Koh, E., et al. (Eds.) (2026). Proceedings of the 2nd International Conference on Learning Evidence and Analytics. Asia-Pacific Society for Computers in Education.

Reproducibility Evaluation of Real-World Educational Evidence Extracted by Causal Inference

Koki Okumura [a*],Chia-Yu Hsu [b],Nobuki Sawada [a],Hiroaki Ogata [b]
[a]*Graduate School of Informatics, Kyoto University*
[b]*Academic Center for Computer and Media Studies, Kyoto University*
*okumura.kouki.27m@st.kyoto-u.ac.jp

Abstract: For enhancing the reliability of Evidence-Based Education (EBE), the evaluation of reproducibility of causal relationships extracted by causal inference from Real-World educational Evidence (RWeE) is crucial. To address this challenge, this study attempted to quantitatively evaluate the reproducibility of educational evidence derived from causal inference by systematically applying the Correspondence Test (CT). Specifically, features of learning behavior were generated from summer vacation learning log data of junior high school students in grades 1 to 3 (6 datasets, n=109-117), and causal relationships associated with changes in deviation scores were extracted using LiNGAM-MMI, which can account for unobserved confounding factors. The results of the CT reproducibility evaluation revealed that factors related to "maintaining a learning rhythm" exhibited high reproducibility with a practically significant effect size, leading to the successful extraction of robust RWeE. This study establishes a new method that enables the quality assurance of RWeE by integrating causal discovery and CT, and demonstrates that these reproducible findings provide the foundation for designing high-quality educational interventions.

Keywords: Real-World Educational Evidence, Learning Analytics, Causal Inference, Reproducibility Evaluation, Correspondence Test

1. Introduction

Evidence-Based Education (EBE) is recognized as an indispensable approach for improving the quality of education (Slavin, 2002). Extracting reliable evidence is essential for effective evidence-based feedback. Real-World educational Evidence (RWeE) plays a role in supplementing the evidence extracted by randomized controlled trials (RCTs) (Sherman et al., 2016), which are difficult to implement in educational settings. RWeE has been extracted through causal inference on observational data (Pearl, 2009; Spirtes et al., 2000). While the application of causal inference is advancing, systematic verification of the reproducibility of these causal relationships is lacking. Causal inference alone does not guarantee the reliability of RWeE, and reproducibility evaluation is essential to enhance reliability. Such reproducibility crisis is a problem across the scientific community (Open Science Collaboration, 2015), and it has been pointed out that it is more serious in educational research than in other fields (Makel & Plucker, 2014).

Specifically, this study focuses on the issue that a method for evaluating the reproducibility of causally inferred evidence extracted in an exploratory manner using a unified standard has not been established. In recent years, RWeE has also been extracted by causal discovery, a technique of causal inference. Causal discovery is a technique for estimating the causal structure itself from observational data and statistical assumptions, and its application to educational data is also progressing (Okumura et al., 2026). A strong candidate for the strict reproducibility verification of causally inferred evidence is the Correspondence Test (CT) proposed by Steiner & Wong (2018). CT has been shown to be effective in the educational field by Cohen et al. (2024) and others, but there has been no systematic study applying CT to results from causal discovery.

This study integrates the discovery power of causal discovery and the verification power of CT, and systematically applies CT to a wide range of RWeE for the first time. Through comprehensive causal discovery using LiNGAM-MMI (Suzuki & Yang, 2024) and structural-level reproducibility evaluation using CT, we extract robust causal relationships that are reproducible and free from noise. The objective of this study is to answer the following two Research Questions.

- RQ1: Can highly reproducible RWeE be extracted by applying CT to causal relationships extracted through causal discovery?
- RQ2: What kind of reproducible causal relationships exist between behavior and academic performance change, and how do they vary by context (grade/academic year)?

