Rodrigo, M. M. T. et al. (Eds.) (2021). Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Identifying Students' Stuck Points Using Self-Explanations and Pen Stroke Data in a Mathematics Quiz

Ryosuke NAMAMOTO[a*], Brendan FLANAGAN[b], Kyosuke TAKAMI[b], Yiling DAI[b], & Hiroaki OGATA[b]

[a]*Graduate School of Informatics, Kyoto University, Japan*

[b]*Academic Center for Computing and Media Studies, Kyoto University, Japan*

*s0527225@gmail.com

Abstract: During the process of learning, students face challenging quizzes that require various knowledge and skills, which results in stuck points. Identifying concepts that cause these stuck points can help identify potential areas of remedial study to overcome the students' difficulties. To achieve the goal of generating a model to identify students' stuck points in a mathematics quiz, we attempted to discover highly influential features using self-explanations and pen stroke data collected from a digital reading system named BookRoll. Firstly, we created rubrics as a gold standard of stuck points, and attempted to estimate students' correctness of each rubric step based on their self-explanations. Next, we extracted features by processing the behavioral data including the pen stroke data associated with the self-explanations. Finally, we divided the data into two groups based on the rubric step correctness, and performed statistical analysis and logistic regression analysis on the extracted features. The results showed that rubric based self-explanation scores significantly differed between the groups, and also the behavioral data such as "operation time" and "operation order" had impacts on identifying the students' stuck points. Future research topics include increasing the number of samples for the quiz and developing a model that can identify the stuck points automatically from a predefined knowledge structure.

Keywords: Stuck point, self-explanation, automated self-explanation scoring, mathematics education, pen stroke data

1. Introduction

Recently, the use of digital learning environments are providing increasing insight into learning behavior through the collection of system interaction data. This has led to new fields of research, such as educational data mining and learning analytics.

Meanwhile, self-explanation has been widely recognized for its learning effects for a long time (Bisra, et al. 2018). It is a good indicator of the students' understanding towards solving a quiz because there are cognitive factors behind explanation (Lombrozo, 2006). Self-explanation is defined as explaining concepts, procedures, and solutions in order to deepen understanding of material and to make sense of relatively new information (Chi and de Leeuw, 1994; Rittle-Johnson, 2006).

However, few research has been conducted on the use of self-explanations in learning analytics to identify stuck points, to examine intervention methods, and to use them to recommend quizzes. Students who can well explain their solving processes possibly have clear solution strategies and good understanding of necessary knowledge and skills to solve the quiz. On the other hand, students who cannot explain how they solved the quiz may be stuck in some solving steps. If a system can automatically identify a student's stuck points, the teacher can not only learn about the student's level of understanding, but the system can also support the student to overcome his or her difficulties. In this study, we investigate influential features relating to students' stuck points by utilizing two types of data collected from a digital learning system: a) the self-explanations generated when the students are reviewing their answers, and b) the pen stroke data generated when the students are answering the quiz.

2. Related Work

2.1 Self-Explanation and Handwriting Input Analysis in Mathematics Education

Self-explanations and handwriting logs were utilized in a lot of research in mathematics education. Hodds, Alcock and Inglis (2014) found students who received the training generated higher quality explanations and performed better on a comprehension test in math quizzes. Renkl (1997) stated in mathematics quizzes solving, more successful learners tended to self-explain by predicting the next steps and identifying the overall goal structure and its subgoal objectives.

On the other hand, handwriting data analytics has been studied for detecting answer stuck points (Iiyama et al., 2017). Kishi and Miura (2018) developed a system to detect a learner's weak points using time intervals of pen stroke data in a mathematics quiz, and Ochoa et al. (2013) could discriminate between experts and non-experts in groups of students solving mathematical problems with multi-modal learning analytics including pen stroke data.

To the best of our knowledge, our study is the first to combine self-explanations and pen stroke data in identifying stuck points, where the pen stroke data provides temporal and spatial hints on "where" the students were stuck and self-explanations help to infer specifically "what" knowledge the students were stuck on.

2.2 Automated Scoring of Self-Explanation

McNamara, Levinstein and Boonthum (2004), created an interactive tutoring system called iSTART to support the development of self-explanation skills in reading comprehension. The system guides learners through the exercise to support active reading and thinking by providing automatic evaluation and scoring of self-explanations and appropriate scaffolding. Natural language processing techniques, such as: Latent Dirichlet Allocation (LDA) Analysis topic modeling, was applied to extract the characteristics of student self-explanation artifacts and their similarity with reading materials are used to analyze the self-explanations for automatic scoring. This system was modeled on an effective human-delivered intervention called self-explanation reading training (SERT).

