. a) The Interface Of ConFlow, and b) A sample concept map

When the user reaches the mapping interface, they can see a list of draggable concepts on the right (see Figure 1a). To create a concept map, users can individually drag and drop these concepts onto the main stage. Two concepts can be connected with a linking arrow, indicating the directionality of the relationship, and with a linking phrase, suggesting the relationship between the two concepts. Thus, a map is built after all the concepts have been exhausted and linked to the conceptual network on the stage (see Figure 1b for a sample).

2.2 Categorization of Quadrants

We use two measures to characterize students' conceptual fluency. The number of moves per concept, plotted on the 'X'-axis, which we use as a proxy for the confidence that a user exhibits regarding the conceptual fluency with a concept, and the accuracy score based on their final map evaluation, plotted on the 'Y'-axis. Based on the above two measures, we classify the conceptual fluency of a learner with all the domain concepts used in the study into four quadrants: 'Unknown', 'Fuzzy', 'Partially Understood', and 'Well Understood' (see Figure 2). Unknown concepts have fewer moves and less accuracy, indicating incorrect connections and incorrect linking phrases. Fuzzy concepts have higher moves and less accuracy, indicating slightly correct connections to sibling concepts and correct linking phrases. Partially Understood concepts have more moves than the average and a higher final score, indicating correct connections and correct linking phrases. Well Understood concepts have fewer moves and more accuracy, indicating correct connections and correctly marked linking phrases.

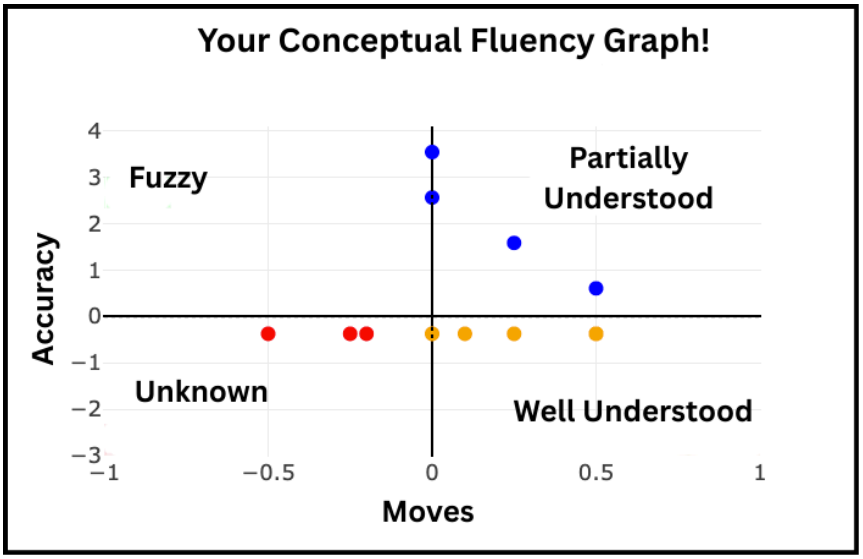


Figure 2. The quadrant graph: dots signify concepts, and the categories signify fluency

Evaluation of the concept maps involved assigning '0' to no linking phrase, '0.5' to incomplete sentences, '0' to meaningless propositions, and '1' to correct propositions.

2.3 Sample and the task

The group of participants in the validation study consisted of 16 consenting adults (Males=10, mean age=21.6 years, SD=2.98; Females=6, mean age=23.3 years, SD=3.01). The participants viewed a 5-minute introductory video on how to use the tool. A training task involving a sample concept map was provided to acclimatize them with the interface. The main task then commenced, where each participant worked with 26 concepts to build the concept map on DNA. Each participant's comprehension of the mapping was assessed through a personal rating after task completion. After the task, participants were asked to classify their understanding of individual concepts into one of four predefined categories: 1. Unknown: You don't know about the concept, 2. Fuzzy: You have a vague understanding of this concept, having encountered it before, but you don't recall much about it, 3. Partially Understood: You know the concept but can't recall it completely, 4. Well Understood: You are very confident about your knowledge of the concept. After the main task, participants were administered a post-test in a multiple-choice questionnaire format to assess their understanding of concepts related to the building task. Lastly, they were asked to complete a feedback form to share their overall experience. The entire experiment was completed within 60 minutes.

3. Analysis

3.1 Evaluation of Concept Maps in ConFlow

We use Cohen's Kappa (κ), a statistical measure, to quantify the level of agreement between the two ratings of conceptual fluency- personal and the tool-based conceptual categorizations. Cohen's Kappa is calculated as:

$$K = (Po - Pe)/(1-Pe)$$

We operationalize 'Po' as the ratio of the common concepts in that category to the total number of concepts marked in that category. 'Pe' is the probability that suggests that a single concept can come in any of the four categories. Here, it is 1/4 (probability of random assignment to any of the four quadrants).

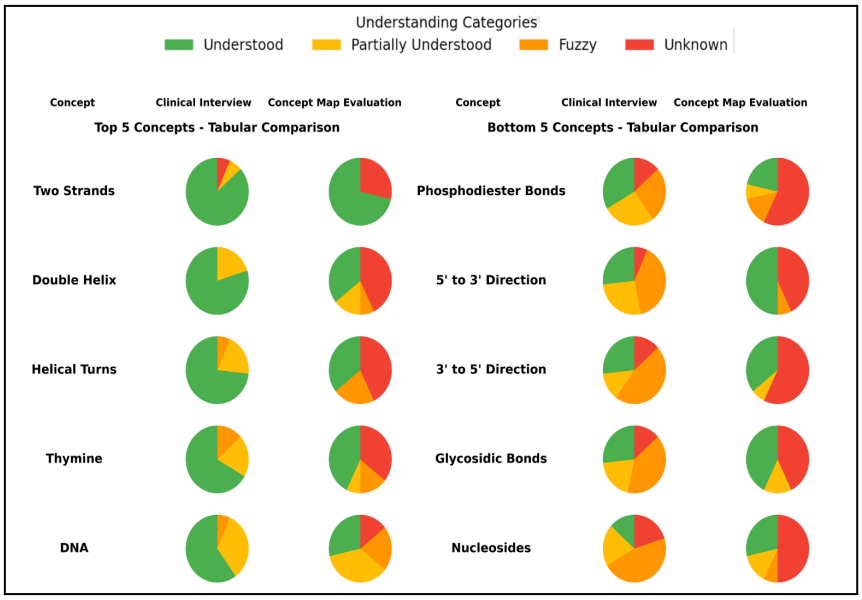


Figure 3. The best and the least well-understood concepts in the study

4. Results

Our validation strategy involved estimating the extent to which our novel concept-mapping-based measurement of gradation in conceptual fluency could reproduce similar gradation inferred via personal ratings. These estimates in turn were obtained from the quadrant plots defined above in the Methods section, with each subject's quadrant plot used to identify whether a concept was well understood, partially understood, fuzzy, or unknown. The pie charts in Figure 3 show the relative fraction of students falling into each category for our measurement instruments. Visually, it is evident that the concept mapping results align well with the clinical interview-based judgment for the study population. The Spearman rank correlation of the degree of conceptual understanding, measured across all 26 concepts using both these methods, is 0.225, further supporting the observation that the concept map measurement is a weak to moderate proxy for personal rating at a cohort level.

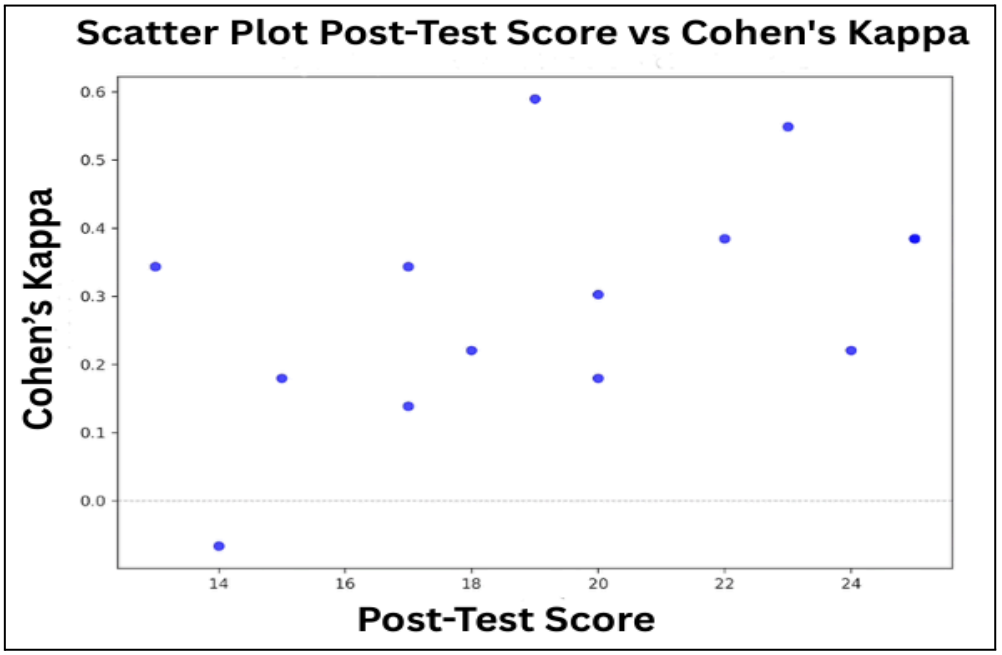


Figure 4. Post Test vs Cohen's Kappa relation

Interestingly, this relationship is stronger at the individual level. Figure 4 plots the individual-level proportion of agreement (measured using Cohen's kappa) for all participants on the Y-axis, with the average (mean = 0.297) significantly greater than zero (p-value: < 0.001). Cohen's kappa calculation uses understanding codes assigned to a candidate based on personal ratings of concepts and codes assigned to a candidate based on the quadrant graph to estimate a within-participant proportion of agreement for the two modalities. We also find that performance in the study post-test highly correlates with the proportion of agreement between the two measurement modalities (r = 0.615). We interpret this correlation to reflect variability in effort level in the study procedure by students since the post-test was intended to be substantially straightforward and thus was difficult to score poorly on without inattentive performance.

5. Discussion, Limitations, & Way Forward

In this work, we examined how concept maps could enhance pedagogical evaluations by producing reproducible, comparable, and interpretable measurements that are significantly more holistic and granular than traditional methods. *ConFlow* demonstrates a way to We measure granular categories, such as "well understood" and "fuzzy," to bridge the gap between qualitative insights and quantitative reproducibility. The digital interface represents a novel assessment approach, allowing broader application and scaling-up possibilities

across domains. The results indicate that individuals have a poorer sense of what they do not know (Figure 3), which could be utilized in future studies to enhance learners' metacognitive abilities. Owing to the limited dataset, generalization is complex. However, this pilot study contributes meaningfully to the literature on process analysis, focusing on learning processes rather than solely on outcomes (Zhang, Wang, Qi, Liu, & Ying, 2023).

This is in line with the position that persistence and effort shown in a domain reflect the development of resilience and grit, which are necessary for achieving a more profound understanding (Dweck, 2006). In our work, we expect concepts to gradually move from the 'unknown' zone to the 'known' zone, based on learners' fluency with those concepts. This behavior would demonstrate the growing competence of learners in the domain. Just as sustained effort over time predicts success more reliably (Duckworth, 2016), we hypothesize that the movement of concepts across quadrants would determine learners' persistence and fluency as well. Taking notes from reviewers' insights, future work will involve empirical testing in real classrooms and modifying the dashboard to include instructional next steps and feedback loops. We also intend to undertake a longitudinal study with a larger participant sample to further validate the system's effectiveness.

Acknowledgments

We are grateful to all the participants for their contribution to this study.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). Standards for educational and psychological testing. Washington, DC: American Education Research Association and National Academy of Education.

Ausubel, D. P. (2012). The acquisition and retention of knowledge: A cognitive view. Springer Science & Business Media.

Baird, M. D., & Pane, J. F. (2019). Translating standardized effects of education programs into more interpretable metrics. Educational Researcher, 48(4), 217- 228.

Dweck, C. S. (2006). Mindset: The new psychology of success. Random House.

Duckworth, A. (2016). Grit: The power of passion and perseverance (Vol. 234). New York, NY: Scribner.

Evans, C. M., & Lyons, S. (2017). Comparability in balanced assessment systems for state accountability. Educational measurement: issues and practice, 36(3), 24-34.

Katagall, R., Dadde, R., Goudar, R. H., & Rao, S. (2015). Concept mapping in education and semantic knowledge representation: an illustrative survey. Procedia computer science, 48, 638-643.

Novak, J. D. (2012). Concept mapping: A strategy for organizing knowledge. In Learning science in the schools (pp. 229-245). Routledge.

Novak, J. D., & Gowin, D.B. (1984). Learning how to learn. Press Syndicate of the University of Cambridge.

Schuwirth, L. W., & Van der Vleuten, C. P. (2018). How to design a useful test: the principles of assessment. Understanding medical education: evidence, theory, and practice, 275-289.

Mayer, R. E., & Moreno, R. (2002). Aids to computer-based multimedia learning. Learning and instruction, 12(1), 107-119.

Machado, C. T., & Carvalho, A. A. (2020). Concept mapping: Benefits and challenges in higher education. The Journal of Continuing Higher Education, 68(1), 38-53.

Maker, C. J., & Zimmerman, R. H. (2020). Concept maps as assessments of expertise: Understanding of the complexity and interrelationships of concepts in science. Journal of Advanced Academics, 31(3), 254-297.

Srivastava, A., Srivastava, N., & Chandrasekharan, S. (2014). Measure concept-mapping, not concept-maps: Procedural analysis elucidates stages in students' understanding of biology concepts. Proceedings of the annual meeting of the American Educational Research Association.

Srivastava, A., Srivastava, N., & Chandrasekharan, S. (2021). Quality of concept maps is affected by map-building strategies. Biochemistry and Molecular Biology Education, 49(1), 129-139.

Zhang S, Wang Z, Qi J, Liu J, Ying Z. Accurate Assessment via Process Data. Psychometrika. 2023;88(1):76-97. doi:10.1007/s11336-022-09880-8

An Exploratory Study of a Generative AI Teacher Coach

Karishma SHANGHVI[a], Sridhar IYER[b]
[ab]*Center for Education Technology, Indian Institute of Technology, Bombay, India*
**24D1204@iit.ac.in*

Abstract: In low-resource educational environments, scalable teacher professional development remains a significant challenge. This exploratory study investigates the usability and perceived usefulness of a Generative AI (GenAI)–based teacher coaching platform designed to provide personalized, feedback-driven instructional support to teachers. The GenAI coach was built using established coaching and AI design models. The GenAI coach analysed audio from teacher talk during classroom teaching to provide personalised feedback and recommendations to teachers, including co-creating future goals and action plans. Teachers from a low-income K–12 school in Mumbai, India used the GenAI coach over multiple sessions. A mixed-methods evaluation included surveys and semi-structured interviews. Results indicated above-average usability and positive perceived value. Teachers highlighted strengths such as unbiased, personalised, evidence-based feedback, privacy and encouragement of self-reflection. Challenges included limited classroom capture, long processing times, and a need for lesson-specific feedback and pre-class planning support. The study offers early evidence that GenAI coaching tools can provide scalable and context-sensitive professional development in settings where access to human coaches is limited. Findings have implications for the future design of AI-enabled teacher development systems in India and other resource-constrained environments.

Keywords: Generative AI, Teacher Professional Development, Teacher Coaching, Scaling, Low-Resource Environments

1. Introduction

1.1 *Teacher Coaching as Professional Development*

Teacher coaching is a form of teacher professional development that enhances teachers' instructional practices and improves student learning outcomes. Extensive research has demonstrated that coaching, especially when individualised, sustained and context-dependent, supports teachers in reflecting on their practice, setting meaningful goals, and applying teaching strategies (Bush, 1984; Knight, 2007; Kraft et al., 2018; Reddy et al., 2021). However, in low-resource educational environments such as India, delivering high-quality instructional coaching at scale remains a substantial challenge (Knight, 2012). Constraints such as limited access to expert coaches, financial limitations, and logistical difficulties hinder the widespread adoption of traditional coaching models.

1.2 *Generative AI Based Teacher Coaching*

To address these challenges, this study explores the potential of leveraging generative artificial intelligence (GenAI) to provide personalised, one-on-one, scalable coaching to teachers. Specifically, a GenAI-based teacher coach was developed to offer teachers actionable feedback on their classroom teaching, recommend specific instructional practices and guide them to create future goals and action plans aimed at instructional improvement. The GenAI coach seeks to democratize access to coaching, particularly benefiting teachers

and schools in low-income contexts where conventional coaching is difficult to implement effectively.

To use the GenAI coach, the teacher registers on the mobile app and engages in an onboarding conversation. For every coaching cycle, the teacher records her voice while teaching in the classroom and uploads the recording to the app, along with a brief description of the class. The GenAI coach transcribes and analyses the recording based on a pre-specified rubric of teaching quality. Feedback is provided to the teacher with specific quotations from the transcript as evidence. The teacher and coach then engage in a collaborative conversation and co-create an action plan for improvement in the next class.

The current exploratory study was conducted to evaluate the usability and usefulness of the GenAI coach, by engaging three teachers from a private K-12 school in Mumbai that serves low-income students to gauge teacher perceptions regarding the value of the app as a professional development tool. The study aimed to uncover both strengths and limitations of the GenAI coach from the perspective of actual users within a real-world educational setting.

1.3 *Research Questions*

The study is guided by the following research questions:

1. How usable is the GenAI coach according to teacher users?
2. How useful do teachers perceive the GenAI coach for their professional development?
3. What improvements do teachers suggest to enhance the effectiveness of the GenAI coach?

2. Generative AI in Teacher Coaching

Teacher coaching is well documented as an impactful method for enhancing teacher practice and student achievement. Coaching models for teachers have evolved over the years, encompassing peer coaching (Joyce & Showers 1980), cognitive coaching (Costa & Garmston, 1994), instructional coaching (Knight 2007) and student centered coaching (Sweeney, 2011) amongst others. Human to human coaching generally involves cycles of human experts working with teachers, observing classrooms, providing feedback, and supporting reflective practice.

Since 2022, the emergence of generative AI (GenAI) offers new possibilities in coaching, with large language models (LLMs) capable of analyzing complex classroom data and generating nuanced feedback. Generative AI can result in more complex, open-ended coaching, providing personalised feedback and adaptive dialogue for teacher professional development.

3. Theoretical Basis for the Generative AI Coach Design

While designing GenAI coach application, the researchers identified models for crafting the AI training prompt (Kamphorst, 2017; Terblanche, 2020), the teacher coaching process (Knight, 2007) and the teacher competency rubric (Danielson, 2011). See Figure 1 alongside for theoretical model use across the GenAI coaching process.

3.1 *AI Training Prompt*

Kamphorst (2017) lists eight essential features of effective e-coaching systems, social ability, credibility, context awareness, personalisation and learning, data integration, proactivity, behaviour change modelling and planning support.

The Designing AI Coach (DAIC) framework (Terblanche, 2020) outlines a structured and ethical approach to developing AI coaching chatbots for organisational settings. The model integrates both coaching science and AI design principles. The model requires specific design considerations for the AI coach such as defining the organisational context, creating boundaries of ethical conduct, designing for the relationship aspects of the AI coach and

coachee, choosing the theoretical coaching model, and specifying the required output of the coaching exercise (Terblanche, 2024).

The prompt design of the GenAI coach platform used in this study was based on both the Kamphorst criteria and the DAIC framework.

3.2 *Teacher Competency Rubric*

Multiple teacher competency rubrics were explored to provide a rich rubric for the GenAI coach to analyse teacher talk in the classroom. The Framework for Teaching (FFT) model (Danielson, 2011) was chosen because of its wide implementation and research validity. Under the FFT, four domains of teacher competency are described with elements and rubrics, which are 1) planning and preparation, 2) the classroom environment, 3) instruction, and 4) professional responsibilities. For the purpose of the instructional coach design, since only classrooms are observed, Domains 2 and 3 were chosen as relevant.

3.3 *Teacher Coaching Process*

While there exist many teacher coaching processes (Costa & Garmston, 1994; Sweeney 2011), the theoretical coaching model chosen for the GenAI coach is popular and research validated Impact Cycle from the Instructional Coaching model (Knight, 2007). The cycle is a structured coaching model for K-12 teachers that guides instructional coaches and teachers through a three-phase process: Identify, Learn, and Improve. In the *Identify* phase, coaches observe teachers and identify strengths and areas for improvement. The *Learn* phase involves a collaborative conversation between the coach and teacher to discuss classroom observations, share feedback and explore teaching strategies. Finally, in the *Improve* phase, the coach and teacher collaboratively create an action plan and teacher applies the strategies in the classroom. The interaction process between the GenAI coach and teacher followed the Impact Cycle (see Figure 1).

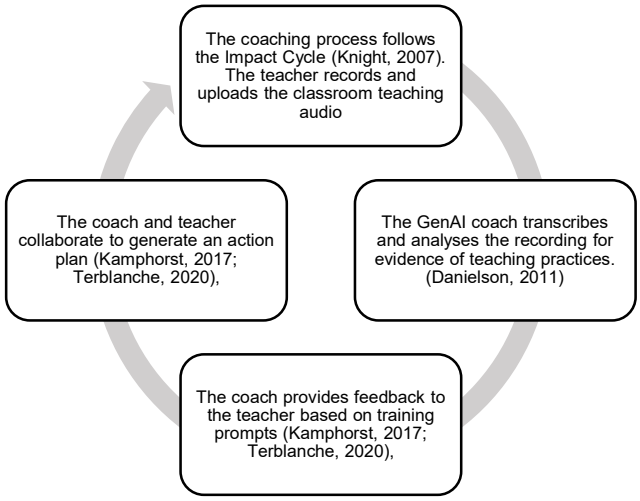


Figure 1. Theoretical models used in the GenAI coach design

4. Research Design

Three teachers from a K-12 school in Mumbai, teaching different grades and subjects, participated in the study. The teachers used the coach for three consecutive classes, after which, each teacher was interviewed by a school-based researcher. The interview protocol was designed to elicit detailed feedback on usability, usefulness and suggestions for improvement (see Table 1). Interviews were video-recorded, transcribed verbatim and analysed to inform the following three research questions.

Table 1. *Surveys alignment with Research Questions*

Survey	Research Question addressed
System Usability Survey	RQ1
Usefulness Survey	RQ2
Suggestions for Improvement	RQ3

A 10-question standardised System Usability Survey (SUS) (Bangor, 2008) generated an average composite score out of 100 across the three teachers. These scores were interpreted using established SUS benchmarks to evaluate overall usability of the AI coach.