2. Methods

The experiment follows the procedure below (Figure 1), which is designed to systematically address the unique challenges of extracting reliable evidence from exploratory, high-dimensional educational data:

1. Causal pairs are extracted for each dataset using LiNGAM-MMI with all features.
2. Common causal pairs across datasets are identified.
3. The Correspondence Test (CT) is performed with dataset pair comparisons.
4. Causal pairs judged as Equivalent with δ=0.2 and δ=1.0, respectively, are identified and interpreted.

This four-step procedure addresses the specific challenges of this study as follows. Step 1 leverages LiNGAM-MMI's capacity to handle all 111 features simultaneously without narrowing down variables based on a priori hypotheses, enabling discovery of unexpected causal relationships. Step 2 filters for structural stability by retaining only causal pairs that appear consistently across independent datasets (different grades and academic years), thereby eliminating noise-driven associations. Step 3 applies CT's dual-test framework (Difference Test and Equivalence Test) to statistically quantify the reproducibility of each causal pair across dataset pairs. Step 4 identifies the most robust evidence by distinguishing strictly equivalent relationships (δ=0.2, fundamental laws) from broadly equivalent relationships (δ=1.0, general tendencies with contextual variation).

Furthermore, for the evaluation of practical significance, we focus on causal pairs with an absolute effect size greater than 0.1.

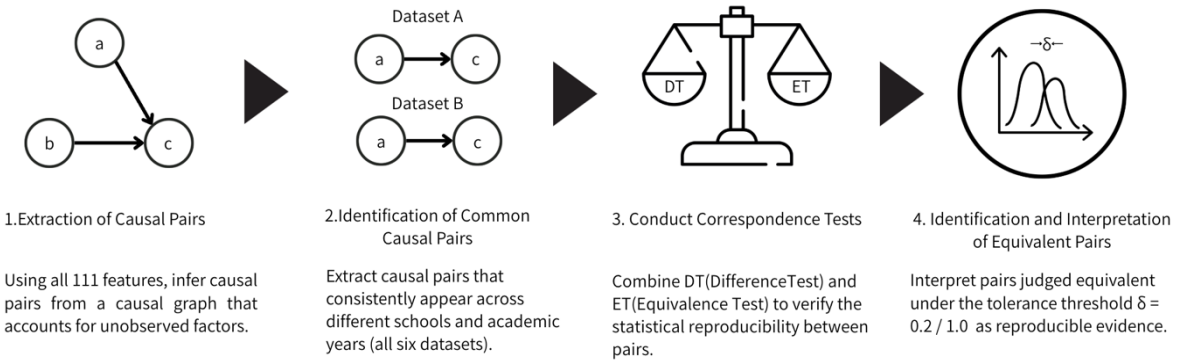

Figure 1: Four Steps for Reproducibility Evaluation

2.1 *Datasets and Features*

This study uses summer vacation learning log data of students in grades 1 to 3 (Academic Years 2023 and 2024) regarding their engagement with summer homework at a Japanese junior high school. A total of six datasets were used: J1_2023 (n=117), J1_2024 (n=112), J2_2023 (n=109), J2_2024 (n=115), J3_2023 (n=109), and J3_2024 (n=114). In addition to this, pre- and post-summer vacation test data are also utilized.

As explanatory variables, a total of 111 features representing learning behavior were generated from the learning analysis platform LEAF (Ogata et al., 2015). LEAF is a platform that records and analyzes learner behavior logs when using digital textbooks. These 111 features include 11 basic learning metrics, 99 advanced learning metrics generated using Sonnet 4.5, and the difference in deviation scores before and after the summer vacation, indicated by deviation_diff. The followings provide examples of the extracted metrics.

- Basic learning metrics: jump_count (total page transitions), yellow_marker_count (number of yellow markers), red_marker_count (number of red markers).
- Advanced learning metrics: max_gap_days (maximum learning interval in days), summer_time_concentration (summer concentration: standard deviation of daily event count during summer vacation), events_per_day (events per day: total events / total learning days), learning_velocity (learning speed: unique page visits / total learning days), wandering_time (wandering time: total time returned to the same page within 1 minute).