Prior research has discussed methods for scoring self-explanations, but there has been little research on using scored self-explanations to identify stuck points, and there has been little research combining self-explanations and handwriting data. The purpose of this study is to examine how to score self-explanations and process pen stroke data on answers to secondary school mathematics quizzes to identify a student's possible quiz concepts or misconceptions that led to stuck points.

3. Method

3.1 Data Collection

We collected the data over the period of May 14-23, 2021 using the LEAF platform (Flanagan and Ogata, 2018), which consists of a digital reading system named BookRoll, and a learning analytics tool LAViEW where students and teachers can monitor and reflect on their learning. The platform was deployed in a Japanese secondary school and has been in use for several years. Firstly, students were asked to view the quiz and write down their answers using a stylus and tablet computer with the handwriting input in BookRoll as shown in Figure 1. BookRoll captures the handwriting data as a series of vectors representing the coordinates and velocity of pen strokes, which allows realistic playback of the handwritten answers and fine-grained analysis of the students' answering process. Preliminary analysis of such data has been examined by Yoshitake, Flanagan, and Ogata (2020) to support group learning. Secondly, when the students had finished answering the quiz, they were asked to review their handwritten answers in LAViEW, and input explanations of how they solved the quiz. As illustrated in Figure 2, the students input a sentence of explanation everytime they think they have completed some step in their answers during the playback. Therefore, the self-explanations are temporally associated with the pen stroke data. In this study, we selected one quiz for detailed analysis which has 217 self-explanation sentences generated by 35 students.

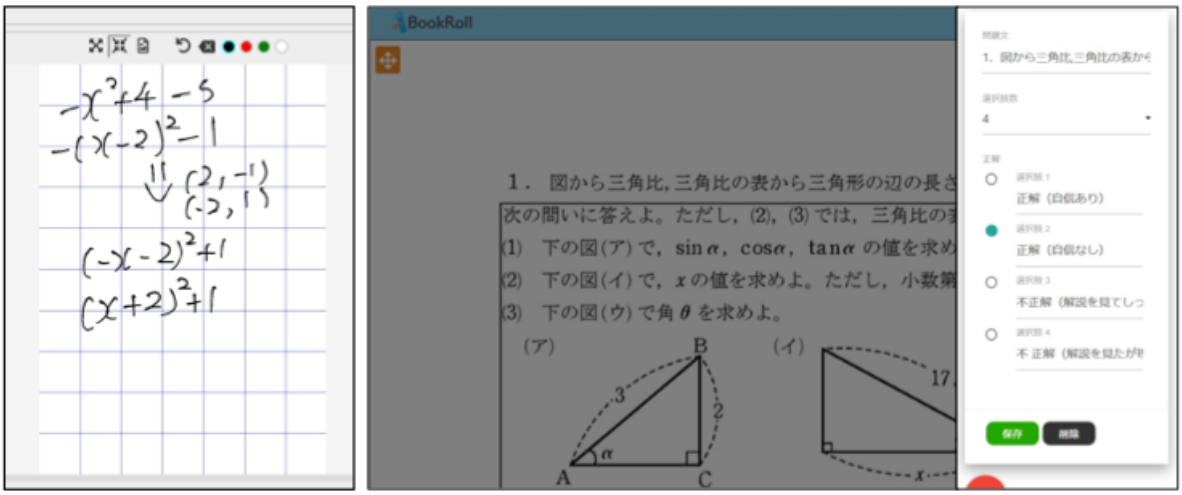


Figure 1. BookRoll user interface: handwritten answer (left), quiz and quiz answer report (right).