A 10-question usefulness survey (Table 2) was designed around social learning theory constructs (Bandura, 1997) and face validated with an external educational researcher. Questions and their corresponding constructs are mentioned in the Results section 5.2 below. The survey followed a Likert Scale of 1 = strongly disagree to 5= strongly agree.

At the end of the interview, teachers responded to open-ended questions on the strengths and weaknesses of the coaching experience. From the questions listed below positive feedback questions (1-3) and negative feedback questions (4-6) were asked.

1. What are you most excited about when using an AI teacher coach?
2. What was the most helpful feature of the AI teacher coach?
3. How did using the AI teacher coach influence your teaching, if at all?
4. What are your biggest concerns about using an AI teacher coach?
5. What was frustrating or difficult about using it?
6. What suggestions do you have to improve the AI teacher coach?

Answers were transcribed and categorized as positive and negative, and tallied for frequency. The five most frequently stated positive and negative aspects of the GenAI coach were listed.

5. Results

5.1 *Usability Survey*

The System Usability Survey (SUS) revealed an average usability score across the three teachers of 70.83 (see Table 2), which implies that the coach is generally considered to be acceptable and above average in terms of usability.

Table 2. *Average Scores on the Usability Survey*

#	Questions (Likert Scale: 1: Strongly Disagree, 5: Strongly Agree)	Average Score
1	I think that I would like to use the coach frequently.	4.00
2	I found the coach unnecessarily complex.	2.00
3	I thought the coach was easy to use.	3.67
4	I think I would need the support of a technical person to be able to use the coach.	2.67
5	I found the various functions in the coach were well integrated.	3.33
6	I thought there was too much inconsistency in the coach.	2.33
7	I would imagine that most people would learn to use the coach very quickly.	4.67
8	I found the coach very cumbersome to use.	2.67
9	I felt very confident using the coach.	5.00
10	I needed to learn a lot of things before I could get going with the coach.	2.67
	SUS SCORE (on 100)	70.83

5.2 *Usefulness Survey*

On a Likert Scale of 1 to 5, the average usefulness score was 3.98, implying significant usefulness for a teacher (See Table 3).

Table 3. *Average Scores on the Usefulness Survey*

#	Questions (Likert Scale: 1: Strongly Disagree, 5: Strongly Agree)	Construct	Average Score
1	The feedback I got was accurate.	Perceived Accuracy	3.33
2	The feedback I got was useful.	Perceived Usefulness	3.67
3	The coach supported my growth as a teacher.	Empowerment	4.00
4	I would recommend this tool to other teachers.	Adoption Intention	4.00
5	I see myself continuing to use this tool in the future.	Long-Term Use Intention	4.33
6	I found the AI's suggestions clear and understandable.	Clarity and Interpretability	4.33
7	I felt confident using the AI teacher coach.	Confidence	4.67
8	I was satisfied with my experience using the AI teacher coach.	Satisfaction	3.83
9	The coach's suggestions aligned well with my teaching goals and context.	Relevance and Goal Alignment	3.67
10	My views on AI in education have changed after using the AI teacher coach.	Belief Change	4.00
	Average Usefulness Score on 5		3.98

5.3 *Suggestions for Improvements*

Answers to qualitative questions were categorised into positive and negative aspects of the coach, and counted for frequency. Table 4 lists the five most common positives and negatives about the GenAI coach design and use.

Table 4. *Top Five Positive and Negative Aspects of the GenAI Coach*

Positive Aspects	Negative Aspects
High potential for self-improvement in teaching practices	Long onboarding process
Confidentiality and privacy of feedback	No record of student talk or classroom video
Personalised, class specific feedback with evidence	Analysis takes too long (20 minutes on average per class)
Feedback is unbiased	Feedback should align to specific lesson goals, not general pedagogy
Feedback is offered in real time	Questionable use in student-centred classroom pedagogy

6. Discussion

The study highlights the promise of generative AI to design a teacher coaching tool. Usability scores (70.83) exceeded the SUS benchmark of 68 (Bangor et al., 2008), and usefulness ratings (3.98/5) suggest teachers found value in the experience. Teachers particularly appreciated unbiased, non-judgmental feedback. The private, self-directed format also offered a safe space for reflection.

Some challenges with GenAI coach use were captured. Teachers noted slow analysis. The AI captured only teacher talk, overlooking classroom dynamics such as non-verbal cues and student interactions, highly relevant in student-centered classrooms where student voice is important. Participants also requested better alignment between lesson goals and feedback,

suggesting a need for customizable, content-specific objectives beyond the Danielson Framework (Danielson, 2011). Extending the coach to pre-class planning support could strengthen the cycle of planning, implementation, and reflection (Knight, 2007).

Limitations include a small sample (three teachers from one school), short study duration, and reliance on self-reported data. Future research should involve larger, more diverse samples, integrate usage logs and classroom observations, and assess impacts on long-term practice and student outcomes, as well as differences across teacher proficiency and subjects. Further research has been planned to gauge accuracy and precision of the GenAI coach feedback in comparison to an expert human coach.

7. Conclusion

This study explored the potential of a GenAI teacher coach for scalable, personalized professional development. The findings suggest that the current application achieves moderate to good usability and is perceived as useful by teachers. The findings suggest that generative AI-based coaching applications have significant potential to supplement traditional professional development approaches, particularly in low-resource contexts.

References

Bandura, A. (1977). *Social learning theory*. Prentice Hall.

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction*, *24*(6), 574-594. https://doi.org/10.1080/10447310802205776

Bush, R. N. (1984). Effective staff development in making schools more effective: Proceedings of three state conferences. Far West Laboratory.

Costa, A. L., & Garmston, R. J. (2002). *Cognitive coaching: A foundation for renaissance schools* (2nd ed.). Christopher-Gordon Publishers.

Danielson, C. (2011). *Enhancing professional practice: A framework for teaching* (2nd ed.). ASCD.

Joyce, B., & Showers, B. (1980). Improving in-service training: The messages of research. *Educational Leadership*, *37*(5), 379-385.

Kamphorst, B. A. (2017). E-coaching systems: What they are, and what they aren't. *Personal and Ubiquitous Computing*, *21*(4), 625-632.

Knight, J. (2007). *Instructional coaching: A partnership approach to improving instruction*. Corwin Press.

Knight, D. (2012). Assessing the Cost of Instructional Coaching. Journal of Education Finance. 38. 52-80. 10.1353/jef.2012.0010.

Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, *88*(4), 547-588. https://doi.org/10.3102/0034654318759268

Reddy LA, Shernoff E, Lekwa A. A randomized controlled trial of instructional coaching in high-poverty urban schools: Examining teacher practices and student outcomes. J Sch Psychol. 2021 Jun; 86:151-168. https://doi: 10.1016/j.jsp.2021.04.001.

Sweeney, D. (2011). *Student-centered coaching: A guide for K–8 coaches and principals*. Corwin Press.

Terblanche, N. H. D. (2020). A design framework to create artificial intelligence coaches. *International Journal of Evidence Based Coaching and Mentoring*, *18*(2), 152-165. https://doi.org/10.24384/b7gs-3h05

Terblanche, N. H. D., & Tau, T. (2024). Exploring the use of a goal-attainment, artificial intelligence (AI) chatbot coach to support first-time graduate employees. *Career Development International*, *39*(3). https://doi.org/10.1177/09504222241287090

Learning Sciences in Action: A Reflective Case on the Pedagogy of a Learning Sciences Course for Postgraduates

DurgaPrasad KARNAM[a*] & Amrit Pal SINGH[a]
[a]*Centre for Educational Technology, Indian Institute of Technology Bombay, India*
*karnamdpdurga@gmail.com

Abstract: Learning Sciences (LS) is an emerging field shaping education research and addressing key divisions in knowledge societies. While its global presence is expanding, teaching LS remains challenging due to the need to question assumptions about learning and to engage with abstract, complex theories. This paper reports on a postgraduate LS course in an Indian institute, demonstrating how such theories can be made accessible. The course employed a conceptual scaffold—the concrete–abstract spectrum—to position learning theories ranging from behaviourism and constructivism (including social constructivism and constructionism) to the emerging 4E paradigm of cognition, which emphasises embodied and situated learning. The design incorporated (a) starting from student motivations, (b) introducing theories through an Action–Reflection–Discussion–Reflection (ARDR) cycle to ground engagement, (c) fostering peer discussion and collective reflection, and (d) assessment milestones culminating in a field observation project that initiated students into LS research practice. Drawing on observations and student feedback, we discuss the efficacy of these elements in addressing key pedagogical challenges and offer insights into advancing LS scholarship and course design.

Keywords: Teaching learning sciences, learning sciences, pedagogy

1. Introduction

Learning Sciences (LS), an emerging and highly interdisciplinary field, draws from education, didactics, cognitive science, psychology, computer science, social sciences, philosophy, engineering, and educational technology. Positioned at the intersection of research and practice, LS addresses key divides in contemporary knowledge society — such as formal vs informal learning, authentic vs inauthentic contexts, and laboratory science vs design research (Packer & Maddox, 2016). With technology transforming how we learn and interact, LS offers critical insights for advancing knowledge societies. It also reflects broader paradigm shifts in science — from positivist behaviourism and cognitivism to post-cognitivist, situated, and interpretative approaches (Rourke & Friesen, 2006) — and the rise of design-based research. Learning sciences programmes, largely at master's and PhD levels, are growing worldwide. A review of 75 programmes (Sommerhoff et al., 2018) found common topics such as how people learn, supporting learning, disciplinary learning, technology-enhanced learning, computer-supported collaboration, and research methodologies. While concentrated in North America and Europe, LS programmes are expanding globally, including in India. Expanding scholarship in underrepresented regions is vital, especially for countries like India with a young population and significant education sector needs.

We conducted a semester-long LS course for postgraduate and PhD students in educational technology, who came from diverse academic backgrounds and varying exposure to education. Since they also take a separate research methodology course, our goal was to help them: (a) understand major learning theories and shifts in how the mind and science are conceptualised, and (b) apply these theories to real-world learning situations. Key challenges included: (1) Students' preconceptions of learning shaped by personal and popular narratives,

which they must critically examine (Tillema, 1994; Joram & Gabriele, 1998); (2) The abstract nature of learning theories, requiring careful contextualisation and nuanced understanding. (3) Navigating institutional constraints that shape course design and delivery. This paper examines how the course addressed key learning challenges while familiarising students with major learning theories, their applications, and underlying philosophical shifts. We report insights on the scope of LS, pedagogical approaches, student responses, and critical reflections from teaching a semester-long course to postgraduate and PhD students. After describing the course design, we: (i) outline the coverage of learning theories; (ii) present the concrete–abstract (CA) spectrum as a scaffold for connecting theories with applications; and (iii) detail the Action–Reflection–Discussion–Reflection (ARDR) cycle and field observation project. We evaluate the ARDR approach through student feedback, reflections, and instructor observations, yielding frameworks and strategies that can guide future LS courses.

2. Description of the Course

The course was organised in three phases:

- Phase 1 (8–10 weeks): Core learning sciences, covering foundational theories, their empirical basis, limitations, and relevance.
- Phase 2 (2 weeks): Widening learning to include motivation, emotion, and identity.
- Phase 3 (2–3 weeks): Application, where students observed real-world learning and applied theories to interpret it.

Methodological aspects were excluded, as students took a parallel research methodology course. After introducing LS and expanding the “what of learning” beyond formal contexts, Phase 1 explored the “how” of learning through three clusters of theories: (i) early theories—Skinner’s behaviourism and Tolman’s cognitivism; (ii) constructivism and related approaches—Piaget, Vygotsky, Papert, and Engeström; and (iii) field models of cognition—Gibson’s ecological psychology and 4E approaches, including situated and socio-cultural learning. This trajectory mirrors shifts in models of the human mind.

Discussions deliberately broadened the “who” of learning, from animals and infants to adults, collectives, systems, and even machines (given AI’s role). Where relevant, underlying philosophical and methodological shifts in science were integrated, though at times challenging for some master’s students. These philosophical strands connected with distinctions noted by Packer and Maddox (2016) and were scaffolded using the concrete–abstract (CA) spectrum, discussed in the following section.

2.1 Concrete–Abstract Spectrum: A Conceptual Scaffold



Figure 1. Concrete-abstract spectrum: A conceptual scaffold.

The concrete–abstract (CA) spectrum, illustrated with a real apple on one end and an abstract representation on the other (Fig. 1), was used as a recurring lens to interpret each learning theory and situate students’ own experiences. This scaffold highlighted how theories make assumptions about what counts as “knowledge” and how it is accessed. This anchored reflections across weeks, covering multiple paradigms of learning theories. For instance, this spectrum can be traced to Piaget’s developmental stages, moving from concrete to formal stages. Further, it helps while discussing Papert’s constructionism and Vygotsky’s social constructivism, where the capacity to interact and manipulate with tools is extended even to

the symbolic world, and also how speech and written language make abstractions accessible, citing Ong & Hartley (2013) and Kirsh (1995). The framework reached its richest interpretation during discussions on the 4E paradigm, from Gibson's affordances to non-dualist, anti-representational stances. Beyond theories, it also supported reflections on formal/informal and authentic/inauthentic contexts of learning, and across domains such as sports, music, and STEM, offering an accessible way into philosophical discussions on learning, knowledge, and mind.

2.2 ARDR Cycles: A Pedagogical Framework

Weekly sessions followed an Action–Reflection–Discussion–Reflection (ARDR) cycle, with students organised into mixed teams of three.

- *Action*: Each week, teams explored a topic using readings, multimedia resources, and their prior experiences. This stage involved an active encounter between existing knowledge and new ideas, often supported by digital and AI tools, with the expectation that prior conceptions would be affirmed, refined, or challenged.
- *Reflection (pre-class):* Individually, students submitted structured reflections addressing prompts such as the theory's central claim, links to their own experiences, connections to other paradigms, persuasiveness, and open questions. This process encouraged critical engagement and progressive refinement of thinking.
- *Discussion*: Weekly three-hour sessions combined (i) peer discussions between action teams, where students debated agreements and concerns, and (ii) whole-class discussions led by the instructor, which clarified contradictions, highlighted philosophical and application-related issues, and situated ideas within the broader LS discourse. These dialogues, built on their reflections, ensured a shared baseline of understanding.
- *Reflection* (*post-class):* Students then wrote post-class reflections, drawing on whole-class discussions and asynchronous exchanges on Moodle and WhatsApp. These provided opportunities for consolidation, peer interaction, and instructor feedback.

2.3 Field Observation Project

In the final phase, teams of 3–4 conducted field studies applying LS theories to real-world learning. Each project involved: (i) a pre-observation plan defining the learning setting, theoretical lens, and logistics; (ii) field observation; and (iii) analysis and presentation to faculty. Chosen contexts included a football coaching session, an entrepreneurship workshop, classrooms in experimental and public schools, and a higher education makerspace. This staged progression—from theory to structured reflection to applied observation—helped students move from abstract reasoning to concrete practice. Weekly ARDR cycles encouraged engagement with learning theories, while reflections and discussions helped connect them to students' reasoning. The field observation project extended this work by allowing students to apply theories in real contexts and practice core LS research skills.

3. Methodology, Observations and Findings

We analysed data from weekly assignments, reflections, and participant feedback to examine perceptions of course content, pedagogy, and the ARDR cycles. Ethical protocols, including consent for anonymised use of data, were followed. The course enrolled 19 students (plus one auditor and two PhD observers) from diverse fields such as entrepreneurship, educational technology, and computer science; 52% were master's students and 48% doctoral. About one-third had no prior exposure to learning sciences. All 19 completed the post-course survey. Seventeen students submitted motivation statements, which clustered into four themes: (i) Understanding learning processes, driven by intellectual curiosity—for instance, "deeper insights into how learning theories, cognition, and social interactions influence the learning process"; (ii) Designing learning environments, particularly among educational technology

students who aimed “to create impactful learning environments” or “design a new tool that fosters curiosity and engagement”; (iii) Enhancing pedagogical strategies, especially among those with teaching experience, such as a physics instructor who saw the course as “an opportunity … to address such challenges through evidence-based approaches”; and (iv) Engaging in learning sciences research, with doctoral students aspiring “to contribute to the field of learning science and entrepreneurial capacity building by coming up with a fresh approach… based on sound principles of learning sciences.”

3.1 Overall Course Experience, Relevance and Effectiveness

Participant feedback on the overall course experience was highly positive, with an average rating of 4.32/5 (Fig. 2). The course content was perceived as relevant, clear, and aligned with personal goals (M = 4.21). Pedagogical effectiveness and the ARDR cycle received slightly lower but still favourable ratings (3.58 and 3.84), with 50–60% rating them above 4. Some Master’s-level students, however, reported challenges in adapting to the design, pointing to a need for additional scaffolding for beginners.

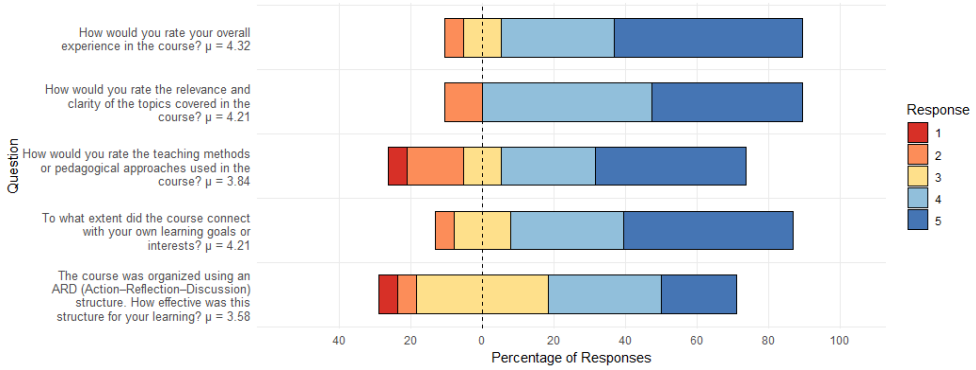


Figure 2. Distribution of responses on overall course experience, relevance & effectiveness.

3.2 Action-Reflection-Discussion-Reflection Cycles

3.2.1 Action: The Role of Prior Readings & Exploration Tasks

The Action phase was rated central by 58% of participants, who valued engaging with readings beforehand for connecting theory with personal experience and field observations. One student explained, “I applied [the theory] to a personal level and saw how my personal experiences included what was there in the learning theory.” Others noted that the flipped design “allowed a longer period to engage with the topic” and was “time-saving since I was not tabula rasa while entering the class.” At the same time, some—particularly master’s students with less experience reading research—struggled with the volume and complexity of texts. A common concern was that “30+ documents over the semester” felt overwhelming. As one participant noted, “I felt confused at multiple points… the readings were overwhelming and did not fit well with my learning style.” Students recommended lighter or summarised readings and introductory guidance before tackling dense theoretical texts. While action teams were intended to provide scaffolding, logistical challenges sometimes hampered their functioning. Yet, when teams worked well, they strengthened peer learning and supported less experienced participants. Overall, the Action phase was foundational, enabling deeper classroom discussions and conceptual grounding, though additional scaffolding could improve accessibility.

3.2.2 Reflection: The Role of Pre & Post-Session Reflections

The Reflection phases (pre- and post-class) were rated least influential by participants, though analysis suggests they played a critical role in fostering readiness. Weekly reflections provided insights into students’ confusions and connections, serving both as feedback to instructors

and a structured space for students to articulate responses. The quality of questions students posed illustrates their critical engagement. For instance, they asked: “Does behaviourism undermine intrinsic motivation? Can it adequately address higher-order thinking?”, After Gibson’s ecological psychology: “How does prior knowledge or culture influence what we perceive as affordances?”; after 4E cognition: “How do we design external representations that enhance rather than hinder cognition?” Though not always recognised by students as impactful, these reflections helped consolidate understanding and seed deeper discussions in class and online forums.

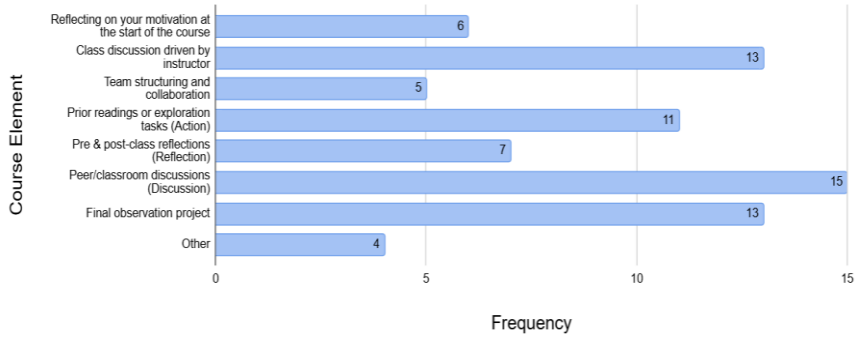


Figure 3. Course elements that were found to be the most influential to students’ learning.

3.2.3 Discussion: The Role of Peer-Led and Instructor-Driven Discussions

Among the course elements, peer group discussions during the first hour of each session emerged as the most influential for learning (Fig. 3). They fostered collaborative meaning-making, active engagement, and a sense of community by giving everyone space to participate. Students noted that peer exchanges helped them uncover nuances and clarify doubts: “Peer interactions helped me get a different perspective on the readings” and “helped to check on if any points were left out.” While widely appreciated, discussions sometimes digressed or caused confusion. As one student observed, “Sometimes I felt that our discussion was going in another direction. We need to narrow down, but yes, narrowing down may not help us to explore everybody's thinking. In the beginning I felt some discussion was meaningless.” Another added, “Sometimes I used to get confused because of multiple interpretations but when the professor explained what exactly it wanted to say.” These limitations were mitigated by instructor-led discussions, which clarified misconceptions and refined interpretations. Students noted, “The class discussion, solidified, my perspective, and sometimes clarified if I had any issues into understanding some parts of the discussion,” and “Sometimes I would take the reading in a different perspective than what was discussed in the class. The class discussion would give me a scaffold and support to understand the nuances of it.” Thus, the discussion phase of the ARDR cycle synthesised ideas at two levels: exploratory peer dialogue with multiple perspectives, and collective convergence led by the instructor. This dual process not only clarified philosophical issues but also strengthened the class as an authentic community of practice.