In this analysis, no feature selection based on specific hypotheses was performed, and all 111 generated features were used as input for causal discovery. The intention here is to discover unexpected causal relationships that are difficult to predict from existing theories, but it also increases the computational cost and the risk of false positives for high-dimensional data. Therefore, as mentioned in the next section, the computational efficiency and robustness of LiNGAM-MMI are indispensable elements for achieving this comprehensive search.

2.2 *Causal Discovery Method*

For estimating causal relationships, we used a Causal Discovery approach, which is a type of Causal Inference that exploratorily estimates the causal structure itself from observational data. Specifically, we adopted LiNGAM-MMI (Suzuki & Yang, 2024), which is based on LiNGAM (Shimizu et al., 2006; Shimizu et al., 2011) that estimates causal structure by assuming Non-Gaussianity, and can uniquely identify the causal direction even in the presence of unmeasured confounders.

The reason for adopting LiNGAM-MMI is its ability to efficiently estimate causal structures even with high-dimensional data while minimizing the influence of unobserved confounding factors (such as learning motivation and home environment) that are unavoidable in educational data. The computational efficiency and robustness of LiNGAM-MMI are extremely important, especially since we did not narrow down the features in this study. LiNGAM-MMI is applied to all 111 features and the objective variable (deviation_diff) to estimate the causal graph. Causal pairs that influence deviation_diff are extracted from the estimated causal graph.

2.3 *Implementation of the Correspondence Test*

The Correspondence Test (CT) is performed on the extracted causal pairs. The Difference Test (DT) tests the null hypothesis H_0: $\beta\_diff = 0$, and the Equivalence Test (ET) tests the composite null hypothesis H_0: $|\beta\_diff| \geqq \delta$ using TOST (Two One-Sided Tests). Based on Steiner & Wong (2018), the reproducibility is determined in the following four ways by combining the test results of DT and ET:

- Equivalent (Reproducible): DT is non-significant and ET is significant. There is no statistical difference, and the difference is within the acceptable range. This indicates the strongest reproducibility.
- Trivial Difference (Minor Difference): DT is significant and ET is significant. There is a statistical difference, but the difference is within the acceptable range δ. It may be considered a successful replication if δ is sufficiently small.
- Indeterminate (Undetermined): DT is non-significant and ET is also non-significant. It cannot be said whether there is a difference or not (e.g., due to lack of power).
- Different (Not Reproducible): DT is significant and ET is non-significant. There is a statistical difference, and the difference is not within the acceptable range.

Following examples of application to educational data (Cohen et al., 2024), two thresholds were set for δ: δ=0.2 (strict) and δ=1.0 (lenient). The choice of the value of δ has been pointed out as a crucial design challenge that affects the conclusion of reproducibility

evaluation by CT (Yeaton & Velasquez, 2022). By setting these two thresholds, it is possible to distinguish between "fundamental rules that are strictly reproducible (δ=0.2)" and "applied rules that include contextual variation but are reproducible as a general tendency (δ=1.0)". In this study, only the Equivalent judgment is considered reproducible.

It is worth noting that equivalence tests are generally sensitive to sample size; with samples of n=109–117, the statistical power to detect equivalence may be limited, particularly for small effect sizes. The Indeterminate outcomes observed in some pairs are likely attributable in part to this constraint. Future studies with larger samples or pooled datasets could further strengthen the power of the CT-based evaluation.