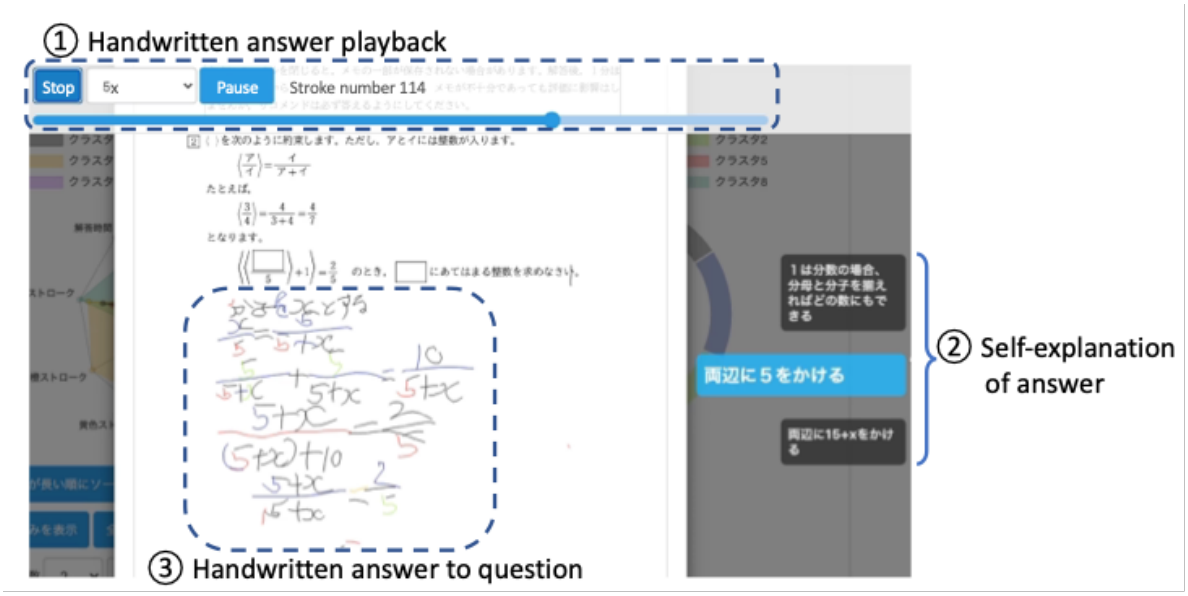


Figure 2. Handwritten answer playback and self-explanation input.

3.2 Evaluation Design Based on Rubrics

Identifying stuck points in the answering process is very ambiguous, and sometimes difficult to evaluate even for teachers. In this study, we attempted to construct an objective and easy-to-understand evaluation system by rubrics. A rubric is an assessment tool which describes varying levels of quality of complex student reasoning, performances, or products from excellent to poor for a specific assignment (Andrade, 2000; Arter and Chappuis, 2007). It also provides valuable information about the subject matter that students are trying to understand and clarify what the learning objectives are. In this study, we applied rubrics using them as a tool to itemize the knowledge required to solve a mathematical quiz, and to visualize students' answering process and difficulties.

Two of the authors created the rubric based on the textbook and the formulas needed to solve the quiz with the following purpose and process. Firstly, the units that make up the quiz were clarified. As shown in Table 2, by breaking down the quiz into units, the rubric aimed to clearly separate the areas where students were doing well from those where they were not. Also, we tried to make human judgments visible in an objective way with the rubric. Secondly, labels which indicate the correctness for each rubric number for each student were created by two of the authors with the aim of using them as objective variables. Cohen's Kappa (Cohen, 1960) was used to determine the inter-rater reliability between the two authors, and indicates a high degree of agreement with a Kappa of 0.882. The

difference of labels was solved by the discussion between the authors. Lastly, sample self-explanation sentences were created to check whether students' self-explanation sentences covered all the necessary elements for solving quizzes. The details of this scoring process will be discussed in Section 3.3.

Table 1 shows the rubric definitions, labels, and sample sentences of self-explanations that were created to match the five rubric numbers. Table 2 shows the rubric for the quiz, which consists of five steps.

Table 1. *Definitions of evaluation tools*

Name	Definitions
Rubric	Can-do descriptors that clearly describe all the essential elements of the quiz and are used to create labels and sample self-explanations for scoring. Ordinal Scale (1-5).
Labels	Labels consist of true or false for each of the rubrics 5 steps, subsequently referred to as "correct step" or "incorrect step" answers. In particular, "Incorrect Step" signifies the point at which the student got stuck.
Sample Sentences of self-explanations	Model answers of self-explanations prepared according to the 5-step rubric number.