3.3 Assessment of Learning Through the Field Observation Project

Learning was assessed in two stages. The mid-semester (2–3 hours) required participants to consolidate their understanding of learning and critically analyse any learning tool, seeding theoretical application midway through the course. The final assessment, a field observation project, was identified as a key element, with 68% reporting it significantly deepened their understanding (Fig. 3). Students applied theories to real-world encounters, noting it “helped me connect everything to the real world” and was “the best way to understand the theory learnt in class.” One called it “a major takeaway of the course,” while others valued the structure—from framing to reports—for integrating theoretical and empirical analysis. A suggestion was to start the project earlier to allow more time for engagement. The project served as a good

anchor, allowing the students to ground the abstract theories and philosophical clarifications into concrete real-world observations in varied contexts: football sessions, maker sessions, regular class sessions. Learning sciences researchers recognised the final presentations for combining contextual breadth with theoretical depth, noting this was achieved within a semester despite students' limited prior exposure.

4. Discussion

This course demonstrated both strengths and areas for improvement in teaching learning sciences. The ARDR cycle supported students in questioning pre-existing notions: readings in the Action phase encouraged sense-making and enquiry, while reflection assignments—though seen as less influential—still showed critical engagement. Peer and instructor-led discussions enabled collective churning of ideas. Conceptual scaffolds, such as the concrete–abstract (CA) spectrum, and the field observation project made abstract theories accessible by linking them to real-world contexts (e.g., football, maker sessions). Presentations reflected this accessibility through the depth of student analysis. The course also balanced institutional constraints with workload and motivation, addressing key pedagogical challenges with varying success. Students' appreciation of discussions and projects, alongside difficulties with readings, echoes findings from a modified flipped classroom study (Chiang & Chen, 2017). Our study extends this by examining pedagogical elements in greater depth.

At the same time, areas needing attention remain. The course began challenging assumptions but did not fully assess conceptual change in students' notions of learning. Some continued to struggle with abstraction in the Action phase—suggesting that more structured action-teams could provide additional scaffolding. Future work should explore these possibilities. We hope that this reflective case on the Learning Sciences course can trigger more conversations in the Learning Sciences community.

Acknowledgements

We thank the students for their participation and contributions, the field observation participants, and CET members for feedback on an earlier draft. We also acknowledge the use of AI assistance in shortening sections of this paper.

References

Chiang, F., & Chen, C. (2016). Modified Flipped Classroom Instructional model in "Learning Sciences" course for graduate students. The Asia-Pacific Education Researcher, 26(1–2), 1–10. https://doi.org/10.1007/s40299-016-0321-2

Joram, E., & Gabriele, A. J. (1998). Preservice teachers' prior beliefs: Transforming obstacles into opportunities. Teaching and Teacher Education, 14(2), 175–191. https://doi.org/10.1016/s0742-051x(97)00035-8

Kirsh, D. (1995). The intelligent use of space. Artificial Intelligence, 73(1–2), 31–68. https://doi.org/10.1016/0004-3702(94)00017-u

Ong, W. J., & Hartley, J. (2013). Orality and literacy. Routledge.

Packer, M. J., & Maddox, C. (2016). Mapping the territory of the learning sciences. In Cambridge University Press eBooks (pp. 126–154). https://doi.org/10.1017/cbo9781107707221.007

Rourke, L., & Friesen, N. (2006). The learning sciences: the very idea. Educational Media International, 43(4), 271–284. https://doi.org/10.1080/09523980600926226

Sommerhoff, D., Szameitat, A., Vogel, F., Chernikova, O., Loderer, K., & Fischer, F. (2018). What do we teach when we teach the learning sciences? A document analysis of 75 graduate programs. Journal of the Learning Sciences, 27(2), 319–351. https://doi.org/10.1080/10508406.2018.1440353

Tillema, H. (1994). Training and professional expertise: Bridging the gap between new information and pre-existing beliefs of teachers. Teaching and Teacher Education, 10(6), 601–615. https://doi.org/10.1016/0742-051x(94)90029-9

Analysis and Comparative of the Academic Draft on Coding and Artificial Intelligence Education in Indonesia Primary and Secondary Schools

Riska SAPUTRA[a]* & Adventhius Immanuel KARO-KARO[b]
[a]*Graduate Institute of Digital Learning and Education, National Taiwan University of Science and Technology, Taiwan*
[b]*Legal Studies Program, Faculty of Law, University of Indonesia, Indonesia*
*riskysaputra17@gmail.com

Abstract: This study evaluates Indonesia's 2025 Academic Draft on Coding and Artificial Intelligence (AI) Education for primary and secondary schools through a qualitative policy review. Using SWOT analysis, cross-national comparison with India, and benchmarking against UNESCO's AI Competency Framework (2024), the analysis highlights both philosophical strengths and operational weaknesses. The draft demonstrates strong alignment with national visions, alongside a flexible modular curriculum and early integration of coding and AI. It also places value on ethical awareness and computational thinking, showing potential to prepare students for a digital future. However, the findings reveal critical challenges in implementation. The draft lacks a phased roadmap, structured teacher competency framework, enforceable data ethics guidelines, and targeted strategies to reduce the digital divide in underserved "3T" regions. Comparative insights from India suggest that modular teacher training programmes, adaptive assessment systems, and public–private partnerships offer practical models Indonesia could adapt. The study concludes that while the draft presents a progressive vision, its success depends on translating ambition into actionable, context-sensitive policies. Strengthening teacher readiness, equity, and governance is essential to ensure inclusive and sustainable AI and coding education.

Keywords: Artificial Intelligence education, coding curriculum, policy analysis, AI Indonesia, educational equity, comparative education, teacher readiness, AI literacy

1. Introduction

The rapid integration of Artificial Intelligence (AI) is reshaping education, with the World Economic Forum (2023) predicting that workforce demands will change by 2027, requiring early digital competence. UNESCO's AI Competency Framework 2024 emphasizes AI literacy, computational thinking, and ethical citizenship (Mutawa & Sruthi, 2025), and many countries have embedded AI and coding in school curricula. While Singapore and India have advanced policies, Indonesia faces a shortage of nine million skilled workers by 2030 (World Bank, 2021) and has only recently introduced the 2025 Academic Draft on Coding and AI Education. The draft outlines curriculum integration from Grade 5, teacher training, and ethical awareness (Basyir, 2014), but lacks clear strategies for underserved *"3T area" (Frontier, Outermost, Underdeveloped)* regions, data ethics, child safety, and teacher readiness. These regions face specific challenges, including a lack of reliable digital infrastructure and insufficient support for educators, which could widen the digital divide. Without an enforceable framework and targeted solutions like investments in inclusive infrastructure and context-sensitive teacher development programs, the scalability and policy alignment of the draft remain limited. This study employs SWOT, comparative analysis with India, and a gap analysis against international benchmarks to assess the draft's strengths, weaknesses, and opportunities for improvement, with a focus on proposing actionable recommendations to bridge these gaps.

2. Literature Review

2.1 Frameworks for Analyzing AI in Education Policies

A growing body of research highlights the importance of analyzing both national and supranational AI policies, particularly in education, as countries differ in how they frame AI competence development within their strategies (Wang et al., 2025; Shi, 2025), offering valuable lessons for Indonesia. Ethical considerations are central to this debate, with Nguyen et al. (2023) and Badawy et al. (2024; 2025) stressing the need for transparent frameworks that safeguard data privacy, ensure algorithmic accountability, and guide responsible AI use in learning environments, all of which must be adapted to the developmental and cultural contexts of young learners. Equally important are competence development and curriculum design, where tools such as the TAICS scale (Chiu et al., 2025) can assess teacher readiness and self-determination theory can enhance educators' motivation to strengthen digital skills. Practical insights, such as Dai et al.'s (2023) collaborative curriculum model, demonstrate how international approaches can be localized, while systemic challenges including assessment integrity, institutional preparedness (Luo, 2024; Jiang et al., 2025), and infrastructure expansion (Williamson, 2019; Luan et al., 2020) must also be addressed. For Indonesia, the literature points to six urgent priorities: scalable teacher training, age-appropriate ethical guidelines, coherent curriculum integration, safeguards against academic dishonesty, equitable access across diverse regions, and long-term policy evaluation. A SWOT analysis informed by these considerations can help reveal gaps and support the development of an inclusive, ethical, and context-sensitive AI and coding education framework for primary and secondary schools.

2.2 Perspectives on AI and Coding Education

The integration of Artificial Intelligence (AI), particularly generative AI tools such as ChatGPT, into coding education offers both transformative opportunities and pressing challenges. Recent studies indicate that AI can support personalized learning by providing real-time feedback, generating code examples, and offering corrective guidance. For instance, Haindl and Weinberger (2024) and Cubillos et al. (2025) show that AI tools, when used effectively, enhance student autonomy and engagement while reducing perceived cognitive load. Similarly, Li et al. (2024) report that AI-driven feedback and automated assessment deliver consistent and accurate results, enabling teachers to concentrate on higher-order instructional tasks. These findings suggest that AI has the potential to expand access to quality coding education when it is thoughtfully integrated into curriculum design. However, concerns persist regarding plagiarism, overreliance on AI-generated outputs, and the potential erosion of critical thinking and academic integrity (Rahman & Watanobe, 2023; Güner et al., 2025). Reliability issues also remain, especially in complex areas such as object-oriented programming (Fernandez-y-Fernandez et al., 2024), and Ouyang et al. (2024) caution that outcomes often vary between human-guided and AI-assisted instruction, highlighting the need for human oversight. Inequalities in digital access and AI literacy further risk widening educational divides, particularly in socioeconomically diverse settings (Zipf et al., 2025; Wang et al., 2024). In the Indonesian context, these challenges are especially relevant as the country begins to implement AI and coding education in schools. The effectiveness of these initiatives will rely on sustained teacher professional development (Moorhouse & Kohnke, 2024), culturally responsive curriculum design (Dai et al., 2023), and policies that promote ethical and equitable AI use. Addressing linguistic and infrastructural diversity will also be crucial, as success depends on balancing technological innovation with inclusive practices to improve learning outcomes for all students.

3. Method

This study addressed three central research questions: (1) What are the strengths, weaknesses, opportunities, and threats (SWOT) in Indonesia's AI and coding education policy? (2) How does Indonesia's approach compare to that of India? and (3) What gaps remain unaddressed within the current academic framework? To answer these questions, a qualitative design was employed, combining content analysis, comparative case study, and critical literature review. The main source of data was the Indonesian Ministry of Education's academic draft *Naskah Akademik Pembelajaran Koding dan Kecerdasan Artifisial Pada Pendidikan Dasar dan Menengah* (Academic Draft on Coding and Artificial Intelligence Learning in Primary and Secondary Education)[1], which was examined through iterative readings to extract policy-relevant insights. For cross-national comparison, official documents from India, including the National Education Policy 2020[2], the National Strategy for AI 2018[3], and related programs, were analyzed to highlight similarities and differences. The gap analysis was benchmarked against global standards, specifically the UNESCO AI Competency Framework for Students and Teachers (UNESCO, 2024)[4], which outlines knowledge, skills, and ethical considerations necessary for AI literacy. This framework provided a reference point for assessing the Indonesian draft's coverage of technical, ethical, and pedagogical dimensions. Gemini 2,5 Pro, a Generative AI tool, was used in this study to assist with comparative policy analysis. Specifically, the prompt *"Comparative Analysis: Indonesia and India in AI and Coding Education Policy" was* used to generate structured comparisons, which were then critically reviewed and validated by the authors against official policy documents.

4. Result

4.1 SWOT Analysis of Academic Draft Indonesia's AI and Coding Education Policy



Figure 1. SWOT Analysis AI and Coding Education Policy

[1] https://kurikulum.kemdikbud.go.id/file/1741766787_manage_file.pdf
[2] https://www.education.gov.in/sites/upload_files/mhrd/files/NEP_Final_English_0.pdf
[3] https://www.niti.gov.in/sites/default/files/2023-03/National-Strategy-for-Artificial-Intelligence.pdf
[4] https://www.unesco.org/en/articles/ai-competency-framework-teachers and https://unesdoc.unesco.org/ark:/48223/pf0000391105 (ai competency framework student)

4.2 Comparative Analysis: Indonesia and India

Table 1. *Comparison Between Indonesia and India*

Feature	India	Indonesia
National Strategy/Policy	- National Education Policy (NEP) 2020: Explicitly includes AI & coding. - National Strategy for AI (2018): Identifies education as key. - "India AI Mission" & "Viksit Bharat 2047": Strong government push for AI leadership.	- Finalizing AI & Coding Curriculum: Targeting 2025/2026 rollout. - "Indonesia Artificial Intelligence National Plan": Focus on AI-driven learning. - Recent high-level push: Aiming to compete globally.
Curriculum Integration	- Coding from Grade 6: As per NEP 2020. - AI as Elective: CBSE offers it in secondary & senior secondary. - Focus: Experiential learning, real-world projects.	- Elective Subjects: AI & coding from Grade 5/6 (Elementary) to High School. - Approach: Blend of device-based and non-device-based learning planned. - Phased Rollout: Likely starting in select 'project schools'.
Teacher Training	- Recognized Challenge: Significant gap in AI-trained teachers. - Initiatives: CBSE training, IBM Skills Build, Google AI for Teachers, Microsoft. - Focus: Upskilling existing teachers, EdTech platform involvement.	- Major Partnership: Google to train at least 1 million teachers in AI & coding (like Gemini Academy). - Involvement: Leveraging private universities & coding providers. - Focus: Rapid, large-scale upskilling via collaboration.
Digital Infrastructure & Access	- "Digital India" Program: Aims for nationwide connectivity & literacy. - Challenge: Ensuring last-mile connectivity and device access, especially in rural areas.	- Digital Literacy Programs: Government initiatives, often with partners. - Challenge: Bridging the digital divide, particularly across its vast archipelago. - Risk Focus: Addressing misinformation & online safety.
Industry Partnerships	- Thriving EdTech Ecosystem: Many platforms offering AI/coding education. - Multiple Collaborations: Government, academia, and industry (MNCs & domestic) work together.	- Strategic Collaborations: Key partnership with Google is central. - Open to Providers: Welcomes involvement from coding service providers and developers. - University Links: Encouraging ties with industry.
Key Challenges	- Infrastructure Gaps - Teacher Readiness Scale - Curriculum Standardization & Flexibility - Equity & Inclusion (Urban-Rural Divide) - Ethical Considerations (Bias, Privacy)	- Infrastructure Gaps (esp. remote islands) - Teacher Readiness Scale - Curriculum Implementation & Standardization - Equity & Inclusion (Access) - Ethical Considerations
Overall Approach & Outlook	- Early Mover: Formalized policy structure. - Leverages Domestic Strength: Strong IT & EdTech sectors. - Goal: Global AI leadership.	- Rapid Mover: Determined push to implement quickly. - Leverages International Partnerships: Key strategy for scaling. - Goal: Global competitiveness & "Golden Indonesia 2045".

4.3 Gap Analysis of the Academic Draft

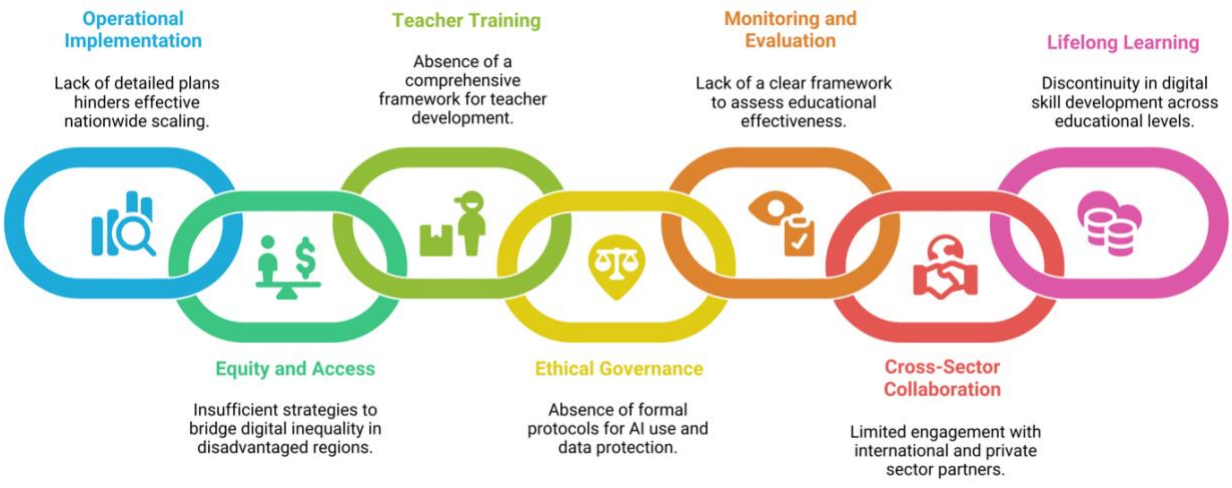


Figure 2. Gap Analysis

5. Discussion and Conclusion

The findings confirm that Indonesia's 2025 Academic Draft on Coding and Artificial Intelligence Education is a forward-looking policy aligned with RPJPN 2025–2045 and Society 5.0, reflecting strong philosophical grounding in Pancasila and a modular curriculum design that is both adaptable and pedagogically relevant (Mutawa & Sruthi, 2025; UNESCO, 2023; Basyir, 2014). Nevertheless, the analysis reveals significant gaps in operational readiness. While the draft acknowledges the importance of teacher training, it lacks a structured competency framework with measurable indicators for teacher preparedness and student outcomes. This gap is evident when compared with India's centralized, certification-based programmes, which provide clear benchmarks for scalability (Wang et al., 2025). Similarly, the policy's commitment to equity is weakened by the absence of concrete strategies to bridge the digital divide in "3T" (Terdepan, Terluar, Tertinggal) regions, raising concerns of deepening educational inequality (Tiwari et al., 2021). Furthermore, its broad mention of data ethics and digital safety remains insufficient without enforceable legal guidelines, potentially limiting both public trust and effective protection for learners.

This study therefore argues that the draft's ambitious vision can only be realised through a stronger emphasis on implementation and equity. Recommended measures include adopting a phased implementation roadmap, establishing certified teacher competency frameworks adapted from international models, investing in infrastructure for low-connectivity regions, and introducing enforceable data ethics guidelines to safeguard learners and build public confidence (UNESCO, 2023). Such reforms would ensure that foundational readiness—particularly teacher training and digital inclusion—precedes large-scale deployment. While the study provides valuable insights, it is limited by its reliance on secondary policy documents and the absence of primary data from teachers, students, and policymakers, which restricts perspectives on local readiness. Future research should therefore combine policy analysis with field-based evidence to strengthen contextual understanding. In conclusion, the draft represents a progressive step toward digital transformation, but its success depends on translating visionary intent into actionable, context-sensitive strategies that guarantee inclusivity and sustainability.

6. Acknowledgement

We thank the Ministry of Education of Indonesia for providing the 2025 Academic Draft on Coding and Artificial Intelligence Education in Primary and Secondary Schools, which guided this study. While this article was written in May 2025, the policy has since been formally launched under *Permendikdasmen No. 13 Tahun 2025*, established on 15 July 2025.

References

Badawy, W., Helal, M. M. I., Hashim, A., Zinhom, H., & Shaban, M. (2025). Ethical boundaries and data-sharing practices in AI-enhanced nursing: An Arab perspective. *International Nursing Review*, 72(1), e70013. https://doi.org/10.1111/inr.70013

Badawy, W., Zinhom, H., & Shaban, M. (2024). Navigating ethical considerations in the use of artificial intelligence for patient care: A systematic review. *International Nursing Review.* https://doi.org/10.1111/inr.13059

Basyir, A. (2014). Pentingnya naskah akademik dalam pembentukan peraturan perundang-undangan untuk mewujudkan hukum aspiratif dan responsif. *Jurnal IUS Kajian Hukum dan Keadilan*, *2*(5). https://jurnalius.ac.id/ojs/index.php/jurnalIUS/article/viewFile/171/147

Chang, C.-Y., Lin, H.-C., Yin, C., & Yang, K.-H. (2025). Generative AI-assisted reflective writing for improving students' higher order thinking: Evidence from quantitative and epistemic network analysis. *Educational Technology & Society, 28*(1). https://doi.org/10.30191/ETS.202501_28(1).TP03

Chiu, T. K. F., Ahmad, Z., & Çoban, M. (2025). Development and validation of teacher artificial intelligence (AI) competence self-efficacy (TAICS) scale. *Education and Information Technologies, 30*(5), 6667–6685. https://doi.org/10.1007/s10639-024-13094-z

Cubillos, C., Mellado, R., Cabrera-Paniagua, D., & Urra, E. (2025). Generative artificial intelligence in computer programming: Does it enhance learning, motivation, and the learning environment? IEEE Access, 13, 40438–40455. https://doi.org/10.1109/ACCESS.2025.3532883

Dai, Y., Liu, A., Qin, J., Guo, Y., Jong, M. S.-Y., Chai, C.-S., & Lin, Z. (2023). Collaborative construction of artificial intelligence curriculum in primary schools. *Journal of Engineering Education, 112*(1), 23–42. https://doi.org/10.1002/jee.20503

Fernandez-y-Fernandez, C. A., Sánchez-Soto, E., Cisnero, J. R. A., & Juárez-Ramírez, R. (2024). Exploring the frontier of software engineering education with chatbots. Programming and Computer Software, 50(8), 796–815. https://doi.org/10.1134/S0361768824700774

Haindl, P., & Weinberger, G. (2024). Students' experiences of using ChatGPT in an undergraduate programming course. *IEEE Access, 12*, 43519–43529. https://doi.org/10.1109/ACCESS.2024.3380909

Jiang, L., Wei, B., Qiu, N., & Huang, L. (2025). What influences ChatGPT's adoption and diffusion in Chinese higher education? A study based on Technology-Organization-Environment framework. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2025.2479166

Lee, S. (2025). *5 Key Statistics: AI's Impact on Global Education Policy Trends.* https://www.numberanalytics.com/blog/5-key-statistics-ais-impact-global-education-policy-trends

Li, J., Jangamreddy, N., Bhansali, R., Hisamoto, R., Zaphir, L., Dyda, A., & Glencross, M. (2024). AI-assisted marking: Functionality and limitations of ChatGPT in written assessment evaluation. *Australasian Journal of Educational Technology, 40*(4), 56–72. https://doi.org/10.14742/ajet.9463

Luan, H., Geczy, P., Lai, H., Gobert, J., Yang, S. J. H., Ogata, H., Baltes, J., Guerra, R., Li, P., & Tsai, C.-C. (2020). Challenges and future directions of big data and artificial intelligence in education. *Frontiers in Psychology, 11*, 580820. https://doi.org/10.3389/fpsyg.2020.580820

Luo, J. (2024). A critical review of GenAI policies in higher education assessment: A call to reconsider the "originality" of students' work. *Assessment & Evaluation in Higher Education, 49*(5), 651–664. https://doi.org/10.1080/02602938.2024.2309963