3. Results

This section reports the factors confirmed to have a statistically significant and robust impact on the improvement of academic performance (deviation score) as a result of the causal discovery analysis. All the following factors satisfy the stability criteria (Equivalence rate = 100%, Sign consistency rate = 100%) and the effect size criterion (|β| >= 0.1) and are limited to causal pairs that affect deviation_diff (deviation score difference). Table 1 presents all causal pairs satisfying these selection criteria; no pairs meeting the criteria were omitted. Table 1 shows the combination of datasets for which reproducibility was confirmed, the name of the causal variable for deviation_diff and the sign of its influence (Positive/Negative), and the inferred internal state and focus. For example, the combination of "J1_2024 and J2_2023" confirmed the reproducibility that max_gap_days (maximum learning interval in days) has a negative influence on deviation_diff, and the inferred internal state is "Disruption of Learning Rhythm".

Only learning_velocity was not extracted with δ=0.2 but was extracted with δ=1.0. Others were extracted even with δ=0.2.

Table 1. *Key Factors Affecting Academic Performance Improvement and Inferred Internal States*

	Key Factors and Influence	Inferred Internal State and Focus
J1_2024 J2_2023	max_gap_days (Negative)	Disruption of Learning Rhythm (Negative)
J1_2023 J1_2024	summer_time_concentration, events_per_day, daily_events_grad (Negative), learning_velocity (Positive)	Lack of Planning (Negative), Shallow Viewing (Negative), Efficient Exploration (Positive). Focus is on Consistency of Learning Quantity and Frequency.
J2_2023 J2_2024	action_density_grad (Negative)	Lack of Planning (Negative)
J3_2023 J3_2024	yellow_marker_count (Positive), jump_count (Negative)	Active Selection (Positive), Skimming (Negative). Focus is on Quality of Information Selection and Processing.
J2_2023 J3_2024	red_marker_count (Negative)	Feeling of Knowing (Metacognitive Illusion) (Negative). Excessive use of simple markers hinders deep comprehension.
J1_2024 J3_2024	wandering_time (Positive), action_count_consistency (Positive)	Deep Processing (Positive), Self-Regulated Learning (Positive).

4. General Discussion

This study established a method to quantitatively assure the reliability of exploratorily extracted causal structures by using the equivalent judgment of CT as a strict criterion. We confirmed that 11 reproducible and practically effective RWeE s are indeed extracted by this method

(RQ1). By interpreting the guaranteed reproducible causal relationships as learner's internal state, we elucidated the mechanism by which factors affecting academic performance qualitatively shift from "consistency of learning quantity and frequency" to "quality of information selection and processing" depending on the developmental stage (RQ2). Furthermore, based on the guaranteed reproducible RWeE and the inferred internal states, we can specifically propose a design guideline for high-quality individualized adaptive education that shifts the focus of intervention from "increase/decrease of behavior" to the "quality of internal state." Table 2 shows the internal states, corresponding strategies, and feedback examples based on those strategies. For instance, the inferred internal state in the first row of Table 1 was "Disruption of Learning Rhythm." A strategy corresponding to "Disruption of Learning Rhythm" is "Reinforce habit formation." Following this strategy, an example of feedback when learning is continuous is, "This is your Xth consecutive day of learning! You are keeping up the consistency."

Table 2. Adaptive Feedback Tailored to Internal States

Internal State	Strategy	Feedback Example
Disruption of Learning Rhythm/Lack of Continuity	Reinforce habit formation	"This is your Xth consecutive day of learning! You are keeping up the consistency."
Shallow Viewing	Suppress unproductive hyper-activity and guide towards positive efficient exploration	"Did you find the key points you need to understand on this page? Let's move on to the next page to proceed with your learning efficiently."
Feeling of Knowing	Impose retrieval effort based on the possibility that simple highlighting behavior may inhibit deep knowledge retention.	"To check your knowledge retention, try summarizing the content you just marked in three keywords without looking at anything."
Deep Processing	Affirm this behavior as persistence without discouraging it, maintaining and strengthening motivation.	"You're tackling difficult problems with persistence. Your persistence will lead to results. Keep up the good work!."