Table 2. *Rubrics and sample sentences of self-explanation.*

Number	Rubric	Sample Sentences of Self-explanations
Step 1	Be able to find the equation of a linear function from two points.	Substituting the y-coordinate of p into the equation of the line AC.
Step 2	Be able to find the equation of the line that bisects the area of a triangle.	Find the area of triangle ABC, and then find the area of triangle OPC.
Step 3	Be able to represent a point on a straight line using letters (P-coordinates).	With the line OC as the base, find the y-coordinate of p, which is the height. p's coordinate is (t,-1/2t+4).
Step 4	Be able to represent a point on a straight line using letters (Q-coordinate).	Since the coordinates of P are (3,5/2), the line OP is y=5/6 and the coordinates of Q are placed as (t,5/6).
Step 5	Be able to formulate an equation for area based on relationships among figures.	Finally, the area of ∆QAC was found from the areas of ∆AQO and ∆OQC, and the coordinates of Q were found.

3.3 Rubric Based Self-Explanation Scoring

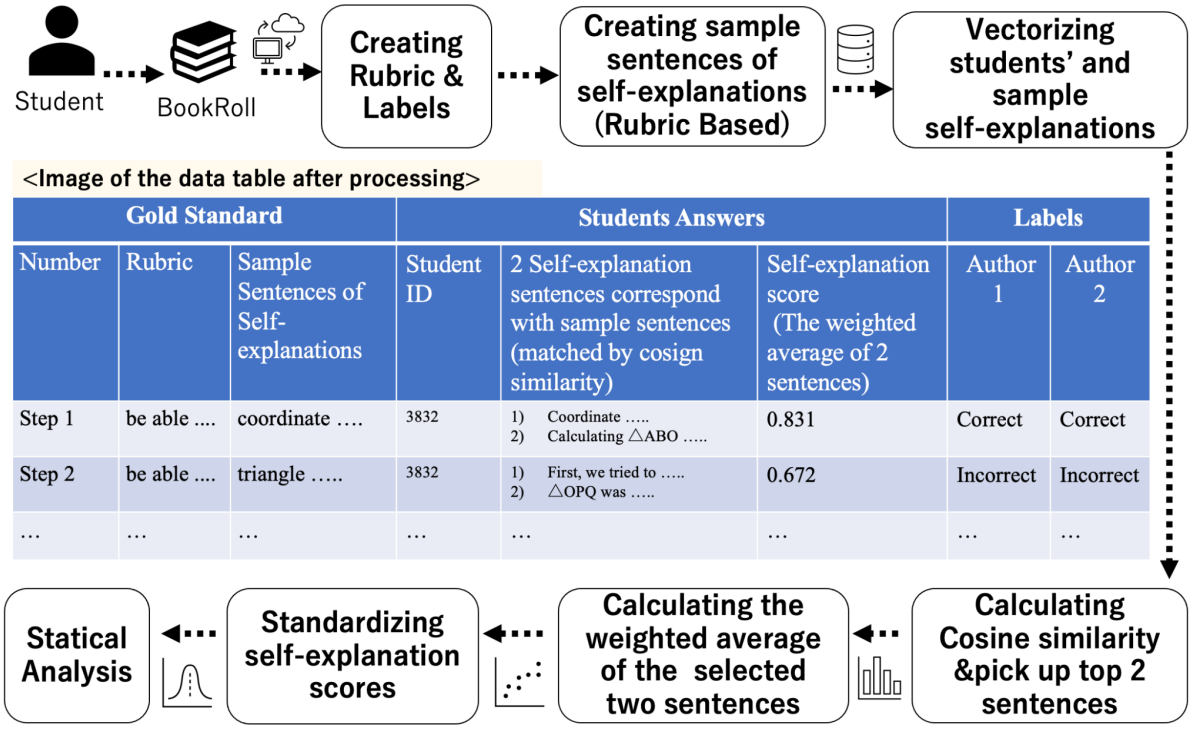
Explanations are accompanied by a sense of understanding(Lombrozo, 2006), and we assumed self-explanation would be a good indicator of the student's knowledge on solving the quiz. If the student wrote poor self-explanations of a rubric step, it is highly possible that he or she got stuck at this step. Therefore, the quality of self-explanations can help identify the stuck points. To estimate the quality of self-explanations for each rubric step, we adopted natural language processing methods to calculate the similarities between the self-explanation sentences and the sample sentences of the rubric steps. The higher similarity with the sample sentence, the better the quality of the self-explanations.