Moorhouse, B. L., & Kohnke, L. (2024). The effects of generative AI on initial language teacher education: The perceptions of teacher educators. *System, 122*, Article 103290. https://doi.org/10.1016/j.system.2024.103290

Mutawa, A. M., & Sruthi, S. (2025). UNESCO's AI Competency Framework: Challenges and Opportunities in Educational Settings. *Impacts of Generative AI on the Future of Research and Education*, 75-96. https://www.igi-global.com/chapter/unescos-ai-competency-framework/358768

Nguyen, A., Ngo, H. N., Hong, Y., Dang, B., & Nguyen, B.-P. T. (2023). Ethical principles for artificial intelligence in education. *Education and Information Technologies, 28*(4), 4221–4241. https://doi.org/10.1007/s10639-022-11316-w

Ouyang, F., Guo, M., Zhang, N., Bai, X., & Jiao, P. (2024). Comparing the effects of instructor manual feedback and ChatGPT intelligent feedback on collaborative programming in China's higher education. *IEEE Transactions on Learning Technologies, 17*, 2227–2239. https://doi.org/10.1109/TLT.2024.3486749

Rahman, M. M., & Watanobe, Y. (2023). ChatGPT for education and research: Opportunities, threats, and strategies. Applied Sciences-Basel, 13(9), Article 5783. https://doi.org/10.3390/app13095783

Shi, L. (2025). Global perspectives on AI competence development: Analyzing national AI strategies in education and workforce policies. *Human Resource Development Review*. https://doi.org/10.1177/15344843251332360

Tiwari, S., Ali, R., Ariadharma, E., Arsana, I., Bennis, L., Beschorner, N., ... & Wibisana, P. S. (2021). *Beyond Unicorns: Harnessing Digital Technologies for Inclusion in Indonesia, World Bank Group*. https://coilink.org/20.500.12592/kx3cpq

Vieriu, A. M., & Petrea, G. (2025). The Impact of Artificial Intelligence (AI) on Students' Academic Development. *Education Sciences, 15*(3), 343. https://doi.org/10.3390/educsci15030343

Wang, S., Zhang, Y., Xiao, Y., & Liang, Z. (2025). Artificial intelligence policy frameworks in China, the European Union and the United States: An analysis based on structure topic model. *Technological Forecasting and Social Change, 212*, 123971. https://doi.org/10.1016/j.techfore.2025.123971

Williamson, B. (2019). Policy networks, performance metrics and platform markets: Charting the expanding data infrastructure of higher education. *British Journal of Educational Technology, 50*(6), 2794–2809. https://doi.org/10.1111/bjet.12849

World Economic Forum. (2023). *The Future of Jobs Report 2023.* https://www.weforum.org/publications/the-future-of-jobs-report-2023/

Zipf, S., Wu, C., & Petricini, T. (2025). Using the information inequity framework to study GenAI equity: Analysis of educational perspectives. *Information Research-An International Electronic Journal, 30*(SI), 533–547. https://doi.org/10.47989/ir30iConf47284

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Configurational Influence of Self-Efficacy, Learning Strategies, and Emotions on MOOC Learning Outcomes: An fsQCA Analysis

Felix Arokiya Raj, A.P.[a*] & Arul Aram, I.[b]
[a] *National Institute of Technical Teachers Training and Research, Chennai, India*
[b] *Anna University, Chennai, India*
*felix@nitttrc.ac.in

Abstract: Measuring learning outcomes in Massive Open Online Courses (MOOCs) is challenging due to the learner diversity and the asynchronous nature of participation. The varying levels of the learner's characteristics, such as self-efficacy, emotions, motivation, learning strategies, and prior knowledge, interact in complex ways to shape the learning outcomes. While ample research exists on the factors influencing outcomes, limited attention has been given to their combined effects. This study employs fuzzy-set Qualitative Comparative Analysis (fsQCA) to investigate the configurations of factors influencing the cognitive, social, and self-growth outcomes in MOOCs. The study uses a dataset of 275 observations to highlight three significant findings: (1) the pathways leading to success differ substantially from those leading to limited outcomes, indicating asymmetry in causal pattern, (2) academic self-efficacy, motivated-learning strategies, and emotions emerge as key interacting conditions that indicate equifinality in achieving successful learning results, and (3) negative emotions, though peripheral, associates with perceived quality and learner satisfaction to support learning achievement. The findings emphasise the value of examining interactive combinations of variables instead of isolated variables and imply practical directions for MOOC design and instructional development.

Keywords: MOOC learning outcomes, self-efficacy, motivated learning strategies, achievement emotions, fuzzy-set Qualitative Comparative Analysis (fsQCA)

1. Introduction

Massive Open Online Courses (MOOCs) have expanded access to higher education by offering scalable and flexible learning options through platforms such as Coursera, edX, NPTEL, and FutureLearn (Guerrero et al., 2021). While democratising education, these online courses also present challenges in achieving consistent learning outcomes due to learner diversity and the self-paced mode of participation (Julia et al., 2021). Prior research highlights the influence of learner characteristics such as motivation, self-regulation, and engagement on academic success in digital environments (Albelbisi et al., 2021).

Despite the vast potential of MOOCs, understanding the factors that drive successful learning outcomes remains complex. Among these factors, **academic self-efficacy (ASE)**, or the belief in one's ability to achieve learning outcomes, is central to persistence and achievement (Allen et al., 2022; Bandura & Adams, 1977). Likewise, **achievement emotions**, conceptualised by the control-value theory (Pekrun, 2006), play a vital role in academic performance. This theory postulates that positive emotions, such as enjoyment and pride, enhance learning engagement, and negative emotions influence learning outcomes in complex ways (King & Areepattamannil, 2014; Pekrun & Linnenbrink-Garcia, 2012).

A third important construct is **motivated learning strategies (MLS)**, which integrate cognitive, emotional, and motivational processes such as goal setting, monitoring, and self-regulation, enabling learners to adapt effectively to online education (Pintrich et al., 1993). Unlike general learning strategies, MLS explicitly connect effort and persistence to

motivational beliefs, making them particularly critical in self-directed online environments. Together, ASE, achievement emotions, and MLS play pivotal roles in impacting **MOOC learning outcomes**, which can be conceptualised across cognitive (knowledge gains), social (collaborative participation), and personal growth (self-regulated learning) dimensions.

Although these factors have been studied individually, less attention has been given to the combined and interactive effects of the antecedents in shaping MOOC learning outcomes. Traditional statistical models often overlook such complex configurations of conditions. This study addresses this gap by employing **fuzzy-set Qualitative Comparative Analysis (fsQCA)** (Ragin, 2006) to explore how ASE, emotions, and MLS influence learning outcomes. These findings provide practical insights for course designers and educators seeking to improve learner satisfaction and achievement in large-scale online education.

2. Conceptual Framework

The conceptual framework (Figure 1) positions **academic self-efficacy, achievement emotions, and motivated learning strategies** as independent variables influencing **MOOC learning outcomes**, with additional moderating effects from perceived quality and satisfaction.

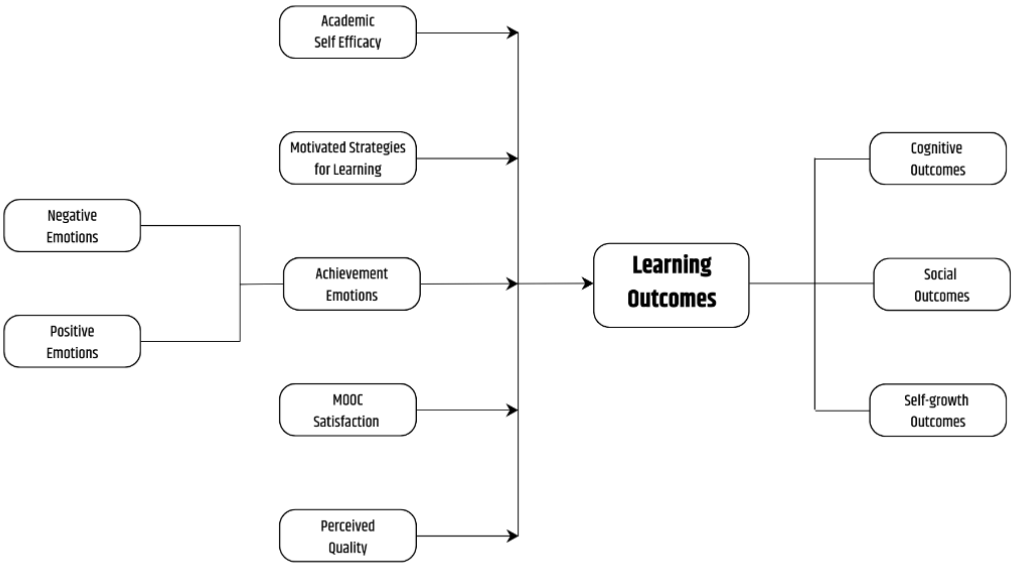


Figure 1. Conceptual framework influencing MOOC learning outcomes.

2.1 Independent Variables and Measures

The following variables were considered for this study based on the literature review of determinants affecting the learning outcomes. These constructs represent psychological and behavioural factors that shape learner engagement and achievements in MOOCs. Each variable was measured using well-established and validated scales to ensure reliability and comparability with prior research.

Academic Self-Efficacy (ASE) refers to learners' confidence in accomplishing academic tasks, influencing engagement and persistence (Honicke & Broadbent, 2016). This study measured ASE using the 8-item Self-Efficacy for Learning and Performance subscale of the Motivated Strategies for Learning Questionnaire (Pintrich et al., 1993). This scale captures the learner's belief in their ability to master course material, perform well on assignments, and achieve high grades, using a 7-point Likert scale format.

Achievement Emotions (AE) are emotional responses to learning activities and their outcomes (Pekrun, 2006). These include both positive emotions (PosEmo), such as enjoyment, hope, and pride, which enhance motivation and satisfaction and negative emotions (NegEmo), such as anxiety, anger, hopelessness, shame, and boredom, which may impede engagement but can also stimulate effort under certain conditions. Negative emotions are further categorised into two types: activating negative emotions (e.g., anxiety, anger, and hopelessness) and deactivating negative emotions (e.g., shame and boredom). A total of 22 items from the shortened standardised Achievement Emotions Questionnaire (AEQ) were employed to assess AE (Bieleke et al., 2021).

Motivated Learning Strategies (MLS) represent the cognitive and metacognitive strategies that help learners to regulate their learning, such as organising content, managing time, rehearsing, and applying problem-solving approaches, all of which contribute to academic success (Credé & Phillips, 2011). The MLS variable was measured using the MSLQ-B subscale, comprising 25 standardised items.

Perceived Quality (PQ) and Satisfaction (SA) were measured using a global, single-item scale to capture learners' overall perception of course design and learning experience. This parsimonious single-item approach was chosen since lengthy instruments may increase dropout rates or survey fatigue (Leung & Xu, 2013).

2.2 Dependent Variable – Learning Outcomes (LO)

Learning outcomes are conceptualised as a multidimensional construct comprising cognitive, social, and self-growth dimensions, providing a comprehensive view of how MOOCs contribute to learner development (Wei et al., 2021).

Each dimension was assessed using a standardised 5-item subscale adapted from (Zhoc et al., 2018), with learners self-evaluating their achievement at the conclusion of the online course.

- **Cognitive outcomes** capture knowledge acquisition, application, and problem-solving skills essential in self-directed online learning (Mayer, 2005).
- **Social outcomes** refer to collaboration, communication, and community-building that support learner engagement and persistence (Greenhow & Lewin, 2016).
- **Self-growth outcomes** emphasise self-regulation, reflection, and personal development, highlighting the transformative role of MOOCs in fostering autonomy and resilience (Dweck, 2007).

3. Methodology

This study employs **fuzzy-set Qualitative Comparative Analysis (fsQCA)**, a case-oriented method suitable for exploring causal complexity (Ragin, 2006). fsQCA identifies multiple pathways called configurations of conditions leading to the same outcome (**equifinality**) and distinguishes between successful and unsuccessful outcomes (**asymmetry**). Configurations are specific sets of causal variables that influence an observed variable of interest. This approach uses set theory and Boolean algebra to assess the complex combination of determinants through comparisons (Fernández-Esquinas et al., 2021).

Data and Sample

Survey data were collected from 275 MOOC participants across three Swayam MOOCs, with the majority having a master's degree and within the 25 to 44 age group. Learning outcomes were assessed through cognitive, social, and self-growth measures. Independent variables included academic self-efficacy (ASE), positive emotions (PosEmo), negative emotions (NegEmo), motivated learning strategies (MLS), perceived quality (PQ), and satisfaction (SA).

Calibration

Variables were calibrated into fuzzy sets using three anchors, namely, 0.95 (full membership), 0.50 (crossover), and 0.05 (full non-membership), based on percentile thresholds. This transformation ensured comparability across constructs.

Analysis Steps

- Necessity analysis tested whether single conditions consistently explained the outcome.
- Truth tables generated all possible condition combinations.
- Logical minimisation identified configurations (recipes) sufficient for high or low learning outcomes.
- The intermediate solutions were identified as the final group of solutions.

4. Results

4.1 Necessary Conditions

Academic Self-Efficacy (ASE) and Satisfaction (SA) emerged as conditions with high necessity consistency (above 0.80), suggesting that strong self-efficacy and a positive course experience often act as fundamental requirements for MOOC success.

4.2 Configurations for High Learning Outcomes

The fsQCA revealed the following multiple sufficient pathways for attaining learning outcomes:

- **Configuration 1:** ASE ● Positive Emotions ● SA → High Learning Outcomes (Learners with strong self-efficacy, enjoyment, and satisfaction achieved consistently significant gains in learning performance.)
- **Configuration 2:** ASE ● MLS ● PQ → High Learning Outcomes (Even in the absence of pronounced emotional states, individuals with strong ASE, effective learning strategies, and perceived course quality attained high outcomes.)
- **Configuration 3:** ASE ● Sub-mean Negative Emotions ● SA → High Learning Outcomes (Manageable levels of negative emotions, e.g., anxiety that motivates persistence, combined with ASE and satisfaction supported achievement.)

4.3 Configurations for Low Learning Outcomes

- **Configuration 4:** Low ASE ● High Negative Emotions ● Low SA → Poor Learning Outcomes
- **Configuration 5:** Low PQ ● Low MLS → Poor Learning Outcomes

The sequence of configurations holds significance, as it reflects the relative magnitude of the outcome variable. Higher-order configurations, such as Configuration 1, correspond to elevated levels of achievement, whereas at the lower end, Configuration 4 exhibits the smallest presence of the study variable, followed by Configuration 5.

5. Discussion

The findings confirm that no single variable ensures MOOC success, but rather, configurations of self-efficacy, emotions, strategies, and course perceptions collectively shape outcomes.

- ASE is a core condition across most pathways, aligning with Bandura's (1977) view that self-efficacy directs effort and persistence.
- Achievement emotions matter, but their role is contextual. Positive emotions (e.g. enjoyment, pride) enhance engagement, while manageable negative emotions (e.g. anxiety) may push learners to invest more effort, which is in line with control-value theory (Pekrun, 2006).
- MLS and PQ act as complementary conditions, even when emotions are mixed, strong strategies and high-quality course design can drive success.
- Satisfaction (SA) appears consistently in high-outcome pathways, reinforcing its role in sustaining engagement and promoting course completion.

These findings demonstrate causal equifinality (different pathways leading to the same outcome) and asymmetry (conditions for outcome success differ from failure), which is central to fsQCA reasoning.

6. Conclusion

This study highlights the importance of configurations of learner and course-related factors in shaping MOOC learning outcomes. Self-efficacy and satisfaction consistently emerge as central drivers, but their impact is amplified when combined with emotional states, motivated strategies, and perceptions of course quality. Importantly, when moderate, negative emotions may not always be detrimental but can contribute to persistence and achievement. By applying fsQCA, this research moves beyond linear models to show how different constellations of factors can yield successful or unsuccessful outcomes.

7. Practical Implications

- For MOOC designers: Enhance learner satisfaction through clear structure, interactive design, and feedback mechanisms.
- For instructors: Foster academic self-efficacy via scaffolding, timely support, and self-regulation prompts.
- For learners: Awareness of emotional regulation and adopting motivated learning strategies can improve success.

Acknowledgements

We thank the SWAYAM project team of NITTTR Chennai for supporting this study.

References

Albelbisi, N. A., Al-Adwan, A. S., & Habibi, A. (2021). Self-regulated learning and satisfaction: A key determinants of MOOC success. *Education and Information Technologies*, *26*(3), 3459–3481. https://doi.org/10.1007/s10639-020-10404-z

Allen, K.-A., Furlong, M. J., Vella-Brodrick, D., & Suldo, S. M. (2022). *Handbook of Positive Psychology in Schools*. Routledge. https://doi.org/10.4324/9781003013778

Bandura, A., & Adams, N. E. (1977). Analysis of self-efficacy theory of behavioral change. *Cognitive Therapy and Research*, *1*(4), 287–310. https://doi.org/10.1007/BF01663995

Bieleke, M., Gogol, K., Goetz, T., Daniels, L., & Pekrun, R. (2021). The AEQ-S: A short version of the Achievement Emotions Questionnaire. *Contemporary Educational Psychology*, *65*, 101940. https://doi.org/10.1016/j.cedpsych.2020.101940

Credé, M., & Phillips, L. A. (2011). A meta-analytic review of the Motivated Strategies for Learning Questionnaire. *Learning and Individual Differences*, *21*(4), 337–346. https://doi.org/10.1016/j.lindif.2011.03.002

Dweck, C. S. (2007). (2007). *Mindset: The New Psychology of Success*. United States: Random House Publishing Group.

Fernández-Esquinas, M., Sánchez-Rodríguez, M. I., Pedraza-Rodríguez, J. A., & Muñoz-Benito, R. (2021). The use of QCA in science, technology and innovation studies: a review of the literature and an empirical application to knowledge transfer. *Scientometrics*, *126*(8), 6349–6382. https://doi.org/10.1007/s11192-021-04012-y

Greenhow, C., & Lewin, C. (2016). Social media and education: reconceptualizing the boundaries of formal and informal learning. *Learning, Media and Technology*, *41*(1), 6–30. https://doi.org/10.1080/17439884.2015.1064954

Guerrero, M., Heaton, S., & Urbano, D. (2021). Building universities' intrapreneurial capabilities in the digital era: The role and impacts of Massive Open Online Courses (MOOCs). *Technovation*, *99*, 102139. https://doi.org/10.1016/j.technovation.2020.102139

Honicke, T., & Broadbent, J. (2016). The influence of academic self-efficacy on academic performance: A systematic review. *Educational Research Review*, *17*, 63–84. https://doi.org/10.1016/j.edurev.2015.11.002

Julia, K., Peter, V. R., & Marco, K. (2021). Educational scalability in MOOCs: Analysing instructional designs to find best practices. *Computers & Education*, *161*, 104054. https://doi.org/10.1016/j.compedu.2020.104054

King, R., & Areepattamannil, S. (2014). What Students Feel in School Influences the Strategies They Use for Learning: Academic Emotions and Cognitive/Meta-Cognitive Strategies. *Journal of Pacific Rim Psychology*, *8*, 18–27. https://doi.org/10.1017/prp.2014.3

Leung, S. O., & Xu, M. L. (2013). Single-Item Measures for Subjective Academic Performance, Self-Esteem, and Socioeconomic Status. *Journal of Social Service Research*, *39*(4), 511–520. https://doi.org/10.1080/01488376.2013.794757

Mayer, R. E. (2005). Cognitive Theory of Multimedia Learning. In *The Cambridge Handbook of Multimedia Learning* (pp. 31–48). Cambridge University Press. https://doi.org/10.1017/CBO9780511816819.004

Pekrun, R. (2006). The Control-Value Theory of Achievement Emotions: Assumptions, Corollaries, and Implications for Educational Research and Practice. *Educational Psychology Review*, *18*(4), 315–341. https://doi.org/10.1007/s10648-006-9029-9

Pekrun, R., & Linnenbrink-Garcia, L. (2012). Academic Emotions and Student Engagement. In *Handbook of Research on Student Engagement* (pp. 259–282). Springer US. https://doi.org/10.1007/978-1-4614-2018-7_12

Pintrich, P. R., Smith, D. A. F., Garcia, T., & Mckeachie, W. J. (1993). Reliability and Predictive Validity of the Motivated Strategies for Learning Questionnaire (Mslq). *Educational and Psychological Measurement*, *53*(3), 801–813. https://doi.org/10.1177/0013164493053003024

Ragin, C. C. (2006). Set Relations in Social Research: Evaluating Their Consistency and Coverage. *Political Analysis*, *14*(3), 291–310. https://doi.org/10.1093/pan/mpj019

Zhoc, K. C. H., Chung, T. S. H., & King, R. B. (2018). Emotional intelligence and self-directed learning: Examining their relation and contribution to better student learning outcomes in higher education. *British Educational Research Journal*, *44*(6), 982–1004. https://doi.org/10.1002/berj.3472

An Online Drill-and-Practice System Anchoring the Cultivation of a Growth Mindset: Design Concepts, Core Components, and Preliminary Evaluation

Fu-Yun YU
National Cheng Kung University, Taiwan
fuyun.ncku@gmail.com

Abstract: Given the far-reaching effects of a growth mindset on cognitive, social-emotional, and professional development, this work aimed to construct the world's first online drill-and-practice learning system designed to cultivate a growth mindset. The central ideas valued by growth mindset advocates (e.g., viewing mistakes as opportunities for cognitive growth, focusing on learning processes over outcomes) were incorporated by designing, developing, and integrating an error-correction subsystem—consisting of six productive error-correction steps—into an existing drill-and-practice platform. This paper primarily explains the design concepts and core components that support the productive error-correction process. An evaluation study involving four classes of sixth-grade students (*n*=104) was conducted to provide preliminary data on the system's effectiveness in fostering a growth mindset.

Keywords: Drill-and-practice activity, growth mindset, learning from errors, online learning system design and development, practice with feedback

1. Introduction

Numerous researchers (e.g., Dweck, 2007; Sousa et al., 2024) emphasized the significance of cultivating a growth mindset for human personal and professional development. Additionally, studies have substantiated the educational value of a growth mindset for promoting academic improvement, mental health (Burnette et al., 2023), and social-emotional development (Jiang et al., 2023). While explicit strategies, interventions, and tips for fostering the development of a growth mindset have been proposed (e.g., embracing mistakes as opportunities for learning and growth, focusing on effort and learning processes over outcomes, encouraging self-reflection) (Burnette et al., 2022), how to integrate these into routine classroom activities remains insufficiently addressed.