5. Conclusion and Future Work

This study established a new method integrating causal discovery and CT to quantitatively assure the reproducibility of Real-World educational Evidence. This method successfully extracted robust causal relationships with confirmed reproducibility, and by inferring learners' internal states from these findings, provided insights for realizing high-quality adaptive individualized education. To leverage the knowledge gained from this study in educational practice and establish its effectiveness, several issues must be continuously addressed.

First, to evaluate the extracted causal relationships, implementation experiments of feedback strategies for the next summer homework must be conducted to verify their internal validity. Furthermore, as the datasets used for analysis are limited to summer vacation learning logs, RWeE's external validity must be evaluated in the future using data from different learning contexts, including regular semester classroom settings. The findings may also vary across different cultural contexts, and cross-cultural validation is an important direction for future research.

Second, to enhance the validity of the internal states inferred in this study, the interpretation must be supplemented and verified through evidence found in cognitive science and educational psychology, qualitative observations by teachers, and qualitative research such as think-aloud protocols.

Finally, future research is suggested to further implement the proposed adaptive feedback strategies on an educational platform and empirically evaluate the intervention effects using experimental methods such as Randomized Encouragement Designs.

Acknowledgements

This work was supported by CSTI SIP Grant Number JPJ012347 and JSPS KAKENHI Grant Number 23H00505.

References

Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. Psychological Methods, 21(1), 1-12.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). Introduction to meta-analysis. John Wiley & Sons.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative theory. Psychological Bulletin, 132(3), 354.

Cohen, J., Wong, V. C., Krishnamachari, A., & Erickson, S. (2024). Experimental evidence on the robustness of coaching supports in teacher education. Educational Researcher, 53(1), 19-35.

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. PLOS ONE, 11(2), e0149794.

Hedges, L. V., & Schauer, J. M. (2019). Statistical analyses for studying replication: Meta-analytic perspectives. Psychological Methods, 24(5), 557-570.

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. Social Psychological and Personality Science, 8(4), 355-362.

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. Educational Researcher, 43(6), 304-316.

Okumura, K., Nishioka, K., Koike, K., Horikoshi, I., & Ogata, H. (2026). Causal discovery for automated real-world educational evidence extraction. Research and Practice in Technology Enhanced Learning, 21, 020.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349(6251), aac4716.

Pearl, J. (2009). Causality: Models, reasoning, and inference (2nd ed.). Cambridge University Press.

Peters, J., Mooij, J. M., Janzing, D., & Schölkopf, B. (2014). Causal discovery with continuous additive noise models. Journal of Machine Learning Research, 15(1), 2009-2053.

Sherman, R. E., Anderson, S. A., Dal Pan, G. J., Gray, G. W., Gross, T., Hunter, N. L., ... & Woodcock, J. (2016). Real-world evidence—What is it and what can it tell us? New England Journal of Medicine, 375(23), 2293-2297.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., & Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. Journal of Machine Learning Research, 7, 2003-2030.

Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., ... & Bollen, K. (2011). DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. Journal of Machine Learning Research, 12, 1225-1248.

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. Educational Researcher, 31(7), 15-21.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). Causation, prediction, and search (2nd ed.). MIT Press.

Steiner, P. M., & Wong, V. C. (2018). Assessing correspondence between experimental and nonexperimental estimates in within-study comparisons. Evaluation Review, 42(2), 214-247.

Suzuki, J., & Yang, T.-L. (2024). Generalization of LiNGAM that allows confounding. arXiv preprint arXiv:2401.16661v4. [Preprint, not peer-reviewed]

Takii, K., Liang, C., & Ogata, H. (2025). Defining the Scope of Learning Analytics: An Axiomatic Approach for Analytic Practice and Measurable Learning Phenomena. arXiv preprint arXiv:2512.10081.

Yeaton, W. H., & Velasquez, G. (2022). Using correspondence tests to assess replicability of Open Science Collaboration results: Inferences from a SMART design-based, meta-analytic approach. Journal of Methods and Measurement in the Social Sciences, 13(2), 41-69.