There exist various language models to represent text, ranging from bag-of-words such as TF-IDF (Salton and Buckley, 1988) to word-embedding such as bidirectional encoder representations from transformers (BERT; Devlin et al., 2018). In this study, we adopted Sentence BERT (SBERT; Reimers and Gurevych, 2019) and BERT Japanese pre-trained model (Suzuki, 2019) to represent the sentences for the following reasons. Firstly, BERT is a deep learning model developed on top of the transformer architecture (Vaswani et al., 2017), which outperforms almost all existing models in the field of natural language processing on various tasks (Devlin et al., 2018). It has also been widely used in the field of

educational technology (Yang et al., 2021). SBERT fine-tunes BERT in a siamese/triplet network architecture, and it has achieved a significant improvement over state-of-the-art sentence embedding methods (Reimers and Gurevych, 2019). Secondly, since students are not well trained in writing self-explanations, it was assumed that there would be no uniformity in the description, expression, and content of them. So the scoring system needed to be flexible enough to evaluate them from different angles. One of the advantages of using the BERT model is its versatility. It can be applied to a variety of tasks without having to change the structure of the model, which we thought was appropriate for this research.

Figure 3 shows an overview of the rubric based self-explanation scoring process. Firstly, both sample sentences and student self-explanations were vectorized with BERT. Secondly, we calculated the cosine similarity between them to match the rubric steps and the student explanations. Finally, the weighted average of the two sentences' similarity was calculated as the score corresponding to the rubric. Through these processes, self-explanation scores were created for each rubric, which means each student has 5 scores based on each rubric step. They were used for later analysis and model generation.

When scoring the self-explanations, we observed that a self-explanation sentence did not correspond to one specific rubric step defined by the authors. For example, one student wrote four self-explanation sentences, which contained enough knowledge to complete the quiz. Another student wrote ten sentences, but some of the steps were missing and the final answer was wrong. To solve this problem, for each rubric step, we selected two self-explanation sentences with the highest similarity scores as the representative sentences and calculated the weighted average. By calculating a weighted average score from the two sentences, it was possible to score the self-explanations based on whether it meets the required knowledge units, without being limited by the length or quantity of the sentences. The two selected representative self-explanation sentences were further used to extract other features, which will be explained in the next section.



Gold Standard			Students Answers			Labels	
Number	Rubric	Sample Sentences of Self-explanations	Student ID	2 Self-explanation sentences correspond with sample sentences (matched by cosign similarity)	Self-explanation score (The weighted average of 2 sentences)	Author 1	Author 2
Step 1	be able	coordinate ….	3832	1) Coordinate ….. 2) Calculating △ABO …..	0.831	Correct	Correct
Step 2	be able	triangle …..	3832	1) First, we tried to ….. 2) △OPQ was …..	0.672	Incorrect	Incorrect
…	…	…	…	…	…	…	…

Figure 3. An overview of the rubric based self-explanation scoring process.

3.4 Preprocessing of Pen Stroke Data and Feature Extraction

We initially assumed that Self-explanations would be a good indicator of the students' understanding of what was required to solve a quiz, however, some unexpected cases were found. For example, there were students who were very good at solving quizzes, but very bad at writing explanations. In other cases, some students were good at solving quizzes and perfect at writing their self-explanations, but they made careless mistakes in some parts. This indicates that self-explanations can help identify stuck points, but self-explanations alone cannot fully identify stuck points. Therefore, we attempted to create additional features by combining the learning behavior logs including pen stroke data with the self-explanations.

16,202 learning logs of 35 students were collected for the quiz from the learning analytics tool LAViEW, and the frequency of each learning behavior is shown in Table 3. Note that "ADD Handwriting", "UNDO Handwriting", and "REDO Handwriting" are deemed as pen stroke data.

Table 3. *Frequencies of learning behavior in LAViEW*

Learning Behavior Type	Learning Behavior	Frequency
Handwriting	ADD Handwriting	15,216
	UNDO Handwriting	233
	REDO Handwriting	3
Memo	ADD MEMO	41
	DELETE MEMO	14
	CHANGE MEMO	16

Firstly, we processed the behavioral data into longitudinal series data associated with each student's self-explanations. We thought it would be easier to follow the considering process of the students if we processed them according to the self-explanations. Secondly, we excluded the data with low frequencies and high correlations between features. Lastly, we normalized the values into the range of [0,1] for further analysis. Description of the data is given in Table 4.