Given that practice on pre-determined learning objectives with feedback is an essential instructional component, yet explanatory feedback on students' performance is predominantly provided by the system or the teacher, this work focuses on developing a student-centered error-correction subsystem to augment online drill-and-practice activities. Since engaging students in error correction aligns with the principles of a growth mindset, this subsystem was designed to promote active learner involvement in the process.

2. Design Concepts and Core Components to Guide the Development of a Productive Error-Correction Process

As several researchers have urged, to maximize learning, students and teachers should focus on more than simply correcting errors, ensuring that other important aspects of the error-correction process are not neglected. For instance, Byrne (1988) and Cogie et al. (1999) encouraged students to think about what types of errors are made, besides providing the

correct answer. Arias (2004) suggested that reasons for errors should be highlighted, and the correct use should be explained. Iseni (2011) similarly suggested that students explain their errors and attend to the types and frequency of errors made during error correction. Khansir and Pakdel (2018) pointed out that error correction involves not only the analysis of the source and location of errors but also the provision of explanations for misconception clarification.

Researchers in feedback and problem analysis have also highlighted other crucial areas for consideration. Kulhavy and Stock (1989) proposed four types of task-specific elaboration messages in response to question-answering, among which the 'instruction-based elaboration'—concerning information covered in the curriculum—is most pertinent. Mason and Bruning (2001) distinguished four types of elaborated feedback. Of these, three are worth considering: 'topic-contingent' messages provide definitive, basic information on the study topic; 'bug-related' messages supply error-specific information; and 'response-contingent' messages offer explanations for the correct answer. To help learning through problems, Pelley and Dalley (1997) proposed explicit steps for problem analysis and emphasized the importance of identifying the topic(s) of the problem and understanding why the correct answers are correct and the incorrect answers are incorrect.

Drawing on the above literature, we proposed six core components and subsequently incorporated them into our system to guide students in productive error correction. They are: (1) correcting mistakes, (2) pinpointing the main ideas tested in the question, (3) explaining why the submitted answer is incorrect, (4) justifying why the correct answer is correct, (5) identifying the sources of the errors, and (6) retrieving and interlinking pertinent knowledge pieces to the question. These highlighted error-correction processes embody the core values of a growth mindset and are expected to foster such a mindset in learners.

3. Preliminary evaluation of the Effects of an Online Error-Correction System on the Cultivation of a Growth Mindset

Four sixth-grade classes (n=104) participated in this study over ten weeks. In the first week, students filled out a scale to assess their beliefs on fixed and growth mindsets, before the teacher gave a brief training session on the aim of the newly integrated error-correction activity (i.e., promoting a deeper understanding of the study topic and elaborative knowledge construction) and system use. Starting from the second week, as a routine in each week, students received two 40-minute instructional sessions on social science delivered by the teacher. Afterwards, in another 40-minute instructional session, the teacher gave brief whole-class feedback on students' performance from the previous week, highlighting model student work and clarifying common misconceptions. Then, students completed a 10-item online multiple-choice quiz on the social science topic covered during the current week. Upon completion, they proceeded to the online error-correction task, in which they worked through each of the six components for any questions they had answered incorrectly, following the step-by-step procedures provided in the system. Finally, students completed the same scale again at the end of the 10th week. Jiăng's scale, with established validity and reliability, was adapted to ensure contextual fitness.

The results based on pair-sample t-tests did not substantiate that students engaged in online drill-and-practice with error-correction activities led to a significantly lower score on a fixed mindset, t=0.17, $p > 0.05$, nor a significantly higher score on a growth mindset, t=-1.49, $p > 0.05$. Several factors may account for the unexpectedly unconfirmed findings. First, it appears that the training session centered more on the technical (i.e., completing the task in the system) rather than the conceptual (i.e., the meaningfulness and usefulness of each error-correction component) and motivational aspects (i.e., positive climate management) of the integrated activity. Second, an instructional implementation consisting of 15 minutes of student-centered error correction for nine weeks (i.e., 135 minutes) may not be sufficient to produce substantial results. As mindset researchers have suggested, several contextual factors (e.g., differences in implementation procedures) influence which growth mindset interventions are effective, for whom, and under what conditions (Burnette et al., 2023). Third, having students engaged in online drill-and-practice with error-correction activities on social

science (i.e., one specific domain) may not be impactful enough to affect students' general beliefs in their intelligence and aptitudes, given that people could possess different mindsets and beliefs about themselves in different areas (Dweck & Master, 2009).

4. Conclusion

An online system to extend a drill-and-practice system for the cultivation of a growth mindset is the aim of this work. Given the unconfirmed findings and the implementation procedures used in this study, future research should include a training session that explicitly conveys the value of each error-correction component—potentially through persuasive messages that tap into students' valuation system (Falk, 2025)—and extend the intervention period to support online drill-and-practice activities across multiple subject areas. Incorporating qualitative methods, such as in-depth unstructured interviews exploring students' perceptions of each component's role in fostering a growth mindset, would also help validate and refine the proposed error-correction framework.

Acknowledgements

This work was supported by the Ministry of Science and Technology, Taiwan, under Grant No. 111-2410-H-006-023-MY3.

References

Arias, I. J. (2004). Treating students' error in oral production. *LETARS, 36*, 25-35.

Burnette, J. L., Billingsley, J., Banks, G. C., Knouse, L. E., Hoyt, C. L., Pollack, J. M., & Simon, S. (2023). A systematic review and meta-analysis of growth mindset interventions: For whom, how, and why might such interventions work? *Psychological Bulletin, 149*(3-4), 174-205. doi: 10.1037/bul0000368.

Burnette, J. L., Knouse, L. E., Billingsley, J., Earl, S., Pollack, J. M., & Hoyt, C. L. (2022). A systematic review of growth mindset intervention implementation strategies. *Social and Personality Psychology Compass, 17*, e12723. https://doi.org/10.1111/spc3.12723

Byrne, D. (1988). *Teaching writing skills*. Longman Group UK Ltd.

Cogie, J., Strain, K., & Lorinskas, S. (1999). Avoiding the proofreading trap: The value of error correction process. *The Writing Center Journal, 19*(2), 7-32. https://www.jstor.org/stable/43442834

Dweck, C. S. (2007). *Mindset: The new psychology of success*. Ballantine Books.

Dweck, C., & Master, A. (2009). Self-theories and motivation: Student's beliefs about intelligence. In K. R. Wentzel & A. Wigfield (Authors), *Handbook of motivation at school* (pp. 123–140). Routledge

Falk, E. (2024). What we value: The neuroscience of choice and change. W. W. Norton & Company.

Iseni, A. (2011). Assessment, testing and correcting students' errors and mistakes. *Language Testing in Asia, 1*(3), 60-90.

Jiang, X., Mueller, C. E., & Paley, N. (2023). A systematic review of growth mindset interventions targeting youth social–emotional outcomes. *School Psychology Review, 53*(3), 251–272. https://doi.org/10.1080/2372966X.2022.2151321

Khansir, A., & Pakdel, F. (2018). Place of error correction in English language teaching. *Educational Process: International Journal, 7*(3), 189-199.

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review, 1*(4), 279-308.

Mason, B., & Bruning, R. (2001). *Providing feedback in computer-based instruction: What the research tells us.* Pelley, J. W. & Dalley, B. K. (1997). *Success types for medical students: A program for improving academic performance*. Consulting Psychologist Press.

Pelley, J. W., & Dalley, B. K. (1997). Success types for medical students: A program for improving academic performance. Consulting Psychologist Press.

Sousa, B. J., & Clark, A. M. (2024). Growth mindsets in academics and academia: A review of influence and interventions. *Journal of Higher Education Policy and Management*, 1–19. https://doi.org/10.1080/1360080X.2024.2384003

Attaining the Program Educational Objectives of the IT Program Three to Five Years After Graduation: An Evidence-Based Approach

Arlene Mae Celestial-VALDERAMA[a*], Roel C. TRABALLO[ab], Sharmaine Justyne R. MAGLAPUZ & Don Erick J. BONUS[a]
[a]*Jose Rizal University, Philippines*
[b]*University of Makati, Philippines*
*arlene.valderama@jru.edu

Abstract: Faculty and administrators responsible for setting academic standards, such as institutional outcome (IO), program educational objectives (PEO), student outcomes (SO) and course learning outcomes (CLO) are internal key stakeholders in universities, especially when aligning with national and international competencies. This paper presents an evidence-based attainment of these objectives. The focal point of this study is on the Bachelor of Science in Information Technology (BSIT) program of the university. It addresses the research question of how program educational objectives are attained for the BSIT degree towards results on each identified PEO where the target point is to attain them 3 to 5 years after their graduation at the university. Quantitative descriptive analysis was employed, and 52 alumni from their graduation year between 2013-2018 were identified respondents of the study under the aforementioned degree program. A PEO-inclined tracer survey with first indicators states that 71% of the respondents' current jobs are in line with their college degree (PEO 1). 48% have rated that they have impacted "very much" contribution in the IT field given their current line of work (PEO 2). Lastly, 42% rated that they have "very much" involvement level in research and development projects in the IT field and allied disciplines (PEO 3). The results have stated positive and the majority of indicators of the attainment of the three PEOs identified by the IT department under the College of Computer Studies and Engineering.

Keywords: program educational objectives, BSIT program, evidence-based objectives attainment, IT graduates

1. Introduction

School administrators, stakeholders, and the teaching sector have been constantly selecting identifiable visions for respective higher education institutions (HEIs). These items aligning to the vision and mission include institutional outcomes, degree program educational objectives, student outcomes, and course learning outcomes. In order to produce graduates of social importance, they wished these outcomes would be attained on the appropriate timeline, including those based on national and international competencies, as they are crucial to the internal stakeholders in universities (Shenderova, 2022). Monitoring and attainment of defined objectives and outcomes is tedious but must adhere to its conformities and attainment.

1.1 Program Educational Objectives Quality

The university has maintained certification for ISO 9001:2015 for the past 9 years, it has conducted internal quality and external audits to the current version of the globally recognized standard for Quality Management Systems (QMS). The last recent school year 2024-2025, it

has applied and passed the ISO 21001:2018 Educational Organization Management System (EOMS) emphasizing importance for maintaining the institution's certification and enhancing the quality management system within the educational organization. In line with the attainment of the program education objectives, the review of these goals is maintained in accordance with the HEI's ISO document coded as ACM 4.2 Curriculum Development, Review and Validation for Higher Education, where ACM is for the Academic Affairs Manual.

In the context of Jose Rizal University, an Institutional Outcome is defined along with Vision and Mission. Following this is the identification of Program Education Objectives (PEO) for each program offering. Tshai et. al. (2014) stated that broad statements outlining the professional and career achievements that the program is preparing the graduates to attain should be included in PEOs. The next salience on this path is how can the university design curriculum and programs that effectively prepare graduates for their desired careers? More so, how can these be measured? For Aldowaisan & Allahverdi (2015), it is important that performance metrics and target attainment levels are established as part of an evaluation procedure used for every PEO. As the university identifies at least three Program Educational Objectives for the degrees, these objectives must match an attainment through a well-designed curriculum. This paper presents an evidence-based attainment of these objectives. The focus is on the Bachelor of Science in Information Technology program of the university. There are 3 identified PEOs and 13 Student Outcomes (SO), cross-matched with each course in the curriculum with at least 3 Course Learning Outcomes (CLO). As this paper focuses on the 3 PEOs of the BSIT program, it focuses mainly on the data of BSIT graduates. The process of monitoring alumni's locus and acquiring information about graduates' job performance from employers is time-consuming and labor-intensive. Universities conduct tracer studies to monitor graduates and evaluate the effectiveness of their education in preparing them for their careers. For accreditation purposes, the findings are also used to evaluate the caliber of programs in higher education (Dewi, et. al., 2021). To streamline this process, the university Guidance and Testing Office, Alumni Office, and the College of Computer Studies and Engineering have conducted a graduate tracer initiative via a survey form floated online and was distributed to alumni from Batches 2013-2018 via the creation of an online graduate tracer approach, taking advantage of internet and mobile technologies.

The ability of college graduates to enter the job market serves as a critical indicator of learning outcomes and the alignment of higher education with societal needs. This paper focuses on the research question of how Program Educational Objectives are attained for the BSIT degree towards results on each identified PEO. The Program Educational Objectives (PEO) of the BSIT program in Jose Rizal University includes:

Three to five years after completing the BS Information Technology, the graduates will:

1) *"Excel in their profession/career utilizing the knowledge acquired in the Information Technology courses."*
2) *"Become effective collaborators and innovators in information technology, applying professional/technical skills and competencies to make a positive impact on society;" and*
3) *"Be engaged in life-long learning and professional development through self-study, continuing education, or graduate and professional studies."*

Higher education institutions have a shared responsibility, two of which are: to impart necessary competencies to their graduates and to facilitate their transition into employment. In order to attain these objectives three to five years after graduation, the university has been implementing the achievement of thirteen Student Outcomes (SO) throughout the four-year stay of the BSIT student in the institution. A way to achieve this is the inclination of at least three course learning outcomes of every course the student takes each semester.

2. Methods and Analysis

2.1 Profile

An alumni tracer documentation data was floated through Survey monkey, retrieved August 2024. It is composed of 32 questions that were distributed to alumni of the BSIT program. The

tracer questions were subjected to a validation process for face and content validity. It was defined by McBurney (1995) as the following:

- Face validity is the idea that a test should appear superficially to test what it is supposed to test; and
- Content validity is the notion that a test should sample the range of behavior represented by the theoretical concept being tested.

In the validation process, content validity was used. The 32-item questionnaire followed in this study was formulated by six (6) experts of the alumni tracer project study of the HEI. Mentioned experts were composed of the alumni director, college coordinator for alumni, dean of the college, and 3 department chairs of the college. Five (5) of the experts excluding the alumni director have formulated and tapped the questions as essential with the intention of measuring the relevance of the objectives of the study. Following the formula in calculating the content validity ratio (CVR) (Taylor, 2017), the questionnaire's worth is based on the ratings of the experts. The equation is for Content Validity Ratio using the formula: CVR = [E - (N / 2)) / (N / 2)], where: E is number of experts who rated the questionnaire as essential, N is total number of experts. The CVR is 0.67 which is an acceptable result. Respondents age range is 20-35, (n = 52; 14 females and 38 males). Their graduation year starts from 2013-2018. A distribution of the graduation year is exhibited in Table 1.

Table 1. *Year of Graduation Distribution*

Year of Graduation	Number	Distribution	3 Years After Graduation	4 Years After Graduation	5 Years After Graduation
2013	4	8%	2016	2017	2018
2014	13	25%	2017	2018	2019
2015	10	19%	2018	2019	2020
2016	7	13%	2019	2020	2021
2017	10	19%	2020	2021	2022
2018	8	15%	2021	2022	2023
Total	**52**	**100%**			

As Table 1 presents the alumni data from the aforementioned years of graduation, it is salient to note that PEO 1 has measurable results of BSIT degree holders' data must be 3 to 5 years after their graduation. Venturing on the 3-year mark, the graduates allowable present information ranges from 2017-2022. For the 4-year mark, information ranges from 2017-2022, and lastly for the 5-year mark, it ranges from 2018-2023.

2.2 Employability

Pheko & Molefhe (2016) proposed a structured approach for various players in the labor market to ensure the advancement of attributes as well as skills in employability skills. In their study, fostering employability is a collaborative effort involving training providers, employers, existing employees, and potential employees. A study gap that may be addressed in this paper is the experience of the higher education institution (HEI) in a decline of the relevance of the BSIT graduate competencies towards their past and present employment. The study of Segbenya, et. al. (2023) acknowledged four key areas of skills ensuring employability, it includes intellectual or scholarly skills, personality dispositions, meta-skills and those that focus on the job itself. Research on employability now encompasses all forms of employment, not only unemployment. Government agencies, higher education institutions (HEIs), and businesses must all work together to promote skills associated with employability of graduates (Tasker & Packham, 1994). As JRU identifies the attainment of the course learning outcomes, the diversity of skills may be attained through them as it crossmatches alongside the student outcomes. In today's competitive global job market, universities must prioritize their graduates'

employability. This is key for graduates to secure satisfying careers and for employers to stay ahead of the competition (Su & Chang, 2015).

Out of 52 respondents, only 1 replied of non-employment, all other 51 alumnus are currently employed on the date of survey data extraction. Jose Rizal University has been monitoring employment among its graduates; the alumni respondents have been employed after graduation in their jobs within the following time range. In less than 6 months after graduation, 83% or 43 out of 52 of the alumni have been employed. Within 6 months to 1 year, 15% or 8 out of 52 were employed. Lastly, between 1 to 2 years, 2% or only 1 of 52 have been employed. Among these, their employment varies with local and foreign employment. Data shows that 8% are from foreign employment, 90% are locally employed, and 2% are self-employed. Another factor worth noting is the number of employments the alum has occupied since obtaining their college degree. There are 38% who mentioned that they only had 1 job since college, 31% had 2, 25% had 3, and only 6% said that they already had more than 3. To conclude, this paper also states how long the alumni have stayed in their present employer. 37% of them have been on their present employment in less than 6 months. 29% have stayed for 1 to 2 years already. 13% have been working for 6 months to 1 year. 10% have been employed within 2 to 3 years, and lastly also 10% have been working for more than 3 years in their companies.

3. Results and Discussion

3.1 Program Educational Objectives

The nature of the company the alumni are engaged in at present includes 42% in services and consultancy, 4% in commercial and retail, 2% in education, and 52% have mentioned others but clearly all have stated their technical positions in their jobs. These job titles include programmers, database administrators, information security specialists, quality assurance specialists, systems analysts, technical support specialists, and web and application developers.

The BSIT graduates were asked to assess the applicability of their JRU curriculum to their current professional endeavors. They rated the relevance of their obtained JRU curriculum in college to their current line of work. Table 2 shows a summary.

Table 2. *Rating of Curriculum Relevance to Current Employment*

Rate (10-highest)	# of Responses	Distribution
10	5	10%
9	4	8%
8	13	25%
7	8	15%
6	6	12%
5	9	17%
4	1	2%
3	3	6%
2	2	4%
1	1	2%
Total	**52**	**100%**

The most common rating was 8 out of 10, with 25% of respondents choosing this option. It suggests that the respondents found their curriculum relevance to their current employment has its rating to be very good, but not perfect. In the context of a survey or assessment, it might mean that the respondents believe that their BSIT program is highly effective and useful.

In terms of satisfaction with the quality of education they have received, 46 of them have replied positively, which comprises a high 88% and only 6 or 12% have stated otherwise.

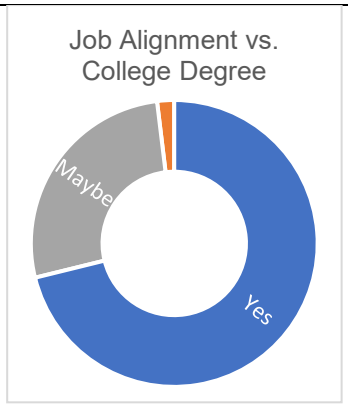
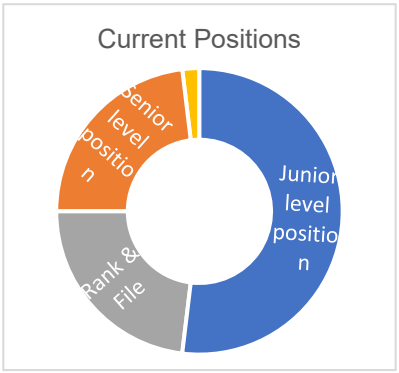
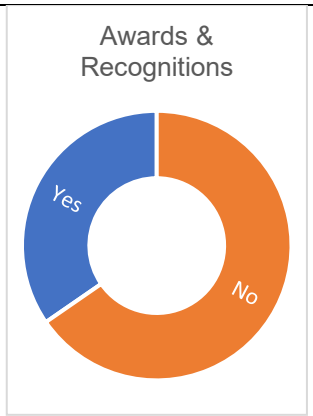
3.1.1 Program Educational Objective 1

The university has defined PEO #1 as “Excel in their profession/career utilizing the knowledge acquired in the Information Technology courses.”

Historically, a “career” was often seen as a linear progression, with individuals moving from one job to another in a way that led to higher positions and status within an organization (Ojala et. al., 2021). In accordance with the BSIT first program educational objective, graduates were given the opportunity to assess the congruence of their jobs’ alignment with their degree. Seventy-one percent of respondents affirmed the statement, while 2% negated it. Twenty-seven percent expressed uncertainty.

For their present positions, more than half of them are in junior level positions or 52%. There are 23% who are in senior level positions and rank and file, respectively. And 2% are in a top management position. A considerable number of graduates have been honored and acknowledged for their contributions to the field of information technology and allied disciplines. Thirty-five percent of the graduates have attained such recognition. To further illustrate, Table 3 presents a summary of the charts.

Table 3. *PEO 1 Indicators*

PEO 1 Indicator 1	PEO 1 Indicator 2	PEO 1 Indicator 3
		

3.1.2 Program Educational Objective 2

The university has defined PEO #2 as “Become effective collaborators and innovators in information technology, applying professional/technical skills and competencies to make a positive impact on society.”

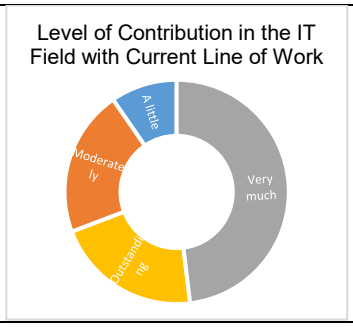
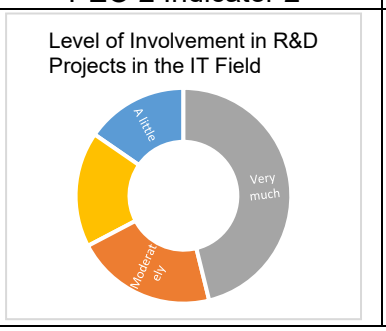
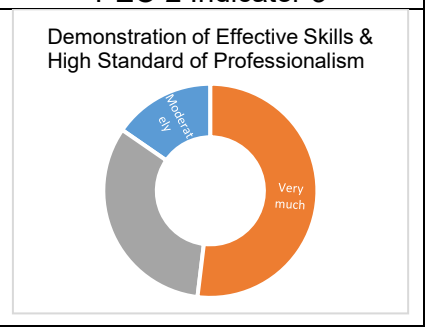
The graduates were also given the opportunity to rate the level of their contributions in the field of Information Technology. This clearly states that they are assessing the impact and significance of their work in that area as well. Essentially, this is how the BSIT degree holders are evaluating how valuable their contributions have been to the advancement of Information Technology. The highest remark of “Outstanding” was responded to by 21% of the graduates. About 48% of the graduates have expressed that they have contributed “very much”, 21% have mentioned that they contributed “moderately”, and only 10% have responded that they have “little” contributions.