Table 4. *Description of features for analysis*

Feature Name	Description
Self-explanation score	The similarity score estimated in Section 3.3, which corresponds to the 5-step rubric number.
Self-explanation length	The weighted average number of characters of two representative self-explanation sentences.
Rubric step number	Rubric step number(Ordinal Scale; 1-5).
Operation time	The weighted average of operation time associated with self-explanation sentences.
Operation order	The weighted average of operation orders associated with a self-explanation sentence.
Handwriting Frequency	The weighted average of frequency of ADD Handwriting associated with self-explanation sentences.
Add Memo Frequency	The weighted average of frequency of Typed Memo associated with self-explanation sentences.

3.5 Results of Rubric Based Self-Explanations Score and Related Features

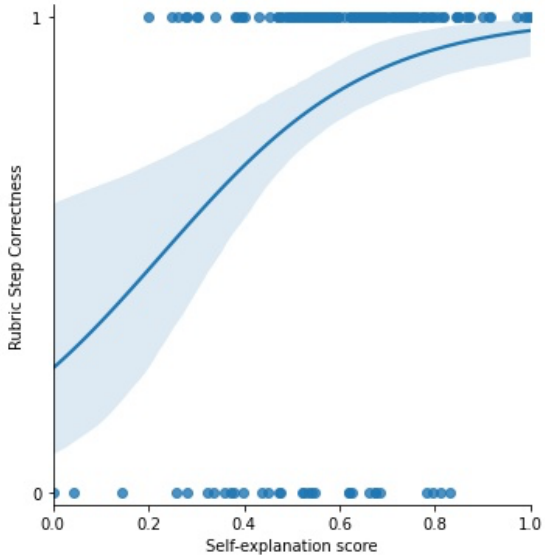
To investigate the correlations between the features and the objective variable, which is the correctness of the rubrics, we conducted Welch's t-test and Spearman correlation analysis on the feature values between the groups with different rubric correctness. Since we collected the data from 35 students and we defined 5 rubric steps, there were 175 samples of data in total. For each sample of the data, we had 7 feature values and one label indicating the correctness of the rubric. The data was divided into two groups, Correct Step (n=144) and Incorrect Step (n=31) based on the label values. Since this study had data on correct and incorrect answers to the quiz itself, this method of analysis was employed to identify the stuck points.

Table 5 shows the result of Welch's t-test, which was employed under the condition that the same variance was not assumed for the two groups. Figures 4 and 5 show the visualized result of rubric based self-explanation score between groups. Table 6 shows the results of the Spearman correlation analysis.

Table 5. *Statistics of rubric based features divided by rubric step correctness.*

	Correct Step (n=144)		*Incorrect Step (n=31)*		*Welch's t-test*	
	M	*SD*	*M*	*SD*	*DF*	*t*
Self-explanation score	0.630	0.157	0.490	0.213	**37.37**	**3.4761*****
Self-explanation length	0.292	0.213	0.236	0.174	51.22	1.576
Rubric step number	0.464	0.343	0.669	0.362	**42.45**	**-2.900*****
Operation time	0.020	0.086	0.050	0.178	33.10	-0.945
Operation order	0.396	0.193	0.569	0.331	**34.49**	**-2.812*****
Handwriting Frequency	0.142	0.151	0.156	0.170	40.85	-0.420
Add Memo Frequency	0.104	0.220	0.242	0.285	**38.08**	**-2.534****

Note. ***$p < 0.01$, **$p < 0.05$



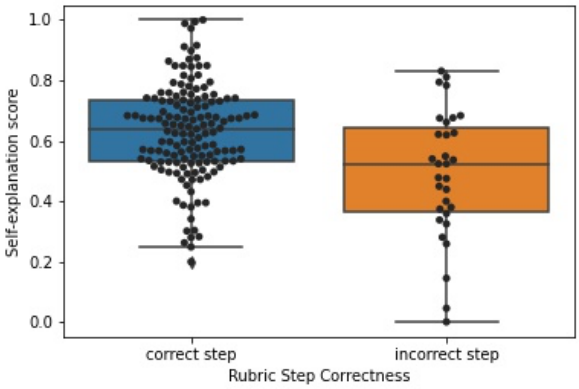


Figure5. *Box plot of the rubric step correctness and the self-explanation score.*

Figure 4. Scatter plot of the rubric step correctness and the self-explanation score.