Involvement in research and development (R&D) projects in Information Technology means active participation in projects that aim to create new or improved IT products, services, or processes. These respondents were also able to contribute to these. About 17% of the respondents have been outstandingly involved. A high proportion of respondents with 46% have been inclined to research and development “very much” as the second top box score. 15% have been “moderately” inclined and only 15% have been involved “a little.”

For this second PEO to take effect, the graduates were also asked if they have ever highlighted their ability to use technology effectively and maintain a prominent level of

professionalism in their work. It is good to note that more than half of them have responded to the second top box score, “very much” which is a massive 52%. A high of 33% also have positively replied to the perfect top box score of “outstanding.” And only 15% have replied “moderate” showcase abilities. No one has responded only “a little” ability in the utilization of technology effectiveness. To further illustrate, table 4 presents a summary of the charts.

Table 4. *PEO 2 Indicators*

PEO 2 Indicator 1	PEO 2 Indicator 2	PEO 2 Indicator 3
		

3.1.3 Program Educational Objective 3

The university has defined PEO #3 as “Be engaged in life-long learning and professional development through self-study, continuing education, or graduate and professional studies.”

The respondents have received several training opportunities and advancement in their professional development. They have obtained these in their current line of work. To demonstrate evidence of this, they have rated the level of training received that enabled them to achieve continuous education and professional advancement. Second top box score of “very much” has gained the highest responses of 42%, top box score of “outstanding” has gained 27%. The “moderate” level of training has 29% and only 2% gained “a little” level of training in professional education advancement. Figure 1 presents a summary of this indicator.

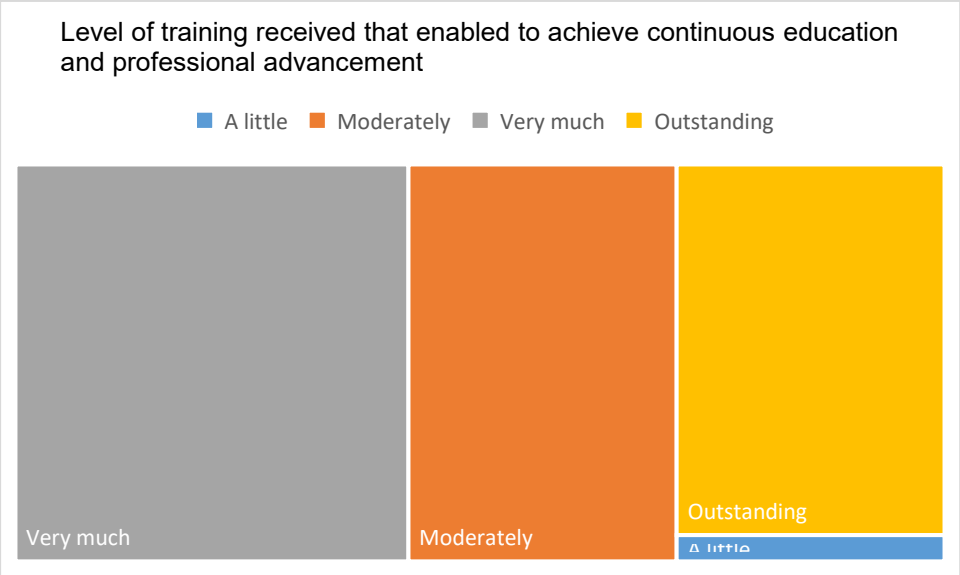


Figure 1. *PEO 3 Indicator 1*

In reference to their professional education, there are 5 alumni who have pursued professional education. This data includes 2 technical training courses and 3 graduate studies. These comprise of DevOps/B2B Expertise/AWS SysOps and Cisco training for technical inclination. The graduate programs pursued included MBA, MIT, and LLB.

Certifications relevant to the Information Technology field and allied disciplines are formal qualifications that validate an individual's knowledge and skills in specific areas of IT.

Recognized professional organizations or technology vendors typically award these certifications. 12 respondents receive these certifications:

1) Data Privacy
2) CRM Training Certification and Security Network Certification
3) Certification for IBM B2B Integration Technical Mastery, Certification for IBM Managed File Transfer Mastery
4) Technical certification for Lenovo, HP, Asus, and Epson Services
5) Drupal Apprentice Program
6) Google Analytics, HubSpot Certified
7) Robotics
8) Cisco training completions
9) Salesforce Certifications: Salesforce Certified Administrator & Salesforce Certified Platform App Builder
10) Software Developer – JAVA
11) Programming- TESDA
12) Training certificate for Java, Cobol, HTML, CSS3, BOOTSTRAP, JavaScript, and Network Management

4. Conclusion

Bhele (2020) provides insights into how specific academic initiatives can boost student and graduate employability, offering measurable metrics for assessing their effectiveness. Their framework addresses the challenges higher education institutions face in influencing employability outcomes. By examining their case studies, it demonstrated how variations in the labor market can significantly impact employability rates. Ultimately, a comprehensive understanding of employability requires a multifaceted approach that considers a range of indicators, including labor market factors, to present a more nuanced picture. Pham et. al. (2024) They suggested that measuring graduate employability immediately after graduation is insufficient, as different forms of capital impact career development at various stages.

An online tracer survey for the Jose Rizal University for the Bachelor of Science in Information Technology program is in place. As of present time, the Alumni Office of the university has central control and management of these data, and every college of the Academic Division has their share of providing data. These initiatives are in place to trace data on graduates and to assess the educational program objectives (PEO) of the degree programs offered by the university.

The results of the responses in this documentation have stated very good indicators of the attainment of the three PEOs identified by the IT department under the College of Computer Studies and Engineering. The usefulness and usability of the college education system are well-indicated by the tracer study. To ensure its accessibility, the system should be designed to serve both current and former graduates. In this predicament, the academic council can immediately obtain graduate statistics and user feedback.

References

Shenderova, S. (2022). Collaborative degree programmes in internationalisation policies: the salience of internal university stakeholders. European Journal of Higher Education, 13(2), 197–215. https://doi.org/10.1080/21568235.2022.2120035

Tshai, K. Y., Ho, J. H., Yap, E. H., & Ng, H. K. (2014). Outcome-based Education – The Assessment of Programme Educational Objectives for an Engineering Undergraduate Degree. Engineering Education, 9(1), 74–85. https://doi.org/10.11120/ened.2014.00020

Aldowaisan, T., & Allahverdi, A. (2015). Improving educational objectives of the Industrial and Management Systems Engineering programme at Kuwait University. European Journal of Engineering Education, 41(3), 252–262. https://doi.org/10.1080/03043797.2015.1056100

Dewi, N. R., et al. (2021). J. Phys.: Conf. Ser. 1918 042010

McBurney, Donald H. (1995). The Problem Method of Teaching Research Methods. First Published February 1, 1995. Accessed from https://repository.up.ac.za/bitstream/handle/2263/28684/03chapter3.pdf?sequence=4

Taylor, C. (2017). How to Calculate Content Validity Ratios. Updated September 26, 2017, https://bizfluent.com/how-7448484-calculate-content-validity-ratios.html

Pheko, M. M., & Molefhe, K. (2016). Addressing employability challenges: a framework for improving the employability of graduates in Botswana. International Journal of Adolescence and Youth, 22(4), 455–469. https://doi.org/10.1080/02673843.2016.1234401

Segbenya, M., Oppong, N. Y., Nyarko, E. A., & Baafi-Frimpong, S. A. (2023). Demographic characteristics and employability skills among tertiary graduates in Ghana: Evidence from the National Service Scheme. Cogent Economics & Finance, 11(2), Accessed from https://doi.org/10.1080/23322039.2023.2225915

Tasker, M., & Packham, D. (1994). Government, higher education, and the industrial ethic. Higher Education Quarterly, 48, 182–193.10.1111/hequ.1994.48.issue-3

Su, W., & Zhang, M. (2015). An integrative model for measuring graduates' employability skills—A study in China. Cogent Business & Management, 2(1). https://doi.org/10.1080/23311975.2015.1060729

Ojala, K., Isopahkala-Bouret, U., & Varhelahti, M. (2021). Adult graduates' employability and mid-career trajectories after graduation with Finnish UAS Master's degree. Journal of Education and Work, 34(1), 67–80. https://doi.org/10.1080/13639080.2021.1875125

Behle, H. (2020). Students' and graduates' employability. A framework to classify and measure employability gain. Policy Reviews in Higher Education, 4(1), 105–130. https://doi.org/10.1080/23322969.2020.1712662

Pham, T., Soltani, B., & Singh, J. K. N. (2024). Employability capitals as essential resources for employment obtainment and career sustainability of international graduates. Journal of Further and Higher Education, 48(4), 436–448. https://doi.org/10.1080/0309877X.2024.2344771

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Effect of ISIED Education Activities on The Development of Students' Key Competency

Yuhan ZHOU[a], Jiaru WU[b] & Guang CHEN[a*]
[a] *Faculty of Education, Beijing Normal University,* Beijing, China
[b] *The Chinese University Of Hong Kong, Shenzhen Caitian School* Shenzhen, China
*guang@bnu.edu.cn

Abstract: Since the 21st century, countries and regions have regarded collaborative learning, critical thinking, creative thinking, and problem-solving skills as key competencies for students. Teaching methods with integrated multidisciplinary approaches, akin to STEM education, are emphasized to create real-world problem scenarios for students and integrate engineering practices with science teaching into traditional science curricula. However, teachers in science courses often overlook engineering practices, and there is currently a lack of an operational educational model to support the integration of engineering practices into the classroom. This study improves and optimizes the STEM teaching model, integrating scientific inquiry and engineering design. Taking the science teaching unit "Heat" as an example, it designs a course to explore the impact of the ISIED educational activities on students' key competencies. The results indicate that the ISIED educational activities promotes students' understanding of scientific concepts and the application of knowledge in real-world problem-solving, thereby advancing the development of 21st-century key competencies.

Keywords: Key Competency, Collaborative Learning, Critical Thinking, Creative Thinking, Problem Solving Ability, Science Education, ISIED

1.Introduction

Cultivating students' key competency is central to 21st-century global science education, with integrating engineering practices a reform focus. The U.S. integrates scientific and engineering practices under STEM, but localized implementation often isolates them due to educators' limited engineering design experience, hindering problem-solving development. Yet engineering design is irreplaceable for scientific learning and key competencies (Wendell & Rogers, 2013), and comprehensive integration models are lacking.

The U.S. promotes integrating K–12 engineering-science instruction via STEM. NGSS (2013) highlight engineering design's interdisciplinarity and role in problem-solving, spurring related studies. Kolodner et al. (2003), Wendell and Rogers (2013), and Siew et al. (2016) showed engineering design enhances reasoning, collaboration, science learning, creativity, and problem-solving. But Crismond and Adams (2012) note integration challenges for non-engineering teachers, requiring feasible models. English et al. (2017) propose a five-phase sequence: problem scoping, ideation, design & prototyping, evaluation, and redesign/reconstruction, where interdisciplinary STEM knowledge is intentionally included.21st-century key competency frameworks (Bruxvoort & Jadrich, 2016) stress their integration.

This study optimizes the model by integrating scientific inquiry and engineering practice, creating the "Integrated Scientific Inquiry-Engineering Design (ISIED) STEM Instructional Model" (Figure 1).

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

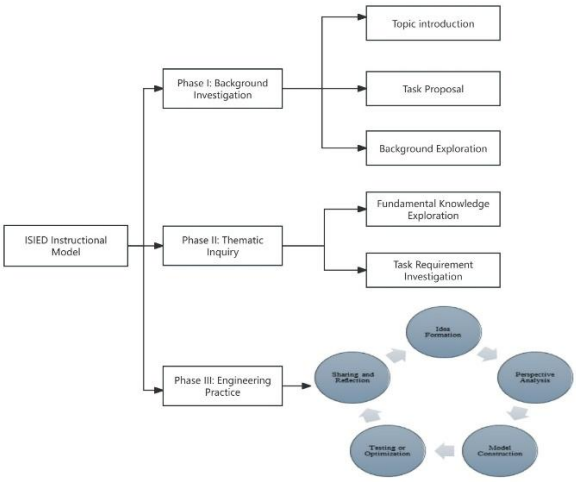


Figure 1. ISIED Instructional Model

The ISIED model, an optimized STEM update centered on engineering problem-solving, retains the original three-phase structure: Phase I (Background Investigation) emphasizes authenticity in three progressive components to lay the groundwork for inquiry and design; Phase II (Thematic Inquiry) links scientific inquiry and engineering practice via theme-centered investigations to foster teamwork and critical thinking; Phase III (Engineering Practice) is an iterative cycle for applying knowledge and collaboration capabilities, with engineering design spanning all phases. Unlike traditional science classes, ISIED adopts an investigation-inquiry-design approach aligned with key competency cultivation, assessing progress in applying knowledge rather than just knowledge acquisition or product quality.

Though optimized from the "STEM model with engineering problem-solving as the main thread," ISIED's application in science classrooms and its ability to promote students' 21st-century key competencies and scientific knowledge need verification, forming the study's main research questions:

RQ1: Does the ISIED instructional model enhance students' scientific knowledge acquisition?

H1: Students participating in ISIED educational activities will demonstrate significantly more significant improvement in scientific knowledge compared to the control group.

RQ2: Does the ISIED instructional model improve students' collaborative learning, critical thinking, problem-solving abilities, and creative thinking?

H2: Students participating in ISIED educational activities will exhibit significantly enhanced collaborative learning skills compared to the control group.

H3: Students participating in ISIED educational activities will show significantly greater development in critical thinking compared to the control group.

H4: Students participating in ISIED educational activities will have significantly superior problem-solving capabilities compared to the control group.

H5: Students participating in ISIED educational activities will demonstrate markedly higher levels of creative thinking than those in the control group.

2. Method

This study used a quasi-experimental pre-test/post-test design, with 20 fifth-graders in Classes 1-3 (experimental group, ISIED activities) and 20 in Class 4 (control group, traditional science courses) matched by prior science scores. It ran two weeks (60-minute sessions) through pre-test, intervention, and post-test stages. Tools included a science knowledge questionnaire, 21st-Century Key Competencies Scale (Cronbach's α=0.95), and post-test semi-structured interviews.

3. Results

Paired t-tests on heat transfer knowledge showed both groups improved, with the experimental group (t=12.499, p=0.000, d=2.281) gaining more than the control group (t=3.627, p=0.002, d=0.978). ANCOVA (covarying pre-test) confirmed the experimental group's post-test scores remained higher, supporting Hypothesis 1 that ISIED better enhances scientific knowledge.

For key competencies, paired t-tests found the experimental group improved significantly in total scores (t=2.454, p=0.024, d=0.530), collaborative ability (t=2.680, p=0.015, d=0.557), and critical thinking (t=2.320, p=0.032, d=0.618). ANCOVA (Table 4) showed their post-test total scores (F=6.69, p=0.014) and collaborative/critical thinking scores (F=2.952, p=0.094) outperformed the control group, supporting Hypotheses 2 and 3. However, no significant improvements were seen in creative thinking (t=1.774, p=0.092, d=0.391) and problem-solving ability, though interviews indicated ISIED improved problem-solving, supporting Hypotheses 4 and 5.

Student interviews revealed positive feedback, with the unit seen as "fun" and "engaging" compared to traditional lessons. Students noted more active knowledge construction, stronger collective responsibility, and highlighted the focus on real-world engineering practice, recognizing improvements in collaboration, critical thinking, and creativity.

4. Discussion

This study explored ISIED activities' impact on students' scientific knowledge and key competence, drawing main conclusions: ISIED effectively enhanced scientific knowledge (consistent with Wendell & Rogers, 2013; Cunningham et al., 2020) via inquiry-engineering approaches and real-world context, and promoted collaborative learning and critical thinking (supporting Kolodner et al., 2003). However, it had no significant impact on creative thinking and problem-solving (contrary to Siew et al., 2016) due to short duration, weak creativity perception, and inertia, with recommendations including documenting creativity and teaching strategies. Unexpectedly, traditional courses failed to improve key competence and lowered scores, lacking real-world problems and independent decision-making.

This study has limitations affecting applicability: small sample size (20 per group) restricts generalizability, with group assignment via prior scores possibly misreflecting current skills; a two-week intervention too short for competency building; self-reported key-competency data may be biased.More focused studies are needed.

References

Adams, R. (2012). The informed design teaching and learning matrix. Journal of engineering education.

Bethke Wendell, K., & Rogers, C. (2013). Engineering design‐based science, science content performance, and science attitudes in elementary school. Journal of Engineering Education, 102(4), 513-540.https://doi.org/10.1002/jee.20026

Bruxvoort, C., & Jadrich, J. (2016). DON'T" SHORT CIRCUIT" STEM INSTRUCTION: Exploring the goals of engineering and science. The Science Teacher, 83(1), 23-28.https://doi.org/10.2505/4/tst16_083_01_23

Cunningham, C. M., Lachapelle, C. P., Brennan, R. T., Kelly, G. J., Tunis, C. S. A., & Gentry, C. A. (2020). The impact of engineering curriculum design principles on elementary students' engineering and science learning. Journal of Research in Science Teaching, 57(3), 423-453.https://doi.org/10.1002/tea.21601

Kolodner, J. L., Gray, J., & Fasse, B. B. (2003). Promoting transfer through case-based reasoning: Rituals and practices in learning by design classrooms. Cognitive Science Quarterly, 3(2), 183-232.

NGSS Lead States. (2013). Next generation science standards: For states, by states. National Academies Press.

Siew, N. M., Goh, H., & Sulaiman, F. (2016). Integrating STEM in an Engineering Design Process: The Learning Experience of Rural Secondary School Students in an Outreach Challenge Program. Journal of Baltic Science Education, 15(4), 477-493.https://doi.org/10.33225/jbse/16.15.477

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Teachers as Learners: Exploring the Role of Course Structure on Learners' Agency

Sumitra SADHUKHAN[a*], Leena BHATTACHARYA[a] & Sridhar IYER[a]
[a] *Indian Institute of Technology Bombay, India*
*sumitra.sadhukhan@iitb.ac.in

Abstract: Faculty members at higher education institutions are often recruited based on their subject-matter expertise, whereas less attention is paid to building other skills that are required to tackle teaching-learning challenges that arise as a part of the profession. In this paper, we draw insights from a student-led course at a technical institute in India aimed at enhancing teaching beliefs by prompting deep pedagogical thinking of future faculty members through 14 post-course surveys and four semi-structured interviews. The findings highlight specific course structures that enabled participants to exercise agency: deep pedagogically engaged course topics, autonomy in connecting personal experiences to course topics, informal peer discussions for co-constructing ideas, and post-class journaling for critical reflection.

Keywords: participatory design, learner-centred knowledge, teaching

1. Introduction

Instructors in higher-education institutions are often selected into the profession with strong research skills, but they often lack proper training to shape their attitude and address other challenges that occur in the classroom (Jain, 2007). As a result, most teachers struggle to understand and adapt to the role as a facilitator, and facilitate student-centred instructional design and student engagement strategies (O'Neill & McMahon, 2005). It affects motivation and the overall experience of students. Therefore, teacher training programs are needed to promote reflective teaching that is necessary to adapt to rapidly changing educational contexts. Teacher training programs are crucial not only for enhancing pedagogical competence but also for fostering the right professional attitudes that align with learner-centred, outcome-based education. Zeng (2020) highlights that when teachers engage in experience-based peer discussion, they gain insights into different pedagogical strategies and internalise their beliefs through comparing and reflecting on their own experiences. This dialogic process facilitates conceptual change and motivates teachers to refine teacher beliefs (Yang & Xiaochen, 2022). If the structure of a course is overly rigid, content-heavy, or teacher-directed, it limits opportunities for student-teachers to reflect, question, and contribute meaningfully, which are key dimensions of exercising agency (Priestley et al., 2015). In this paper, our research goal is to understand whether a student-centered course structure that was floated to build teaching skills in India helped learners to exercise their agency in learning and to refine their teaching beliefs.

2. Data and Methodology

The study is conducted at the Indian Institute of Technology, Bombay, a premier technical institute in India that caters to a wide range of students from various socio-economic backgrounds who are selected through strict selection criteria. The course was designed by the instructor together with the learners. Students proposed topics and reacted to peers' ideas, and the most-voted topic was selected prior to the class. The topics ranged from identifying teacher practices from case studies, mental health challenges, to the conceptualization of challenges and benefits of generative AI in teaching-learning practices.

The course was rolled out for 8 weeks between January to March 2025. Past and present PhD students and postdocs associated with the Centre for Educational Technology were invited to attend the pilot course voluntarily. The learners were also asked to fill in their post-class reflections in a web-based journal, select their peer groups and engage in discussions.

We use data from a post-course questionnaire and interviews for this paper. The learners reflected on their experience with the course. They expressed their experience by ranking their experience on a five-point Likert scale. Out of the enrolled 31, 14 filled out the questionnaire. We acknowledge the fact that more motivated learners are more likely to voluntarily fill out a questionnaire, creating possible bias.

Based on the responses, researcher 1 conducted a semi-structured interview with 4 learners. The data collected as part of the survey are represented in terms of reflecting the participants' profiles and experiences as learners in this course. Researcher 1 read and coded the responses according to their alignment with one of these four structural components following Braun and Clarke (2006). Following this, researcher 2 cross-verified the analysis. This study employed a qualitative multi-phase analysis, combining both content analysis and thematic analysis to explore how course structure enabled students to exercise agency in peer group discussions and refine their teaching beliefs.

3. Result

In this section, we combine the findings from the post-course survey and interviews. The topics covered in the course were defined by the learners together with the instructor. In the sample, 14% of the learners had no teaching experience, whereas the rest had different levels of experience, with a maximum of 10 years. The topics revolved around common challenges faced by higher-education instructors, or the topics presented the learners with scenarios that would help them reflect on their (teacher) beliefs.

The classes followed a structure that was based on discussion and exploration in peer groups. Almost 57% of the learners had never or rarely attended a student-led course before, making it a new experience for them. Interviews revealed that co-designing the course made the learners feel empowered. Although the learners had the choice to propose topics, this structure was not consistently observed across the course, which could have been the result of the student composition and their inexperience in attending student-led lectures. In the absence of learner-led topics, the instructor proposed a topic which were discussed in the class. The instructor gave agency to the learners to choose their peer groups for discussion, but only 43% of the sample reported that they chose their peer groups; yet, 93% reported being heard by their peers, and 57% reported that their experience was always or often similar to that of their peers.

When learners were not confined within the rigid interpretation of the topic, they connected the topic with their emotional and philosophical experiences, not just cognitively but personally. This autonomy created a space for reflective agency and identity expression, which were essential for meaningful belief development (Biesta & Tedder, 2007). Akkerman and Meijer (2011) emphasise that teacher identity is dynamic and dialogical, constructed through participation, negotiation, and reflection. The instructor did not engage in the peer discussions; rather, they allowed the learners to shape the discussion and reflect and strengthen their (teacher) beliefs, and find implementable contextual solutions to different challenges. Although class participation was voluntary, the learners participated owing to their prior familiarity with the teacher, peers, and their interest in the course structure. The familiarity with the peers' discussion could also be occurring from the nature of the questions discussed in the classes, like the role of teachers in classrooms, teacher's actions when faced with learners with learning differences and mental health challenges, and the use of (generative) AI and the role of teacher in the era of (generative) AI, which has spillovers to other professions. Mahmood et al. (2023) showed in an empirical study that peer comfort and psychological safety are essential for engaging in critical educational dialogue to promote relational agency. In a similar line, Zembylas & Schutz (2009) also discuss how

emotional safety and belonging are crucial for open discourse, especially in emotionally charged or critical topics. Our findings are in the same line that familiarity and safety enhanced discussions and helped learners exercise their agency.

Interviews also revealed that specific features of the course, such as the ability to propose topics, drive discussions, critically engage in topical issues that are not covered by disciplinary knowledge, provided a freedom where they could think critically, engage meaningfully with peers, redefine their beliefs and share their co-constructed beliefs among the other groups. When students are given freedom to select or interpret course topics with their personal experiences, it fosters a strong sense of ownership and authenticity in the topic's dimension selection. The learners also shared that reflecting and documenting their learning played a critical role in shaping their teacher beliefs. This is in line with the literature that this promotes deeper processing and facilitates belief refinement by giving students the time and freedom to revisit what resonated, what challenged them, and what they might carry into future practice, hence exercising their metacognitive agency (Farrell et al, 2013).

4. Discussion and Conclusion

We explore whether courses that are co-designed by instructors and learners are helpful in influencing learners to refine their teaching beliefs and exercise agency. We use a post-class reflective survey and data from interviews to address our questions. The findings of the study emphasise the importance of courses that build on teachers' real-life experiences, promote reflection and are embedded in the local contexts. While our results hold for this specific cohort of learners, we acknowledge the limitation that self-selection of learners into completing the reflective survey could have bias. Future research with a broader group of teachers would help us test the external validity of the results.

Acknowledgements

We would like to thank the participants of the course for participating in the survey and interviews that helped us gain valuable insight into the efficacy of the course.

References

Akkerman, S. F., & Meijer, P. C. (2011). A dialogical approach to conceptualizing teacher identity. *Teaching and Teacher Education, 27*(2), 308–319.

Biesta, G., & Tedder, M. (2007). Agency and learning in the life course: Towards an ecological perspective. *Studies in the Education of Adults, 39*(2), 132–149.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), 77-101.

Farrell, T. S. C. (2013). *Reflective writing for language teachers*. London, UK: Equinox.

Jain, R. (2007). A study of teaching effectiveness of teachers and their attitudes towards teaching profession. *Journal of Indian Education*, *33*(1), 77-89.

Mahmood, M., Talafian, H., Shafer, D., et al. (2023). *Navigating socio-emotional risk through comfort-building in a physics teaching community of practice*. arXiv:2310.16778

O'Neill, G., & McMahon, T. (2005). Student-centred learning: What does it mean for students and lecturers? *Emerging Issues in the Practice of University Learning and Teaching*, 1, 27–36.

Priestley, M., Biesta, G., & Robinson, S. (2015). *Teacher agency: What is it and why does it matter?* Bloomsbury Academic.

Yang, G., & Xiaochen, W. (2022). Towards understanding teacher mentoring, learner WCF beliefs, and learner revision practices through peer review feedback: A sociocultural perspective. *Journal of Language and Education*, *8*(4 (32)), 58-72.

Zembylas, M., & Schutz, P. A. (2009). *Advances in teacher emotion research: The impact on teachers' lives*. Springer

Zeng, L. M. (2020). Peer review of teaching in higher education: A systematic review of its impact on the professional development of university teachers from the teaching expertise perspective. *Educational Research Review*, *31*, 100333.

Panel

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Developing K-12 AI Literacy in Different Countries: Approaches and Challenges

Shuchi GROVER[a*], Balaraman RAVINDRAN[b], Siu-Cheung KONG[c], Maria M.T.RODRIGO[d], Lung-Hsiang WONG[e] & Weena NAOWAPRATEEP[f]
[a]*Looking Glass Ventures, USA*
[b]*IIT Madras, India*
[c]*Education University of Hong Kong, Hong Kong, China*
[d]*Ateneo Laboratory for the Learning Sciences, Philippines*
[e] *National Institute of Education, NTU, Singapore*
[f]*Institute for Innovative Learning, Mahidol University, Bangkok*
*shuchigrover@acm.org

Abstract: As artificial intelligence (AI) becomes increasingly pervasive across sectors, AI literacy has emerged as foundational for K-12 learners worldwide. How are education systems around the world addressing this challenge? This panel brings together representatives from India, Philippines, Singapore, Thailand, and the US to examine their strategies for advancing AI literacy through curriculum design, pedagogical innovation, professional development, and policy frameworks. The panelists will discuss their national approaches against the backdrop of initiatives such as India's AI Samarth program reaching over five million participants, Hong Kong's integration of AI modules into computational thinking curricula, Singapore's coordinated AI for Everyone strategy, Thailand's national AI curriculum development, Philippines' pilot efforts with the Education Center for AI Research (E-CAIR), and the US's 2025 executive order establishing AI education as a federal priority. Despite varied implementation pathways, shared challenges persist: preparing learners equitably, defining learning goals and developmentally appropriate AI concepts, addressing teacher readiness, aligning curricula with evolving frameworks, and ensuring that ethical dimensions receive adequate attention alongside technical competencies. Panelists will explore design principles, opportunities, and challenges in building inclusive, human-centered AI literacy that prepares students to thrive in an algorithmically mediated world.

Keywords: AI Literacy, global perspective, AI education, AI in education

1. Introduction

The development of "AI Literacy" has gained significant global momentum as artificial intelligence (AI) becomes an increasingly pervasive force across sectors shaping many aspects of civic life. It is now emerging as foundational for learners across disciplines and contexts. In the context of K–12 education, AI literacy refers not only to students' conceptual understanding of AI, but also to their ability to critically evaluate its applications, engage ethically with its use, and participate meaningfully in a world shaped by algorithmic systems.

Recognizing this growing imperative, the OECD, in collaboration with the European Commission and Code.org, recently released a draft AI Literacy Framework (OECD, 2025) for primary and secondary education. The effort follows similar, earlier efforts by global organizations such as UNESCO that released AI competency frameworks for students (cite) and teachers (cite). The OECD framework outlines the essential knowledge, skills, and attitudes students need to navigate AI-rich environments. It is organized into four core strands along with ethics being a cross-cutting competency: *Engaging with AI* which involves recognizing AI and using AI as a tool to access new content or recommendations. (2) *Creating*

with AI which involves guiding and refining AI output through prompts and feedback. (3) *Managing AI* which entails intentionally choosing how AI can support and enhance human work. (4) *Designing AI* empowers learners to understand how AI works and connect it to its social and ethical impacts by shaping how AI systems function. The framework is intended to guide national efforts, and will inform the upcoming PISA 2029 Media and AI Literacy test.

Within this emerging international consensus, countries and regions have begun taking concrete steps to integrate AI literacy into K–12 education systems. In India, the AI Samarth initiative represents the country's first large-scale AI literacy program for school-aged learners, providing a multilingual, open-source curriculum reaching over five million students, parents, and educators. In Hong Kong, AI modules have been integrated into existing computational thinking efforts such as CoolThink@JC (https://coolthink.hk/en), with support from the Education University of Hong Kong in designing tools and professional learning experiences for teachers. In the Philippines, the Department of Education's Center for AI Research (E-CAIR) aims to revitalize education by developing AI-driven tools that enhance teaching, learning, and school administration. Singapore continues its coordinated efforts under the AI for Everyone and AI for Students programs, embedding AI literacy in its broader Smart Nation (https://www.smartnation.gov.sg/) and digital readiness strategy. Thailand's main national-level effort is to develop and implement an AI curriculum from primary to secondary school, train thousands of teachers nationwide, and integrate AI literacy across subjects. In the United States, there have been concerted efforts at the federal as well as the state level since 2023. A 2025 executive order formally established AI education as a national priority, catalyzing federal investment in teacher training, ethical guardrails, and equitable access to AI tools in schools.

While the pathways differ across national and regional contexts, shared challenges remain. These include defining developmentally appropriate AI concepts, addressing teacher readiness, aligning curricula with evolving frameworks, and ensuring that ethical and societal dimensions of AI are addressed alongside technical ones. As with earlier waves of K-12 computing education and Computational Thinking efforts, the implementation of AI literacy calls for research-informed guidance, sustained investment, and collaboration among educators, policymakers, and researchers.

In this panel, representatives from six countries will share a glimpse of approaches in each of their geographies to advancing AI literacy through curriculum design, pedagogical strategies, professional development, and policy innovation. The following section shares brief abstracts of their presentations. Their experiences, albeit not exhaustive in terms of efforts in their respective countries, reflect both the local specificity and global urgency of equipping learners with the knowledge and dispositions needed to thrive in a world increasingly shaped by AI. The goal is to contribute to a deeper understanding of the design principles, opportunities, and challenges involved in building inclusive, human-centered AI literacy in schools.

2. Abstracts of Individual Panelists' Presentation

2.1 AI Literacy in India (Balaraman RAVINDRAN)

The Wadhwani School of Data Science and AI at IIT Madras, in collaboration with the Central Square Foundation, has developed AI Samarth—an open AI literacy curriculum framework for students, parents, and educators in government and affordable private schools. Grounded in a comprehensive curriculum landscaping study, the framework focuses on: (1) foundational AI concepts and limitations, (2) human agency and critical thinking, (3) real-world applications, (4) ethical and responsible use, and (5) meaningful, critical engagement with AI tools. The curriculum also incorporates sector-specific AI examples relevant to Indian contexts and bilingual resources to improve accessibility. Teacher capacity-building modules and ready-to-use classroom activities support effective implementation.

Key learning outcomes include understanding AI systems, recognising ethical issues, exercising judgment to avoid over-reliance, and applying AI to address real-world challenges while upholding human-centred values. Developed with input from educators, academics, and practitioners, and reviewed by K–12 teachers, AI Samarth is practical, relevant, and designed for scalable adoption across diverse school settings.

2.2 AI Literacy in Hong Kong (Siu-Cheung KONG)

Traditional education emphasizes guiding students to actively inquire, develop concepts, and construct their own knowledge. However, the operation of generative AI chatbots, based on large language models, differ significantly from this. These tools do not merely provide information; instead, they directly generate complete answers to specific questions. These “answer-oriented” responses, while convenient and efficient, also harbours the risk of weakening students’ abilities for independent thought and inquiry (Kong & Yang, 2024). One of the core objectives of education is to cultivate students’ ability to ask profound questions and to deepen their understanding through further questioning (Kong & Yang, 2025). If students become accustomed to easily obtaining ready-made answers from generative AI, their motivation and opportunities for deep thinking, critical evaluation of information, and even active questioning may diminish. Recent trends in the development of generative AI, such as features that automatically break down and refine a user’s question into multiple sub-questions, may seem helpful but take over the thinking and questioning processes that students ought to complete themselves. This developmental trend appears to run counter to the fundamental educational goal of inspiring thought and fostering a spirit of independent inquiry.

For learners aged 6 to 18, whose cognitive and thinking abilities are still developing, a great deal of practice is required. If AI tools designed for adults are applied indiscriminately to primary and secondary school students, it may hinder the development of students’ thinking skills. The education sector must be highly vigilant to prevent depriving students of the opportunity to experience necessary intellectual struggles and thereby build deep understanding and capabilities. Therefore, the application of generative AI in primary and secondary education settings must ensure that the locus of control for thinking and inquiry remains with the students (Kong et al., 2024) The role of generative AI tools should be to assist, not to replace. Opportunities should be provided for students to ask questions and find answers based on their own learning contexts and needs. In this process, if generative AI can provide options or initial responses, students must still bear the responsibility of integrating, evaluating, and deepening their understanding, transforming it into a part of their own knowledge system.

To address the teaching and learning challenges and responsibly integrate generative AI into basic education, a framework comprising three core dimensions is proposed: 1. Understanding the technology: popularising foundational knowledge of AI. 2. Using the technology: making good use of generative AI in teaching and learning. 3. Unleashing the potential of students: enhancing independent learning and thinking abilities (Kong, 2025). We need to cultivate students who not only know how to use generative AI, but who can also understand it, critique it, and ultimately, with its help, become more passionate and independent lifelong learners and unleash their potential. This requires teachers to possess the corresponding competencies and teaching strategies to guide students in the age of AI, enabling them to master technology while maintaining and developing unique human wisdom and judgement. This is a formidable but extremely important task, crucial for the core competitiveness of the next generation and the future development of society.

2.3 AI Literacy in the Philippines (Dr. Maria Mercedes T. RODRIGO)

In international standardized tests such as the PISA and the TIMSS, Filipino learners are among the poorest performers. In the 2018 PISA, Filipino students scored an average of 340 points in reading literacy, 353 points in math, and 357 points in science. In contrast, OECD countries averaged 487, 489, and 489 in the same areas respectively. As part of a

multipronged educational intervention effort, both the public and private sector are turning to AI to improve learning outcomes.

AI applications in education have three general orientations: student-facing, teacher-facing, and system-facing. Student-facing applications support learning by providing personalized learning paths and feedback. Teacher-facing applications provide assistance with lesson planning, classroom management, and student assessment. System-facing applications assist with data management and support data-driven decision making. We present examples of ways in which the Philippine educational system is attempting to integrate these three AI application orientations to support K12 education.

One application that is both student- and teacher-facing is Khan Academy. In 2023, Geraldine Acuna-Sunshine established Khan Academy Philippines (KAP), a non-profit organization with the goal of empowering Filipino learners with foundational and lifelong learning skills using Khan Academy as a platform. Khanmigo, Khan Academy's AI assistant, is the key technology powering the intervention. Khanmigo provides students with localized content and personalized tutoring in a number of local languages. Khanmigo also supports teachers by automating lesson planning, student progress tracking, and plagiarism detection. In partnership with the Philippines Department of Education (DepEd), KAP began with an initial cohort of 34 schools and 3000 learners in 2023 and has since extended its reach to over 1500 schools and 1.4 million learners in 2025.

Other examples of AI-based, student- and teacher-facing applications are the Reading Progress and Reading Coach systems of Microsoft. These applications enable teachers to assign reading assignments by uploading new material or selecting from a catalog. They allow customization of assignments based on learner needs. They also automate measurement of reading accuracy, speed. DepEd launched the use of Reading Progress and Reading Coach in schools in June 2025.

An initiative that addresses all three types of applications is DepEd's Education Center for AI Research (E-CAIR). E-CAIR's mission is to "[revitalize] basic education by developing AI-driven tools that enhance teaching, learning, and school administration." (DepEd, 2025). Among its programs are the systematic use of DepEd data to identify schools with infrastructure needs, the use of computer vision to screen undernourished children, enhanced detection of learners with disabilities, determination of training needs for faculty and administrators, and optimization of strategic public-private partnerships.

2.4 AI Literacy in Singapore (Lung-Hsiang WONG)

Singapore's approach to AI literacy in K–12 education reflects a hub-and-spoke ecosystem, where government, universities, industry, and schools contribute through differentiated yet interconnected initiatives. While there is no standalone national AI literacy curriculum, multiple entry points have emerged to embed AI-related competencies within existing educational structures. At the national level, AI Singapore—a government-funded national programme—offers a "LearnAI Student Outreach Programme", aimed at secondary and tertiary students. The programme introduces foundational concepts such as how AI works, its applications and limitations, and real-world ethical dilemmas. Sessions are delivered by trained AI Singapore ambassadors and adapted to school-specific contexts. For the general public, the "AI for Everyone" (AI4E) programme includes a free online AI Ethics & Governance for Youths course, targeting learners aged 16–19, and covering topics like algorithmic bias, privacy, and human-AI responsibility.

AI literacy has also been embedded into existing STEM and computational thinking efforts through programmes such as the Applied Learning Programmes (ALPs). Originally launched to deepen interdisciplinary thinking through STEM projects, many schools now incorporate AI-themed design challenges, robotics with vision or voice AI, and basic machine learning toolkits. These ALPs function as contextualized entry points for cultivating AI awareness and responsible use, without requiring formal curriculum overhauls.

In a recent development, the Ministry of Education announced that from 2026, *Computing* will become an elective subject for Secondary 3 (9th grade) students, replacing the former *Computer Applications* subject. This curriculum shift aims to raise students'

awareness of emerging technologies including AI, with differentiated depth across learning tracks. For instance, students will be introduced to cloud computing and AI as part of their digital and computational literacy education.

Beyond student-facing programmes, teacher capacity-building plays an important indirect role in supporting AI literacy. Some schools have initiated bespoke AI workshops for teachers, often inviting experts from universities such as National Institute of Education, Nanyang Technological University (NIE, NTU) to introduce current AI trends, tools, and pedagogical implications. These one-off sessions raise awareness of how AI can be responsibly introduced into school learning environments. At a more structured level, the "Leadership in Artificial Intelligence for Education" module within NIE's *Management and Leadership in Schools (MLS)* programme prepares middle leaders to conceptualize school-wide AI action plans. The module includes: (1) Foundations of AI and generative AI; (2) Adaptive/personalised learning systems; (3) Ethical issues in AI (bias, transparency, privacy); (4) AI-supported lesson planning and use-case design; (5) Strategic planning for school-level AI integration.

Singapore's decentralized yet coordinated efforts reflect an ongoing negotiation between local school-level initiatives and broader national ambitions. Rather than imposing a standardised curriculum, the hub-and-spoke model allows schools to explore AI literacy within existing STEM or coding-related programmes. This approach aligns with broader regional thinking that "the relentless progress in educational technologies… prompts us to reconsider the dynamics of AI-empowered learning" (Wong & Looi, 2024, p.1).

2.5 AI Literacy in Thailand (Weena NAOWAPRATEEP)

Thailand recognizes artificial intelligence as a critical component for national innovation and economic growth, particularly in preparing students for an AI-driven future. However, significant gaps exist in K-12 AI literacy, with over 50% of teachers having never used AI or operating only at basic levels according to recent surveys (OBEC, 2025).

A comprehensive three-tier approach has been adopted for AI literacy development that aligns with international best practices. The first tier, *learning with AI*, focuses on applying AI tools to improve teaching, learning, and school management. The second tier, *learning about AI*, builds students' understanding of AI fundamentals, capabilities, limitations, and ethical considerations. The third tier, *building AI*, emphasizes hands-on creation where students design and develop AI solutions to foster innovation. This framework is consistent with the OECD and European Commission's *AI Literacy Framework for Primary and Secondary Education* (OECD, 2025), which similarly emphasizes engaging with AI, creating with AI, managing AI, and designing AI.

Multi-agency collaboration between the Office of Basic Education Commission (OBEC), the Institute for the Promotion of Teaching Science and Technology (IPST), and the National Science and Technology Development Agency (NSTDA) has produced a five-module AI curriculum. These modules cover AI basics, machine learning, natural language processing, generative AI, and ethics, spanning elementary to high school. This aligns with UNESCO's AI competency framework that highlights human-centered mindsets and ethical considerations (UNESCO, 2024). Pilot programs have already reached 47 schools and 2,134 students, with teacher reports showing that 70% of elementary learners could explain AI concepts taught in class. In addition, the "Learn to Earn" program has enrolled 33,664 students in AI creation activities, while the Adaptive Education Platform supports 45,000 students with personalized learning.

Despite this progress, several challenges persist. Rural schools face infrastructure gaps that limit equitable access. Teacher preparedness varies: although curriculum materials are available, many educators require more training to confidently integrate AI into teaching. Assessment frameworks also lag, with limited tools to measure AI literacy. Moreover, cultural adaptation of international frameworks to Thai contexts requires ongoing refinement to ensure local relevance while maintaining global standards.

2.6 AI Literacy in the US (Shuchi GROVER)

Over the past five years, U.S. K-12 AI literacy has advanced through NSF-funded initiatives such as AI4K12.org that articulated the "Five Big Ideas in AI" (Touretzky et al., 2019) that have guided resource development, curriculum design, and educator networking (Grover, 2024). Complementary programs have expanded access to free curricula, teacher professional learning, and student outreach. Recent federal executive orders have elevated AI literacy as a national priority, directing agencies to support AI-integrated curricula, teacher training, and public–private partnerships. One outcome is the launch of the National Academy for AI Instruction (AFT, 2025), a national effort backed by industry partners to provide large-scale professional development (PD) for teachers.

31 US states have official guidance or policy on the use of AI in K-12 education as of this writing (TeachAI, 2024). Guidance currently is mostly teacher-facing, and policies typically emphasize data privacy, ethical use, oversight, and integration into existing curricula rather than standalone AI courses. Concurrently, individual school districts, especially large ones, are publishing their own guidance documents and policies that clarify AI (especially Generative AI use). Somewhat related, some states (such as Ohio) are mandating that all K-12 school districts create local policies to govern the use of AI. Initiatives such as TeachAI.org have created resources to guide policymakers, district administrators, as well as educators.

At least five states are actively piloting classroom-based curricula focused on helping students develop foundational understanding of AI concepts and applications. For example, Connecticut's pilot program provides grade 7–12 students with opportunities for hands-on inquiry into how AI systems function and impact society, while Indiana is supporting teacher training and digital learning environments that integrate AI into tutoring and everyday coursework (ECS, 2025).

Meanwhile, the K-12 computing education community is reimagining their role in preparing children for a future not only with computing, generally, but AI, more specifically. The Computer Science Teachers' Association (CSTA) has published "AI Learning Priorities for All K-12 Students" to guide the revision of K-12 CS standards with a view to covering the general and technical aspects of AI literacy that are imperative for all students. These include learning goals under themes: Humans and AI, Representation and Reasoning, Machine Learning, Ethical AI System Design and Programming, and Societal Impacts of AI.

Despite this momentum, the highly decentralized U.S. education system poses significant challenges. State and district autonomy results in uneven adoption. Rural and under-resourced schools face infrastructure gaps compounding equity concerns.