Table 6. *Correlations between rubric step correctness and rubric based features.*

	Step Correct-ness	Self-explanation score	Self-explanation length	Rubric step number	Operation time	Operation order	Hand-writing Frequency	Add Memo Frequency
Rubric step correctness	-	**0.31*** **	0.1	**-0.22*** **	-0.11	**-0.29*** **	-0.03	**-0.22*** **
Self-explanation score		-	**0.21*** **	0.09	0.08	**-0.26*** **	-0.0	**-0.25*** **
Self-explanation length			-	0.01	0.02	**0.13***	-0.1	-0.11
Rubric step number				-	0.04	0.1	0.02	**0.14***
Operation time					-	-0.06	0.02	0.01
Operation order						-	**-0.13***	**0.17****
Handwriting Frequency							-	**0.49*** **
Add Memo Frequency								-

Note. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

3.6 Logistic Regression Analysis

We further applied logistic regression to explore the influential features relating to the rubric correctness. The multivariate logistic regression model was performed with adjustments for selected features as listed in Table 7, and contains both the adjusted OR(P=.005), and the Akaike Information Criterion (AIC) which was 139.82.

Table 7. *The results of logistic regression analysis.*

	Estimate	Std.Error	z value	Pr(>|z|)	OR	OR95%CI	
(Intercept)	1.3705	1.0799	1.2691	0.2044	3.9373	0.4742	32.6901
Self-explanation Score	4.2371	1.5314	2.7668	**0.0057*** **	69.2069	3.4401	1392.2938
Self-explanation length	1.4785	1.3140	1.1252	0.2605	4.3866	0.3339	57.6262
Rubric step number	-2.0821	0.6970	-2.9873	**0.0028*** **	0.1247	0.0318	0.4887
Operation time	-3.1635	1.7474	-1.8104	**0.0702***	0.0423	0.0014	1.2988
Operation order	-2.5505	1.0358	-2.4623	**0.0138****	0.0780	0.0102	0.5944
Handwriting Frequency	-0.0781	1.5802	-0.0494	0.9606	0.9249	0.0418	20.4746
Add Memo Frequency	-0.7704	1.1034	-0.6982	0.4851	0.4628	0.0532	4.0243

Note. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

4. Discussion

From Table 5, we observed a significant difference between the mean values of self-explanation score for the groups of Correct Step (0.630) and Incorrect Step (0.490). Therefore, we assume that self-explanations can be a useful indicator to show students' understanding towards the necessary knowledge in solving a quiz. Besides, from table 6, the order of operations and the number of steps were negatively related to the rubric step correctness, which implies that students are prone to get stuck during the later process of answering a quiz. This is understandable since mathematics quizzes get harder in later steps as the required knowledge that needs to be applied is compounded.

Table 7 shows the result of a logistic regression model with the adjusted odds ratio (OR) and 95% confidence interval (CI). There were significant differences in self-explanation score ($OR = 69.2069$, $p < 0.01$), rubric step number ($OR = 0.1247$, $p < 0.01$), operation time ($OR = 0.0423$, $p < 0.1$), and operation order($OR = 0.0138$, $p < 0.05$). In the effect size statistics, among the four features that affected the objective variable, the self-explanation score was the most influential feature. However, the confidence interval(95%) is very large, suggesting that it is not yet suitable for predicting the rubric step correctness. The main reason lies in the small sample size of self-explanations, which is sensitive to calculation errors. Enlarging the sample size is necessary to show the robustness of the model.

On the other hand, "rubric step number," "operation time," and "operation order" all had a negative impact on the objective variable. As mentioned earlier, the later the rubric occurs in the quiz the more likely mistakes will be made as the required knowledge is compounded, and this will increase depending on the difficulty level as the quizzes are often longer. These interpretable results can be used to improve for future studies.

5. Conclusion and Future Work

In this study, we attempted to create and discover highly influential features to generate a model to identify students' stuck points in a mathematics quiz using self-explanations and learning behavioral data including pen stroke data. Firstly, we created rubrics as a gold standard of stuck points, and attempted to score the self-explanations based on rubric step numbers using the Japanese Sentence Bert pre-trained word-embedding model. Next, the behavioral data was processed according to self-explanation sentences to generate a model for predicting students' stuck points. The results showed that rubric based self-explanation scores significantly differed between groups with different rubric step correctness, and the effect size was large, however the Confidence Ratio (95%) was quite large which indicates the need for further analysis and research on a larger dataset. We also found that "rubric step number", "operation time", and "operation order" all had a negative effect on the objective variable as items that occur later in the quiz tend to have more mistakes made. These results will help us to improve this line of inquiry for future studies.

There are several limitations and issues that still need to be improved. Firstly, students might not have acquired the skills required for writing good self-explanations, and this may affect the quality of their self-explanations. We plan to support students in writing self-explanations by providing hints or instructions in future work. Secondly, the sample size was small and the selected quiz is one of many units in the curriculum. Variation in the quizzes and a sufficient sample size are needed for a deeper analysis and generalization of this study. Thirdly, in the preprocessing of self-explanations, the similarity between the sample sentences and the students' were used as the score and the weighted average of the top two sentences was calculated. However, this process cannot handle cases such as when a student writes more than two sentences for each rubric step. For more accurate scoring, we need to identify how many sentences correspond to each rubric step and aggregate them before finding the similarity. Finally, although the authors made sample sentences to score the self-explanations, creating a rubric and sample sentences for each quiz at a large scale would not be practical. Therefore, an automatic generation system of rubrics and sample sentences is needed. In the future, we plan to create automatically generated rubrics from the knowledge structures contained in the knowledge map being developed at Kyoto University.

Acknowledgements

This work was partly supported by JSPS Grant-in-Aid for Scientific Research (B) 20H01722, JSPS Grant-in-Aid for Scientific Research (Exploratory) 21K19824, JSPS Grant-in-Aid for Scientific Research (S) 16H06304 and NEDO JPNP20006 and JPNP18013.

References

Andrade, H. G. (2000). Using rubrics to promote thinking and learning. Educational leadership, 57(5), 13-19.

Arter, J. and Chappuis, J. (2007). Creating and recognizing quality rubrics. Upper Saddle River, Pearson/Merrill Prentice Hall.

Bisra, K., Liu, Q., Nesbit, J. C., Salimi, F., & Winne, P. H. (2018). Inducing Self-Explanation: a Meta-Analysis. *Educational Psychology Review*, 30(3), 703-725.

Chi, M., Leeuw, N., Chiu, M., & Lavancher, C. (1994). Eliciting self-explanations improves understanding, *Cognitive Science,* 18(3), 439-477.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1), 37-46. doi:10.1177/001316446002000104.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* arXiv:1810.04805.

Flanagan, B., Ogata, H. (2018). Learning Analytics Platform in Higher Education in Japan. *Knowledge Management & E-Learning (KM&EL)*, 10(4), 469-484.

Hodds, M., Alcock, L., & Inglis, M. (2014). Self-Explanation Training Improves Proof Comprehension. *Journal for Research in Mathematics Education*, 45(1), 62-101. doi:10.5951/jresematheduc.45.1.0062

Iiyama, M., Nakatsuka, C., Morimura, Y., Hashimoto, A., Murakami, M., & Minoh, M. (2017). Detecting Answer Stuck Point Using Time Intervals of Pen Strokes. *Transactions of Japanese Society for Information and Systems in Education,* 34(2), 166-171.

Kishi, K., Miura, M. (2018). "Detecting Learners' Weak Points Utiliz-ing Time Intervals of Pen Strokes". *International Journal of Learning Technologies and Learning Environments*, 1(1), 61–77.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in cognitive sciences*, 10(10), 464-470.

McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy training for active reading and thinking. *Behavior Research Methods, Instruments, & Computers*, 36(2), 222-233.

Ochoa, X., Chiluiza, K., Méndez, G., Luzardo, G., Guamán, B., & Castells, J. (2013). Expertise estimation based on simple multimodal features. In *Proceedings of the 15th ACM on International conference on multimodal interaction* (pp. 583-590).

Reimers, N., & Gurevych, I.. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, *arXiv preprint* arXiv:1908.10084

Renkl, A. (1997). Learning from worked-out examples: A study on individual differences, *Cognitive Science*, 21(1), 1-29.

Rittle-Johnson, B. (2006). Promoting transfer: effects of self-explanation and direct instruction. *Child development,* 77(1), 1-15.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513-523.

Suzuki, M. (2019). Pretrained Japanese BERT models, *GitHub repository*, https://github.com/cl-tohoku/bert-japanese

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30, 5998-6008.

Yang, A. C. M., Chen, I. Y. L., Flanagan, B., & Ogata, H. (2021). From Human Grading to Machine Grading: Automatic Diagnosis of e-Book Text Marking Skills in Precision Education. *Educational Technology & Society*, 24 (1), 164–175.

Yoshitake, D., Flanagan, B., & Ogata, H. (2020). Supporting Group Learning Using Pen Stroke Data Analytics. *In 28th International Conference on Computers in Education Conference Proceedings* (pp. 634-639).