3. Discussion and Conclusion

The six panel presentations reveal critical connections between policy decisions, local needs, teacher preparation, and curriculum implementation across diverse educational contexts. The perspectives from various countries also reveal a consistent theme, especially in Asia: the collaborative role being played by higher education institutions, researchers, and other partners in designing AI Literacy curricula and nationwide rollout efforts. Three themes will be explored in depth during the conference: the boundaries between technical skill development and curricula focused on AI awareness and use; the role of teacher PD as both catalyst and bottleneck for implementation; and the challenge of creating culturally responsive AI literacy that addresses local contexts while meeting global competency standards. The discussion will synthesize insights from varied national approaches to identify design principles and implementation strategies. The significance of these contributions lies in advancing contextualized frameworks for AI literacy that balance learning goals, informing policy and practice as countries worldwide navigate this emerging educational imperative.

References

American Federation of Teachers (AFT). (2025). AFT to launch National Academy for AI Instruction with Microsoft, OpenAI, Anthropic and United Federation of Teachers [Press release]. https://www.aft.org/press-release/aft-launch-national-academy-ai-instruction-microsoft-openai-anthropic-and-united

AI4K12.org n.d. Grade band progression charts. https://ai4k12.org/gradeband-progression-charts/.

CSTA & AI4K12 (2025). AI Learning Priorities for All K-12 Students. New York, NY: Computer Science Teachers Association. Retrieved from https://csteachers.org/ai-priorities.

Education Commission of the States (ECS). (2025). AI pilot programs in K-12 settings. https://www.ecs.org/ai-artificial-intelligence-pilots-k12-schools/

Kong S. C., Cheung W. M. Y. & Tsang O. (2024). Developing an artificial intelligence literacy framework: Evaluation of a literacy course for senior secondary students using a project-based learning approach. Computers and Education: Artificial Intelligence, 100214. https://doi.org/10.1016/j.caeai.2024.100214.

Kong, S. C., & Yang, Y. (2024). A human-centered learning and teaching framework using generative artificial intelligence for self-regulated learning development through domain knowledge learning in K-12 settings. IEEE Transactions on Learning Technologies, 17, 1562-1573. https://doi.org/10.1109/TLT.2024.3392830

Kong, S. C., & Yang, Y. (2025). A study on the empowerment of senior secondary students in artificial intelligence literacy. Research and Practice in Technology Enhanced Learning, 20. https://doi.org/10.58459/rptel.2025.20024

Kong, S.C. (2025). The challenges of using generative artificial intelligence in school education. In Exploring How to Drive Interdisciplinary Co-creation and Collaborative Learning in the Era of Generative AI. Hong Kong: Sino United Electronic Publishing Ltd.

Grover, S. (2024). Teaching AI to K-12 learners: Lessons, issues, and guidance. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education* V. 1 (pp. 422-428).

OBEC (Office of the Basic Education Commission, Thailand). (2025). *AI guidance* [PDF]. NITEDCR1. https://nitedcr1.go.th/wp-content/uploads/2025/03/OBEC-AI-Guidance_.pdf

OECD (2025). Empowering learners for the age of AI: An AI literacy framework for primary and secondary education (Review draft). OECD. Paris. https://ailiteracyframework.org

TeachAI. (2025). Foundational policy ideas for AI in education. https://www.teachai.org/policy

Touretzky, D., Gardner-McCune, C., Martin, F., & Seehorn, D. (2019). Envisioning AI for K-12: What should every child know about AI?. In Proceedings of the *AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 9795-9799).

Wong, L.-H., & Looi, C.-K. (2024). Advancing the generative AI in education research agenda: Insights from the Asia-Pacific region. *Asia-Pacific Journal of Education, 44*(1), 1-7. https://doi.org/10.1080/02188791.2024.2315704

UNESCO (2024). *AI competency framework for teachers and students*. UNESCO Publishing.

Jiang, B. et al. (Eds.) (2025). Proceedings of the 33rd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Professional Development of STEM Teachers Using the DECODE Model in India and Taiwan

P. S. SREEDEVI [a*], P RAJASEKARAN [b], Chun-Yen CHANG [c], Marison Sudianto MANALU[d]

a,b-Department of Education, The Gandhigram Rural Institute, India
c-Graduate Institute of Science Education and Institute for Research Excellence in Learning Sciences, National Taiwan Normal University, Taiwan, Graduate School of Education, Chung Yuan Christian University, Taiwan, Department of Biology, Universitas Negeri Malang, Indonesia
d- National Taiwan Normal University, Taiwan

Abstract: This study will explore the impact of the DECODE model for STEM teacher professional development, drawing on empirical evidence from a year-long India–Taiwan collaborative project. The DECODE framework integrates structured teacher learning cycles where Demonstration (DE) provides model practices, Co-training/Co-design (CO) engages teachers collaboratively in lesson design and peer learning, and Debriefing (DE) offers reflection and feedback for continuous improvement. Together, these components integrate technology, pedagogy, and content knowledge to prepare teachers for rapidly evolving educational contexts, particularly in AI-supported scientific inquiry and technology integration. Workshops conducted with pre-service and in-service teachers in India, alongside field visits and collaborative research in Taiwan, revealed significant improvements in teachers' competencies, confidence, and ability to integrate innovative tools such as CloudClassRoom (CCR) and AI platforms into STEM curricula. The research findings compare professional development models from India and Taiwan, and present strategies for designing adaptable, cross-cultural training programs. The study recommendations for policymakers, educators, and researchers interested in scaling hybrid teacher training models for global STEM education needs.

Keywords: STEM education; STEM teacher professional development; online training model; DECODE

1. Introduction

The DECODE model for STEM teacher professional development combines Demonstration (DE), Collaborative co-design/training (CO), and Debriefing/feedback (DE) cycles to foster technological pedagogical content knowledge (TPACK). In 2024–25, a joint India–Taiwan project applied this model to enhance pre-service and in-service teacher competencies in integrating innovative technologies such as AI-Supported Scientific Inquiry (AISI) and CloudClassRoom (CCR).STEM-TPD is designed to engage teachers in active learning and focus on teachers' needs to STEM teaching successfully. When teachers learn actively in quality PD, teachers would increase their self-efficacy in STEM teaching, gain STEM experiences (Thibaut et al., 2018), and realise the relevance of teachers' professional lives (Cavlazoglu & Stuessy, 2017). Integrated STEM teaching advocates that active learning is an essential environment for students to construct new ideas for authentic problem-solving and designing(Heba et al., 2017; Nadelson et al., 2013). Teachers perhaps need to become active learners in the quality PD opportunities before implementing STEM education that requires an active learning environment. Researchers sustained that the rationale for organizing STEM-TPDs is to improve teacher practices by enhancing superior knowledge needed for the betterment of STEM implementations. The authentic and up-to-date content knowledge was

transformed in TPDs, such as earthquakes or nanotechnology (Fore et al., 2015).STEM-TPD aims to transform instructional models such as inquiry-based teaching, how to teach social scientific issues (SSIs) or engineering challenges (Macalalag et al., 2020; Nadelson et al., 2012), and how to integrate technology in STEM classrooms to facilitate awareness of STEM careers (Parker et al., 2015).

1.1 ***Workshop Description***

The workshops conducted under the DECODE professional development framework have been intentionally designed as purely online programs, enabling teachers to participate regardless of geographical or institutional constraints. The virtual nature of these workshops ensures accessibility and inclusivity, allowing educators from rural, urban, and international contexts to engage with the same training quality.

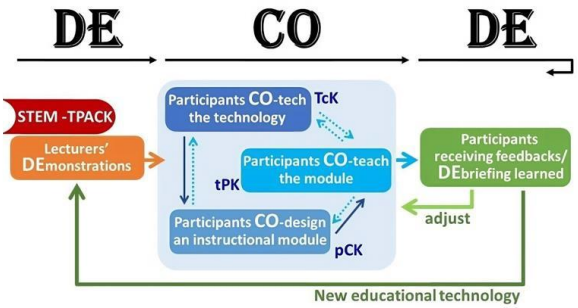


Fig-1: DECODE model (Adopted from Cheng et al.,2022)

1.1.1 Demonstration-DE (Session 1)

Taiwanese experts introduced and demonstrated the AISI and CCR platforms, highlighting the importance of scientific inquiry and artificial intelligence in education. Participants explored the platforms to understand their features. (Rajasekaran et al., 2024)

1.1.2 CO-teach&CO-train (Sessions 2, 3, and 4)

Participants were divided into groups based on subjects in these sessions, ensuring diverse contributions. Facilitated by the Indian research team, these offline sessions involved exploring the platforms, topic selection, and collaborative planning. Groups finalised topics with research team consultations, prepared lesson plans incorporating AISI and CCR, and selected group leaders to coordinate tasks.

1.1.3 DEbrief-DE (Session 5)

Participants presented their topics in a hybrid mode to a large audience, including Taiwanese experts, institution teachers and students, and the research team. Feedback from Taiwanese experts aimed to improve presentations and assess the professional development of Pre-Service teachers post-workshop.

1.2 *Literature Review*

Information technology capacity is a prerequisite in the 21st century with the continuous development of science and technology. All experts need to use information technology in their work from doctors or scientists. Teachers are also one of them. Teachers need to use information technology to integrate into their lessons, especially during the Covid-19 epidemic. Technology is a tool in teachers' lessons and becomes the goal of classes. Technology plays an even more essential role in STEM lessons. Technology (T) is a critical element in STEM lessons. Consequently, the technological pedagogical and content knowledge (TPACK) framework was advocated as the framework to examine the technology,

pedagogy and content that the current STEM-TPD is targeting (Chai, 2019). Teachers utilized instructional practices to access students achieve such valuable STEM competencies. For example, the 6E instructional model, project-based learning (PjBL), problem-based learning (PBL), and engineering design process (EDP) (Wahono et al., 2020). Teachers dynamically choose any instructional model to implement STEM education successfully.

Thibaut et al. (2018) reviewed 23 STEM interventional papers in terms of instructional practices to synthesize five principles in STEM teaching. Such five principles are rooted in the social constructivist view of learning theory with student-centred pedagogies. STEM teaching could be successful when teachers could adapt pedagogical approaches based on the social constructivist view, inquiry and engineering design.

Vrasidas and Zembylas (2004) agreed with the effectiveness of technology integration in TPD but note the lack of research-based frameworks to develop and evaluate TPD. In addition, there is a dilemma in balancing between philosophical and pragmatic focusing on TPD in Science Education (Mundry & Loucks-Horsley, 1999) because of the Nature of Science. The rapid developments of science and technology call for innovative teaching to educate individuals well-equipped with academic knowledge as well as 21st-century skills. Integrate Science is advocated as a promising place for better education such as science with engineering integrated. Antink-Meyer and Meyer (2016) indicated Science teachers' misconceptions of Science and Engineering distinctions in online TPD.

The ultimate of STEM-TPDs positively increase students' learning outcomes once TPDs would directly improve teachers' both perceptions and practice (Han et al.,2015; Nadelson et al., 2013). Besides, teachers' specific professional knowledge such as pedagogical knowledge (PK), content knowledge (CK), pedagogical content knowledge (PCK), and technological pedagogical content knowledge (TPCK) should be activated and developed for successful STEM implementation (Awad et al., 2019; Chai, 2019). In the following paragraphs, we identified key features of effective STEM- TPD in response to the STEM assessment, STEM teaching, and the global trend. Based on our research, the Technological Content Knowledge in response to technological advancement, such as teaching techniques, and assessment technology, but not including learning management system.

DECODE model (Figure 1) can facilitate teachers' critical reexamination of the affordances of the innovative technologies for their teaching practices from the views of subject matter selection, motivation empowerment, information presentation, activity design, and pedagogy transition. It was also found that the DECODE model can facilitate teachers' TPACK towards a more connected model that addresses accessible technologies, pedagogy, and subject matter jointly (Cheng et al., 2022). The DECODE model includes three stages: (1)DE: teacher's DEmonstrations, (2)CO: students CO-train the use of CloudClassRoom, students CO design an educational technology-integrated course, (3)DE: students CO-teach, eventually students receive feedbacks and DEbrief what they have learned through the stages mentioned above. For every type of technology, we will run through DECODE once. The length of time determined for running through the DECODE will be based on the type and characteristics of the technology. In each DECODE, at least two rounds of DE-CO-DE- CO-DE will be conducted to promote and strengthen teachers' familiarity and mastery of the technology used in STEM education.

The study conducted a comparative analysis of STEM teacher training models in India and Taiwan, focusing on professional development programs designed to equip educators with the skills required to teach in rapidly evolving scientific and technological landscapes. The study's objectives are twofold: (1) to identify the existing professional development frameworks and approaches utilized for STEM teachers in both countries and (2) to evaluate the effectiveness of these models in preparing educators to address current and future educational demands. With an increasing emphasis on STEM education globally, this analysis highlights each country's approach to integrating recent advancements, such as Artificial Intelligence, Internet of Things (IoT), and interdisciplinary STEAM strategies, into professional training.

2. Abstracts of Individual Panelists' Presentation

2.1 International Collaboration and Cross-Cultural Framework Adaptation (Chun-Yen CHANG)

This presentation highlight the role of international collaboration in adapting the DECODE framework to diverse cultural and educational contexts. Drawing on Taiwan's expertise in competency-based STEM education and AI-enhanced learning environments, the talk will outline how tools such as CloudClassRoom (CCR) and AI-supported Inquiry (AISI) can be localized for effective integration in India. Prof. Chang will also share lessons from field visits and collaborative workshops, emphasizing how cross-national partnerships strengthen professional development models and promote scalable, globally relevant practices in STEM education.

2.2 Applying the DECODE Model in Pre-Service Teacher Training (P. S. SREEDEVI)

This presentation focus on empirical evidence gathered from training 187 pre-service teachers across Kerala and Tamil Nadu using the DECODE model. The sessions integrated CCR and AI platforms, enabling participants to engage in collaborative lesson design, active inquiry, and technology-enhanced pedagogy. The findings reveal substantial improvements in teacher confidence, digital literacy, and capacity to design STEM lessons responsive to real-world contexts. Prof. Sreedevi will discuss the challenges and opportunities of implementing the DECODE model in Indian teacher education institutions, offering insights for adapting it in developing-country contexts.

2.3 Technology Integration and Curriculum Innovation in STEM Education (Dr. Marison Sudianto Manalu)

This presentation address how technology integration and teacher professional development can be enhanced through the DECODE model. Drawing from postdoctoral research at National Taiwan Normal University, Dr. Manalu will explore how CloudClassRoom (CCR) and AI-supported Inquiry (AISI) as online collaborative environments facilitate authentic scientific inquiry in STEM classrooms. Case studies from Indonesia will illustrate how the integration of the technology enhances teacher professional development, ensuring the sustainability and adaptability of the model across varied institutional settings.

2.4 Instructional Design and Teacher Training through DECODE (Mr.P. Rajasekaran)

This presentation examine the instructional design dimension of the DECODE framework, focusing on how structured cycles of demonstration, co-design, and debriefing foster deeper professional growth among teachers. Based on research with pre-service and in-service teachers, Mr. Rajasekaran will present data on how collaborative workshops enhanced reflective practices and peer-supported learning. The presentation will also address the scalability of DECODE-based training through online and hybrid delivery models, emphasizing its potential to overcome geographical and institutional barriers in teacher professional development.

3. Discussion

Data collected quantitatively and qualitatively from teachers in order to understand the effects of the DECODER TPD on teachers' technological competencies, as well as perceptions and practice of technology integration in their science teaching. The research team conducted three workshops in preservice teacher education institutions (B.Ed, Special B.Ed & D.El.Ed),

in two states (Kerala and Tamil Nadu) for 187 Preservice teachers with the DECODE Model. The professional development was measured through a rating scale constructed based on five stages of professional development.

The results indicated significant improvements in their professional skills. The preservice teachers gained skills like incorporating Artificial intelligence, scientific inquiry, and cloud classrooms in their regular classrooms. This study highlights the significant impact of professional development for pre- service teachers and suggests the DECODE model of Professional development training for teachers in India.

About the comparative analysis of STEM teacher training models in India and Taiwan, we focus on professional development programs designed to equip educators with the skills required to teach in rapidly evolving scientific and technological landscapes. Data gathered from a series of focus group discussions and qualitative surveys reveal the structure, accessibility, and content diversity within each model, showcasing how different educational contexts influence program design. Findings underscore the strengths and limitations of each country's training model, offering insights into the adaptability of these approaches across cultural and educational settings. Recommendations are proposed for cross-collaborative training initiatives between India and Taiwan to enhance STEM education through shared expertise and resources. This study contributes to the global discourse on professional development in STEM education, emphasizing the need for adaptable, accessible, and forward-looking training programs for teachers. In addition, the research team also had a field visit & observation of professional development Training for STEM Teachers in Taiwan in the first year and continued to plan future cooperation.

4. Conclusion

The panel underscores the effectiveness of the DECODE model as a transformative framework for STEM teacher professional development across diverse cultural and institutional contexts. By combining structured cycles of demonstration, co-design, and debriefing, DECODE not only strengthens teachers' technological pedagogical content knowledge (TPACK) but also fosters collaboration, reflection, and innovation in teaching practice. Evidence from India and Taiwan demonstrates that pre-service and in-service teachers gained significant improvements in confidence, digital integration, and inquiry-based pedagogy. Moreover, the cross-national partnership highlights the adaptability of the model to varying educational systems and resource settings. Looking ahead, the panel advocates for scaling DECODE as a hybrid, accessible, and sustainable professional development approach capable of preparing educators for AI-rich, globally connected STEM classrooms. The insights generated will serve as actionable recommendations for policymakers, teacher educators, and researchers committed to building resilient and future-ready teacher training ecosystems.

Acknowledgements

This work was financially supported by the National Science Council of Taiwan under contracts the MOST 111-2410 H003-032-MY3, and the NSTC 111-2423-H-003-004 and the "Institute for Research Excellence in Learning Sciences" of National Taiwan Normal University (NTNU) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan and 113-2923-H-003 -001 -MY2.
Professional Development of STEM Teachers Uisng the DECODE in India and Taiwan & Joint Research project under ICSSR(India) NSTC(Taiwan)/JRP4/2024-25/IC).

References

Antink-Meyer, A., & Meyer, D. Z. (2016). Science Teachers' misconceptions in science and

Engineering Distinctions: Reflections on modern research examples. *Journal of Science Teacher Education*, *27*(6), 625–647. https://doi.org/10.1007/s10972-016-9478-z

Awad, N. A., Salman, E., & Barak, M. (2019). Integrating Teachers to Teach an Interdisciplinary STEM-Focused Program about Sound, Waves and Communication Systems. *European Journal of STEM Education*, *4*(1). https://doi.org/10.20897/ejsteme/5756

Cavlazoglu, B., & Stuessy, C. (2017). Changes in science teachers' conceptions and connections of STEM concepts and earthquake engineering. *The Journal of Educational Research*, *110*(3), 239–254. https://doi.org/10.1080/00220671.2016.1273176

Chai, C. S. (2018). Teacher Professional Development for Science, Technology, Engineering and Mathematics (STEM) Education: A Review from the Perspectives of Technological Pedagogical Content (TPACK). *The Asia-Pacific Education Researcher*, *28*(1), 5–13. https://doi.org/10.1007/s40299-018-0400-7

Cheng, P.-H., Molina, J., Lin, M.-C., Liu, H.-H., & Chang, C.-Y. (2022). A New TPACK Training Model for Tackling the Ongoing Challenges of COVID-19. Applied System Innovation, 5(2), 32. https://www.mdpi.com/2571-5577/5/2/32

Fore, G. A., Feldhaus, C. R., Sorge, B. H., Agarwal, M., & Varahramyan, K. (2015). Learning at the nano-level: Accounting for complexity in the internalization of secondary STEM teacher professional development. *Teaching and Teacher Education*, *51*, 101–112. https://doi.org/10.1016/j.tate.2015.06.008

El-Deghaidy, H., Mansour, N., Alzaghibi, M., & Alhammad, K. (2017). Context of STEM Integration in Schools: Views from In-service Science Teachers. *Eurasia Journal of Mathematics Science and Technology Education*, *13*(6). https://doi.org/10.12973/eurasia.2017.01235a

Macalalag, A. Z., Johnson, J., & Lai, M. (2019). How do we do this: learning how to teach socioscientific issues. *Cultural Studies of Science Education*, *15*(2), 389–413. https://doi.org/10.1007/s11422-019-09944-9

Mundry, S., & Loucks-Horsley, S. (1999). Designing Professional Development for Science and Mathematics Teachers: Decision Points and Dilemmas. *Nise Brief*, *3*(1), n1.

Nadelson, L. S., Seifert, A., Moll, A. J., & Coats, B. (2012). I-STEM Summer Institute: An Integrated Approach to Teacher Professional Development in STEM. *Journal of STEM Education*, *13*(2), 69–83. https://eric.ed.gov/?id=EJ978728

Nadelson, L. S., Callahan, J., Pyke, P., Hay, A., Dance, M., & Pfiester, J. (2013). Teacher STEM Perception and Preparation: Inquiry-Based STEM professional Development for elementary teachers. *The Journal of Educational Research*, *106*(2), 157–168. https://doi.org/10.1080/00220671.2012.667014

Parker, J. G., Rubin, K. H., Erath, S. A., Wojslawowicz, J. C., & Buskirk, A. A. (2015). *Peer Relationships, Child Development, and Adjustment: A Developmental Psychopathology Perspective*. https://doi.org/10.1002/9780470939383.ch12

Rajasekaran, R., Sreedevi, P. S., & Chang, C. (2024). DECODE-based STEM workshop in improving academic resilience and teaching competency of pre-service teachers. *Eurasia Journal of Mathematics Science and Technology Education*, *20*(2), em2406. https://doi.org/10.29333/ejmste/14243

Thibaut, L., Knipprath, H., Dehaene, W., & Depaepe, F. (2018). The influence of teachers' attitudes and school context on instructional practices in integrated STEM education. *Teaching and teacher education*, *71*, 190-205.https://doi.org/10.1016/j.tate.2017.12.014

Vrasidas, C., & Zembylas, M. (2004). Online professional development: lessons from the field. *Education + Training*, *46*(6/7), 326–334. https://doi.org/10.1108/00400910410555231

Wahono, B., Hariyadi, S., & Subiantoro, A. W. (2022). The development of an online STEM teacher professional development package with the DECODE model: An innovative teacher's quality maintenance. Eurasia Journal of Mathematics, Science and Technology Education, 18(12).https://doi.org/10.29333/ejmste/12647:

