Okumura et al. *Research and Practice in Technology Enhanced Learning* (2026) 21:20
https://doi.org/10.58459/rptel.2026.21020

Research and Practice in Technology Enhanced Learning

RESEARCH Free and Open Access

Causal discovery for automated real-world educational evidence extraction

Koki Okumura [1]*, Kento Nishioka [1], Kento Koike [2], Izumi Horikoshi [3] and Hiroaki Ogata [3]

*Correspondence:
okumura.kouki.27m@st.kyoto-u.ac.jp
Graduate School of Informatics, Kyoto University, Japan
Full list of author information is available at the end of the article

Abstract

There is increasing demand to shift from intuition- and experience-based practices to evidence-based education. However, extracting meaningful evidence from real-world educational data poses significant challenges. Traditional approaches to evidence generation, such as randomized controlled trials and systematic reviews, face limitations in both the medical and educational domains due to high costs and ethical constraints. In response, the concept of real-world evidence has emerged as a promising alternative, particularly in medicine and, more recently, in education. Although this approach may be less robust than traditional methods, it offers the potential to uncover broad and practical insights from naturally occurring data. This study explores the use of deep learning for causal discovery in real-world educational data. Specifically, we apply Structural Agnostic Modeling, a method previously validated in biological datasets, to identify underlying causal relationships. In Study 1, we compare this data-driven approach to a traditional hypothesis-driven method. The results demonstrate that this technique can generate both interpretable and novel causal hypotheses, although it occasionally produces plausible relationships in the reverse direction. To address this limitation, we propose an enhanced model, SAM+, in Study 2. Our findings indicate that SAM+ effectively mitigates the identified shortcomings. This research contributes a new methodology for leveraging large-scale educational data and opens new possibilities for advancing evidence-based education.

Keywords: Evidence-based education, Real-world evidence, SAM, Causal discovery, Causal analysis

Introduction

Evidence-based education (Davies, 1999) is expected to provide education not based on intuition or experience. *Evidence* has generally been extracted from systematic reviews or RCTs (Randomized Controlled Trials). Although these methods provide higher-level evidence, they have difficulties collecting large amounts of evidence because they are often

costly, unethical, or not feasible (Slade & Prinsloo, 2013). Therefore, there has been proposed a concept of real-world evidence in the medical field (Mahajan, 2015), and several attempts to apply it to the educational field as well. Real-world evidence (RWE) is evidence extracted from various real-world data (RWD), and even though it is less reliable than systematic reviews and RCTs, it is expected to be able to collect a large amount of evidence.

Our previous method attempted to compare a control group with a corresponding target group to verify the effect of an intervention (Nakanishi, 2021). While this method had the advantage that the effect of the intervention could be verified without conducting experiments with setting target and control groups. However, it requires manually setting the classes to be compared, and extracting a lot of evidence was difficult.

To address this issue, we developed a method to automatically search for classes with similar contextual information, such as grade level, as a control group (Takami et al., 2022). This is called Automated Similar Lecture Search. This can make hypotheses conceived by humans tested without experiment as if they had been subjected to a comparative experiment. However, since this method is hypothesis-driven, it is first necessary to formulate hypotheses for comparison and validation. However, it takes much time to formulate many hypotheses manually.

If we can automatically generate hypotheses from data, we can extract more evidence. How can we do this? The technology that answers this question is causal discovery. As Kalainathan points out (Kalainathan et al., 2022), observational causal discovery, which is causal discovery for observational data, is attracting attention from the machine learning community, and it is being applied in various fields, including economics and bioinformatics informatics, and other fields. There are also a few examples in education. However, until now, these have been based on simple algorithms, and it is difficult to automatically search through large-scale, complex real-world educational data. We apply observational causal discovery to real-world educational data and attempt to automatically extract evidence.

Therefore, in Study 1, we address the following two research questions.

- RQ1: What do we get from the observational causal discovery method adapted to real-world educational data?
- RQ2: How different are the results between hypothesis-driven and data-driven methods?

For RQ1, we apply the causal discovery algorithm “SAM” to RWD. It has already produced good results in biological RWD. In RQ2, we will compare Automated Similar Lecture Search and SAM-based methods and identify what features and differences exist between them.

We find that SAM has the advantage of generating a large number of unexpected various causal hypotheses. However, it also has the disadvantage of generating causal hypotheses that are not feasible but are interpretable and feasible in the opposite direction. This disadvantage should be improved. Therefore, in Study 2, we propose the following research question.

- RQ3: Can the shortcomings of observational causal discovery be improved?

For RQ3, we created SAM+, an improved version of SAM, which updates SAM to allow for the input of a priori unfeasible causal relationships.

Through these research questions, we aim to show a new path for automatic extraction of more effective RWE from RWD.

Related works

Evidence-based education

Definition of evidence

In this section, we explore the diverse interpretations of what constitutes evidence. Generally, Randomized Controlled Trials (RCTs) are regarded as the most reliable form of evidence (Greenhalgh, 2004). However, there is also a perspective that values the opinions and experiences of experts as valid evidence (Buysse & Wesley, 2006). As defined by Sackett (Sackett et al., 1996), evidence-based medicine integrates clinical expertise with systematic research. This definition, subject to debate since the 1990s, has yet to reach a consensus.

Research-based evidence often takes precedence, but its evaluation can change with new studies, highlighting the limitations of RCTs. Consequently, it is said that a major challenge moving forward is to avoid adherence to a specific type of evidence and instead combine different sources of evidence to test which framework is most effective (Rycroft-Malone et al., 2004).

Despite the various interpretations and conflicts of opinion, there is a consensus that evidence, regardless of how it is interpreted, must be independently observed and verified (Davies & Nutley, 2000). In essence, evidence should be based on data obtained through observation and experimentation, and it must involve the validation of hypotheses and systematic compilation of research findings.

Evidence hierarchy

The evidence hierarchy is a core principle of evidence-based practice that has been used in a variety of forms since 1979, beginning with the “Canadian Task Force” periodic health examination studies (Evans, 2003). The exact form and rank of these hierarchical research

Table 1 Classification of levels of evidence

Levels of evidence	Classification
I	Systematic review/meta-analysis
II	One or more randomized controlled trials (RCTs)
III	Non-randomized controlled trial
IV	Quasi-experimental study
V	Descriptive study

designs have not been determined, and a variety of approaches have been taken. In this study, we refer to a five-level classification in health care (Cook et al., 1995) and Table 1 shows the classification.

Level I evidence refers to results from systematic reviews and meta-analyses, while Level II evidence refers to results from randomized controlled trials. Level I and II are the subject of conventional discussions of general evidence (McMillan & Schumacher, 2010). The reason is that RCTs are considered to minimize the risk of confounding and provide the most reliable evidence when evaluating the effect of an intervention (Evans, 2003). However, in this study, we will not conduct an RCTs, but will include Level III, IV, and V nonrandomized controlled trials, cohort studies, and case studies. Although the evidence for these is lower than for systematic reviews and RCTs, if properly designed, they could approach RCT levels (Burns & Grove, 2010; Sherman et al., 2016). In addition, if enough cohort studies are accumulated, they could lead to systematic reviews and improve the overall level of evidence.

Evidence in education

In line with the trend towards evidence-based medicine, there is now a demand for evidence-based education (Davies, 1999) in the field of education too.

In the realm of evidence-based education, systems that handle high-quality research such as systematic reviews and RCTs include the What Works Clearinghouse (WWC) operated by the U.S. Department of Education. This platform serves as an educational research database, summarizing and providing evidence from studies that meet certain criteria, including conducting RCTs. Another example is the Education Endowment Foundation (EEF), supported by the UK's Department of Education. Established in 2011 as an independent charity, the EEF assists teachers and school leaders by providing resources based on evidence designed to improve practices and foster learning.

Both systems summarize evidence with strong causal relationships and compile them into databases. However, they do not handle lower level evidence, limiting the scope of evidence that can be collected. Moreover, these systems primarily focus on sharing evidence, and the registration of evidence is done manually, presenting a challenge.

Real-World Evidence in education

What is Real-World Evidence

Real-World Evidence (RWE) refers to clinical evidence derived from the analysis of Real-World Data (RWD), which can be generated through various research designs. RWD encompasses medical information obtained from sources beyond the clinical research environment, such as Electronic Health Records (EHRs), disease registries, personal devices, and medical applications (Sherman et al., 2016). RWE has the potential to mimic methodologies like Randomized Controlled Trials (RCTs), observational studies, pragmatic trials, and large simple trials, given the right conditions and data. Moreover, it is hoped to yield insights from innovative research designs and analyses.

The advantages of RWE include its ability to be collected at a much lower cost compared to RCTs and the use of vast samples that allow for statistical generalization (de Lusignan et al., 2015). However, challenges include uncertainties in methodology and evaluation techniques, and the inherent disorder and incompleteness of RWD, necessitating sophisticated statistical methods for accurate insights.

While the use of RWE in the medical field has been increasing (Oyinlola et al., 2016), its research in the field of education has been limited. With the growing use of ICT in educational settings, an increase in available educational data is anticipated. The application of this data to generate RWE could lead to its transformation into big data, enabling the collection of substantial evidence in education. This, in turn, could facilitate the application of evidence in educational settings, driving the realization of evidence-based education and contributing to the improvement of educational quality.

Real-World Evidence in education

One system that handles real-world educational evidence is the Learning Evidence and Analytics Framework (LEAF) (Figure 1), proposed by Ogata and others. LEAF is a Learning Analytics (LA) platform designed to support the processes of data collection, analysis, planning interventions, monitoring, and reflection, with the goal of discovering and accumulating real-world educational evidence from learning log data. Kuromiya (Kuromiya, 2023) proposed and integrated LEAF into a platform that manually extracts and accumulates evidence of effective learning and teaching methods using data stored in LEAF. Nakanishi (Nakanishi, 2021) utilized this platform to improve teaching practices and demonstrated its effectiveness. However, as Nakanishi points out, the manual extraction of evidence makes it challenging to continually gather evidence on a routine basis. Therefore, the authors proposed a method to automatically find control groups for certain interventions (Takami et al., 2022), but the creation of hypotheses still requires

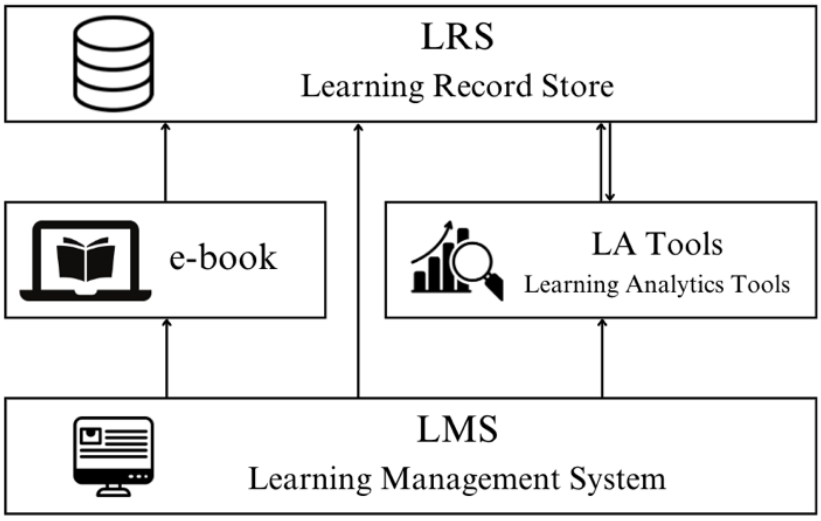

Fig. 1 LEAF system

Table 2 Position of this research

	RCT	RWE extracted by manual	RWE extracted automatically
	WWC/EEF	*Kuromiya/Nakanishi*	*This Research*
Easiness of collection	Low	Middle	High
Evidence level	High	Low	Low but can be promoted to High

human input, highlighting the ongoing challenge of automation. Therefore the problem that this research aims to solve can be positioned as shown in Table 2.

Causal discovery in education

Causal discovery

How can we automatically generate hypotheses? The technology that answers this question is called causal discovery.

Causal discovery is a task that calculates the presence or absence of causal relationships and their directionality from data. It searches for causal relationships between variables in the data. For example, if there are three variables in the data - learning time, sleep time and grades - it can identify whether there is a causal relationship between all three variables, and if so, which direction the causal relationship is in (Figure 2).

There are four broad approaches to causal discovery.

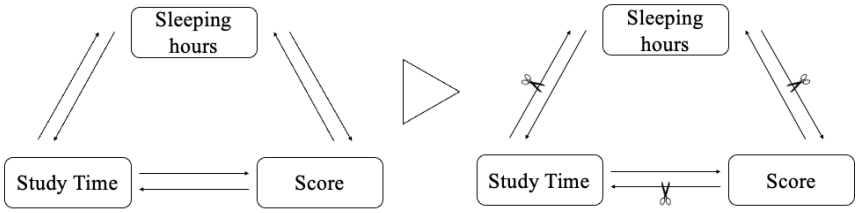

Fig. 2 Image of causal discovery

The first approach is a constraint-based method. This is a method for discovering causal graphs that utilizes conditional independence tests. The basic algorithm for this method is PC (Peter-Clark), and an extended algorithm is FCI (Fast Causal Inference). FCI can consider the existence of hidden variables based on observed data. This approach is simple and powerful, but it is often not possible to specify which direction it is.

The second approach is a score-based method. A score such as BIC or AIC is set, and the graph is evaluated based on the score to search for the optimal causal graph. The basic algorithm is GES. It adds and deletes edges to achieve the best score. GOLEM also uses gradient-based optimization to search for the causal graph with the best score. This approach can identify the direction of causality, but it tends to suffer from the curse of dimensionality when there are many variables.

The third approach is a hybrid of the two approaches above. It narrows down the possible causal graphs based on constraints, and then selects the optimal model based on scores.

The fourth approach is a method that uses asymmetry or traces of causality. This is a method for identifying causal graphs that uses asymmetry and traces of causality based on the process of generating observational data. LiNGAM assumes that the data follows a linear model and uses non-Gaussian noise to determine the direction of causality. SAM constructs causal graphs using conditional independence and distribution asymmetry, and in particular uses neural networks to estimate the distribution of each variable. It is also powerful for non-linear models and high-dimensional data, and can handle complex causal relationships. We will explain the details later.

Causal discovery in education

As Kitto et al. (2023) point out, the history of causal discovery for educational data began before the field of LA. In 2009, Brokenshire used FCI to conduct causal discovery on the theme of SRL (Brokenshire & Kumar, 2009). He compared the causal models output by theoretical causal models and causal discovery. However, at the time, it did not become a mainstream approach. This was due to the difficulty of data collection and implementation. Five years later, in 2014, Fancsali uses PC and FCI to conduct causal discovery on non-

cognitive behaviors and phenomena that affect student learning outcomes (Fancsali, 2014). Nine years later, in 2023, Ouaadi and Ibourk use PC, GES, LiNGAM and GOLEM to conduct causal discovery on tasks that identify the characteristics of students with low grades (Ouaadi & Ibourk, 2023). Also in the same year, Smith used PC and GES to perform causal discovery on student performance data and identified the factors that affect performance (Smith, 2023).

In summary, the causal discovery algorithms that have been used for educational data so far have not been able to apply algorithms that can handle both complex structures and hidden variables. Real-world educational data is complex and involves many hidden variables. Therefore, we propose the use of the algorithm, SAM.

Structural Agnostic Modeling (SAM)

SAM (Kalainathan et al., 2022) utilizes a type of deep learning called Generative Adversarial Network (GAN). GAN is known for its ability to generate fake objects that are close to the real thing by having two networks, Generator and Discriminator, compete. SAM is applying this GAN concept to causal analysis, which is used to find causal relationships among real-world educational data (Figure 3). The SAM Generator focuses on one of the n variables, sets the other n-1 variables to true data, and generates the specified variable from those variables according to the input noise. The Discriminator determines whether 'one variable is generated data and the other n-1 variables are true data' or 'all n variables are true data'. This is repeated n times with different target variables (actually, matrix operations are performed). In doing so, the method of generating each variable is stored in a matrix called Structural Gates. This shows the causal graph (Figure 4).

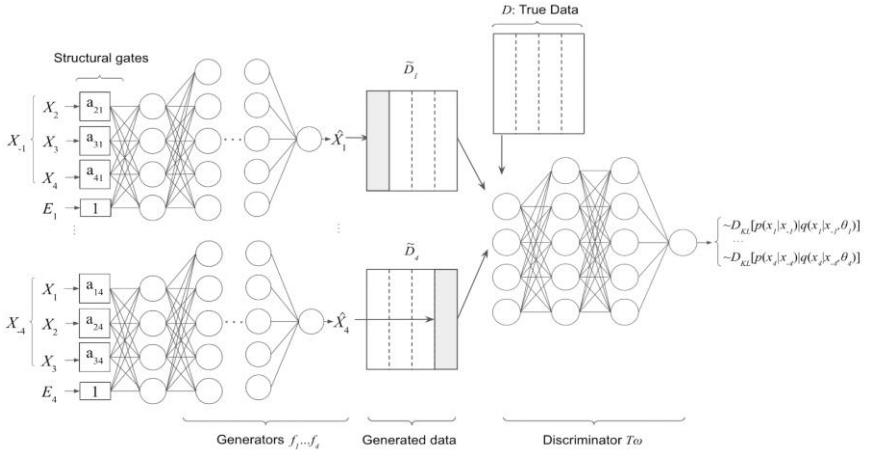

Fig. 3 Algorithm of Structural Agnostic Modeling (SAM)
(Adopted from Kalainathan et al. (2022))

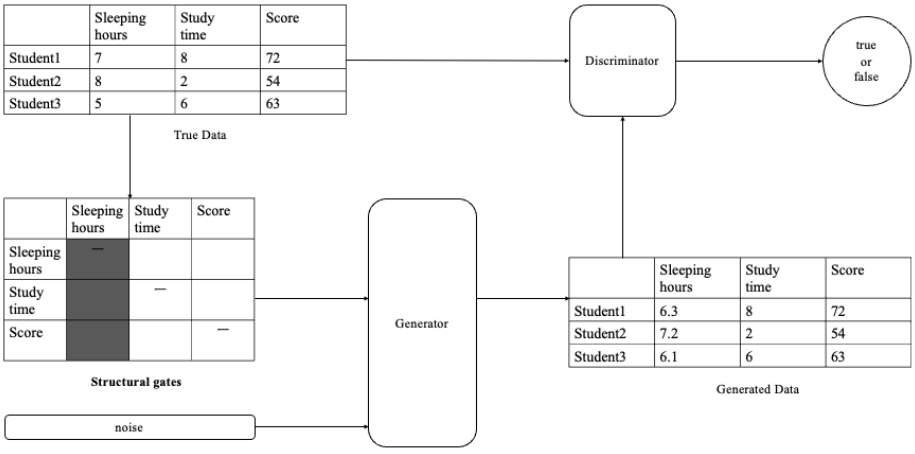

Fig. 4 Overview of SAM

Specifically, SAM quantifies the likelihood of causal relationships between each indicator and examines whether the causal relationships. SAM uses deep learning to enable it to learn complex non-linear relationships and multi-dimensional causal structures. It also uses adversarial learning to achieve highly accurate causal discovery. It is also possible to handle hidden variables, as it introduces unobserved variables as noise terms and calculates the difference between the probability distributions of the generated data and the observed data to minimize the difference. The hyperparameters are the same as those placed in the original paper. The differences between SAM and other algorithms are shown in Table 3.

Table 3 Position of SAM

Feature/Algorithm	PC	FCI	GES	LiNGAM	GOLEM	SAM
Approach	Constraint-based	Constraint-based	Score-based	Uses non-Gaussian	Score-based	Uses traces of causality
Applicable Data	Linear & Non-linear	Linear & Non-linear, Hidden Variables	Mainly linear	Mainly linear, non-Gaussian distribution	Linear & Non-linear, mainly continuous data	Linear & Non-linear, Non-Gaussian distribution, Complex structures
Handling Complex Structures	Low	Low	Low	Medium	High	High (utilizes deep learning)
Handling Hidden Variables	Not possible	Possible	Not possible	Not possible	Not possible	Possible (considers latent variables)
Model Interpretability	High	High	High	Medium	Medium	Medium
Computational Efficiency	High	Medium	High	Low	Medium	Low (requires substantial computational resources)

Method

In this paper, we address three primary research questions (RQs) aimed at exploring the effectiveness of observational causal discovery methods in the context of real-world educational data.

Data collection

LEAF, which stands for Learning Evidence Analytics Framework (Ogata et al., 2018), is an innovative technology design framework that supports evidence-based education. This framework is particularly notable for its comprehensive approach to collecting and utilizing educational data to enhance learning experiences and outcomes. Here are some key aspects of LEAF:

LEAF consists of several integral components. BookRoll is a digital textbook platform that allows the collection of detailed reading logs. It captures data such as time spent on each page, annotations made by students, and more. Log Palette is a system for visualizing and analyzing learning logs. It provides educators and researchers with tools to understand student learning behaviors and patterns. Learning Record Store (LRS) is responsible for storing and managing learning records in a standardized format, facilitating easy access and analysis of educational data.

LEAF has been implemented in various educational settings, demonstrating its versatility and effectiveness. Several schools have been using LEAF for about five years, accumulating rich datasets that are invaluable for research and development in educational technology. One of the strengths of LEAF is its applicability across a wide range of subjects. This versatility ensures that it can be integrated into various curricula, enhancing the learning experience in different academic fields. The data collected through LEAF is not only beneficial for improving educational practices but also serves as a rich resource for educational data analysis contests. These contests encourage the exploration and development of new methods and techniques in data science, specifically tailored to education. Schools using LEAF for an extended period have accumulated several years of real-world data. This data provides a deep insight into long-term educational trends and student learning behaviors, offering a valuable resource for educators and researchers to refine and develop more effective educational strategies.

In summary, LEAF is a multifaceted framework that integrates various technologies to gather and analyze educational data. Its implementation in diverse subjects and schools, along with its role in facilitating educational data analysis contests, highlights its potential in shaping future educational practices and policies.

In Japan, one tablet or another device per student has been distributed with the GIGA School project. Within this project, a Learning Analytics platform named LEAF system

has been used, and log data has been collected and accumulated. Based on this background, this paper utilized the log data from the LEAF system.

Datasets

The used datasets are shown in Table 4. In order to see the effects in a wide range of areas in education, we will prepare three real-world educational datasets for junior high school, high school and university.

Dataset from a junior high school (Dataset X)

As Dataset X, we analyze the logs of the days when Active Reading activities were conducted at the same junior high school in the 2021 and 2022 school years. Active Reading is a reading method in which students read while asking questions (Toyokawa et al., 2024). Active Reading enables one to read quickly and understand the important points, and the class utilized multiple learning tools such as e-Books and Learning Analytics Tools. The target class consists of three classes, Day 1, Day 2, and Day 3 and there is a day of no-class between the first and second days. There were activities with e-Book memo, marker, and timer. There were also quizzes and measuring reading speed (WPM) before and after the class (pre- and post-). There is also a task where students write a summary of the text they have read, and their scores are kept. The indicators are shown in Table 5.

Dataset from a high school (Dataset Y)

Dataset Y focuses on the effectiveness of the problem recommendation system during the summer vacation period from July 20, 2021 to August 23, 2021. This dataset is a study investigating the impact of question recommendation reasons on student learning (Takami et al., 2022). In this setting, six classes were divided into two groups, with three classes receiving question recommendations with reasons for recommendation and the remaining three classes receiving question recommendations without reasons for recommendation. The curriculum was divided into two units, with two classes in Unit A and four classes in Unit B. The main objective was to evaluate the impact of the question recommendation model on the click rate (CVR) of the recommended questions. The indicators are shown in Table 6.

Table 4 Summary of datasets

ID	Data from	Year	Log Count
Dataset X	a junior high school	2021, 2022	38*23
Dataset Y	a high school	2021	103*5
Dataset Z	a university	2019, 2020	206*11

Table 5 Indicators of Dataset X

Abbreviation	Description
ARD D1	Active reading dashboard use on Day 1
ARD D2	Active reading dashboard use on Day 2
ARD D3	Active reading dashboard use on Day 3
MRK D1	Marker use on Day 1
MRK D2	Marker use on Day 2
MEM D1	Memo use on Day 1
MEM NC	Memo use on a day of no-class
MEM D2	Memo use on Day 2
MEM D3	Memo use on Day 3
RDG D1	Reading operations on Day 1
RDG NC	Reading operations on a day of no-class
RDG D2	Reading operations on Day 2
RDG D3	Reading operations on Day 3
TIM D1	Timer use on Day 1
TIM D2	Timer use on Day 2
TIM D3	Timer use on Day 3
QUZ PRE	Score of the pre-quiz
QUZ PST	Score of the post-quiz
WPM PRE	Words per minute before the active reading activity
WPM PST	Words per minute after the active reading activity
SUM SMR	Score of summary after active reading class

Table 6 Indicators of Dataset Y

Abbreviation	Description
TPC	Teaching topic
MDL	With or without reasons for recommendation
RDT	Time spent browsing prior to the start of the teaching period
MRK	Number of markers prior to the start of the teaching period
CVR	Percentage of clicks on recommended questions

Dataset from a university (Dataset Z)

Data from an educational data analysis contest conducted by the Council for Evidence-Driven Education Research (EDE) was used (EDE, 2022).

We use educational data obtained by using the LEAF system at a university. Among them, we used data on digital educational material browsing behavior in a total of four courses offered as Kyu-data in the 2019 and 2020 academic years. The logs were organized by class time and contents IDs, and data were produced for each indicator for each class. The indicators are shown in Table 7.

Table 7 Indicators from Dataset Z

Abbreviation	Description
CNT LP	Count of Lecture Period
SUM OP	Sum of Operations
AVG OP	Average Operations
SUM ATT	Sum of Attending Students
PG LP	Number of Pages Covered in Lecture Period
DV LP	Number of Devices Used in Lecture Period
SUM MRK	Sum of Markers
AVG MRK	Average Markers
MV LP	Number of Teaching Material Variations Used in Lecture Period
AFE ATT	Average Final Exam Score of Attending Students
ONLINE	Whether the Online Course

Research questions

(1) RQ1: What do we get from the observational causal discovery method adapted to real-world educational data?

To answer RQ1, we utilize the causal discovery algorithm "SAM," which employs deep learning to generate hypotheses from large datasets. SAM has shown promising results in real-world biological data, and this study aims to verify its effectiveness in the educational domain.

(2) RQ2: How different are the results between hypothesis-driven and data-driven methods?

For RQ2, the study involves a comparative analysis between Automated Similar Lecture Search as a Hypothesis-Driven Method and the observational causal discovery approach using SAM as a Data-Driven Method. This comparison aims to elucidate the characteristics and differences between these methods, thereby highlighting the strengths and limitations of each.

(3) RQ3: Can the shortcomings of observational causal discovery be improved?

Tackling RQ3 in Study 2, we introduce an updated version of SAM, termed SAM+. This enhanced version allows for the pre-input of improbable relationships. The application of SAM+ to real-world educational data will enable us to assess whether the identified shortcomings of the observational causal discovery method can be effectively addressed.

This research consists of two studies. In Study 1, we address Research Questions 1 and 2 (RQ1 and RQ2), and in Study 2, we respond to Research Question 3 (RQ3). In Study 1, we apply SAM to real-world educational data to explore RQ1 and RQ2. This approach will allow us to generate a diverse range of causal relationship hypotheses that might be difficult to identify based solely on human assumptions. One of the challenges to

be addressed is the generation of improbable hypotheses, which is a critical aspect of our analysis. Study 2 focuses on RQ3, where we implement SAM+. By pre-inputting improbable relationships, we aim to refine the process of hypothesis generation, enhancing the method's overall effectiveness and applicability to educational data analysis. Through these studies, the paper aims to provide a comprehensive understanding of the potential and limitations of observational causal discovery methods in educational settings. We anticipate that our findings will contribute to the refinement of data-driven approaches in educational research and practice.

The remaining sections of this paper are organized as follows. In Section 4, Study 1 is presented, where RQ1 and RQ2 are investigated. This section will detail the methodology, data analysis, and findings relevant to these research questions. Section 5 is dedicated to Study 2, focusing on the exploration of RQ3. It will elaborate on the methodologies employed, the analysis conducted, and the insights gained from this study. In Section 6, a General Discussion will synthesize the findings from both studies, providing a comprehensive understanding of the research questions and their implications in the broader context. The paper concludes in Section 7, summarizing the key findings, discussing the implications, and outlining future work and potential areas for further research.

Study 1

Method

In Study 1, we will answer RQ1 and RQ2. The first step is for RQ1, which is to examine whether SAM applies to educational data and whether it provides effective results. The second step is for RQ2, which is to compare the SAM-based method to the previous method. For this step, we apply SAM to the three datasets.

The detail of the analysis process is shown in Figure 5. First, the log data, which contains information about what and when students have done, are collected. Second, this information is summarized to show how many operations are conducted on digital textbooks in each class. Third, a causal discovery is conducted to determine whether a causal relationship exists. Fourth, the results are displayed as a table in the form of 0s and 1s, by cutting off with threshold. The threshold is set to 0.8. Finally, the results are output as a DAG (directed acyclic graph).

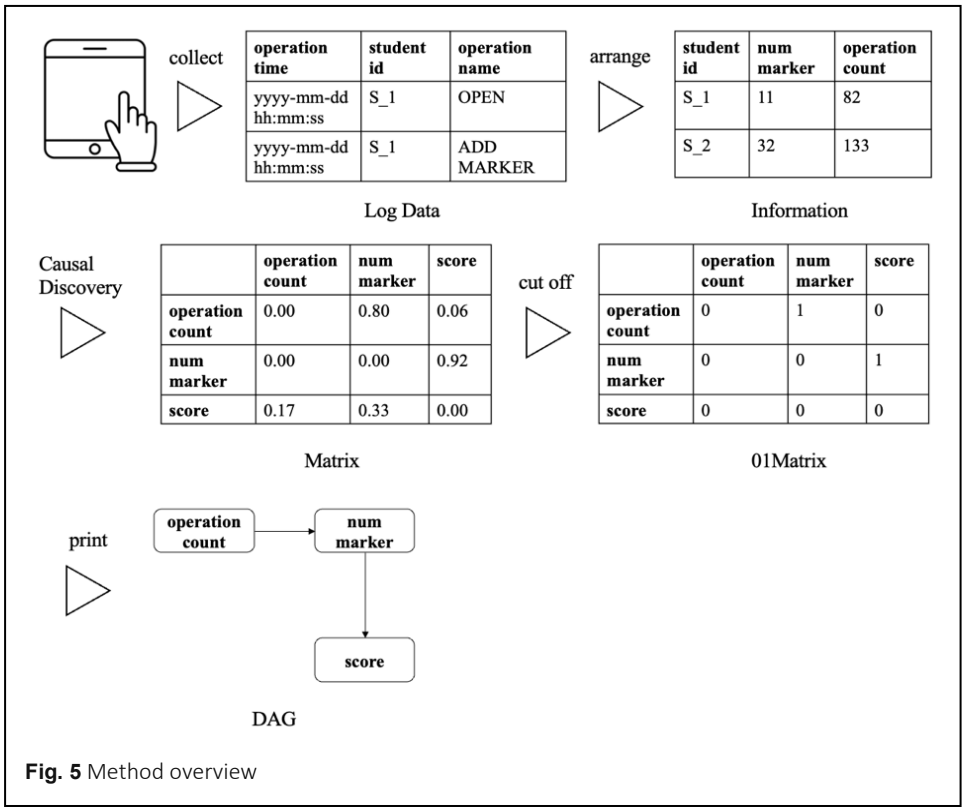

Fig. 5 Method overview

Results

RQ1: What do we get from the observational causal discovery method adapted to real-world educational data?

To answer RQ1, we examined whether SAM applies to educational data and whether it provides effective results. SAM was applied to the three datasets for analysis. Figures 6, 7, 8 show the result of SAM.

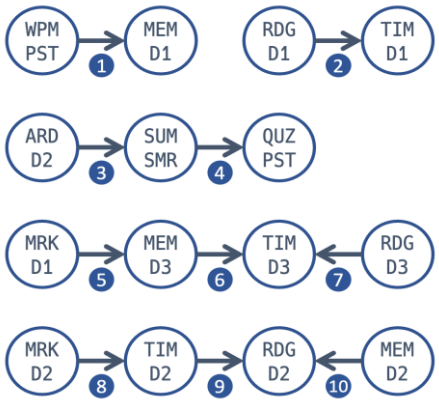

Fig. 6 Result of SAM in Dataset X

(1) "Feasible and Interpretable Causal Relationship"

The results of SAM function as a means to validate hypotheses based on existing knowledge and experience. For instance, it is confirmed that summary scores change as a result of active reading (Figure 6, Arrow 3: ARD D2 → SUM SMR). Additionally, it is evident that the number of attendees affects the number of operations (Figure 6, Arrow 10: MEM D2 → RDG D2). These results are consistent with existing knowledge and experience, and can be intuitively accepted without requiring complex analysis or deep understanding. In other words, SAM successfully captures patterns reflected in the actual data.

(2) "Uninterpretable but Feasible Causal Relationship"

SAM results can also lead to unexpected hypotheses. For example, the number of markers on Day 1 affects the number of memos on Day 3 (Figure 6, Arrow 5: MRK D1 → MEM D3). This relationship might suggest that content deemed important on Day 1 influenced the content on Day 3, but it is difficult to clearly interpret the causal relationship.

(3) "Reverse Causal Relationship"

These causal relationships are unfeasible in the observed direction but may become "feasible and interpretable causation" or "uninterpretable but feasible causation" when reversed. For instance, the number of operations affects the number of times the timer is used (Figure 6, Arrow 2: RDG D1 → TIM D1). Reversing this relationship results in a feasible and interpretable causation. Additionally, the result that WPM after active reading influences the number of memos on Day 1 (Figure 6, Arrow 1: WPM PST → MEM D1) also reverses the time axis. If the number of memos on Day 1 influenced WPM after active reading, it would be an unexpected hypothesis and could be classified as "uninterpretable but feasible causation."

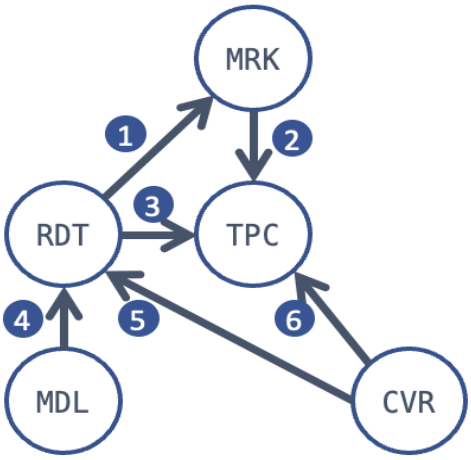

Fig. 7 Result of SAM in Dataset Y

(1) "Feasible and Interpretable Causal Relationship"

The result showing that ReadingTime affects the number of Markers (Figure 7, Arrow 1: RDT → MRK) is both feasible and easy to interpret.

(2) "Uninterpretable but Feasible Causal Relationship"

The results indicate that the number of Markers and ReadingTime before the start of the course influence the Topic (Figure 7, Arrow 2: MRK → TPC, Arrow 3: RDT → TPC). This may be because the Topic is determined by whether the students that entered from high school or had entered from junior high school, with the ReadingTime and Markers differing accordingly before the course starts. However, it is difficult to interpret the specific mechanism behind this causal relationship.

(3) "Reverse Causal Relationship"

The result that the Model affects ReadingTime before the start of the course (Figure 7, Arrow 4: MDL → RDT) is a reversal of the time axis. Additionally, the result showing that CVR affects ReadingTime and Topic before the start of the course (Figure 7, Arrow 5: CVR → RDT, Arrow 6: CVR → TPC) is also a reversal of the time axis.

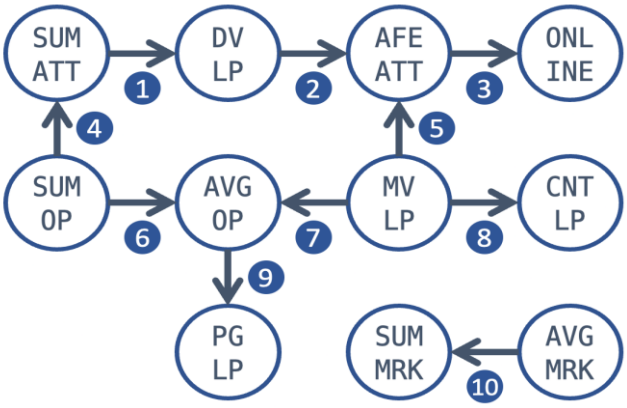

Fig. 8 Result of SAM from Dataset Z

(1) "Feasible and Interpretable Causal Relationship"

The analysis revealed a causal relationship between the number of teaching materials used during class time and the average number of operations (Figure 8, Arrow 5: MV LP → AFE ATT). In other words, the more types of teaching materials used, the greater the number of operations. This result is consistent with existing human knowledge and experience, and is easily accepted without requiring complex analysis or deep understanding. This means that the SAM captures patterns that are evident in the actual data.

(2) "Uninterpretable but Feasible Causal Relationship"

Conversely, there was a causal relationship between the number of device types used in class and the semester grade point average of the students who attended (Figure 8, Arrow 2: DV LP → AFE ATT). In other words, the more types of devices used, the higher the grade. If there is indeed a causal relationship, this is a new educational method worth trying. The causal relationship between whether a course is offered online or not and end-of-semester grades (Figure 8, Arrow 3: AFE ATT → ONLINE) is also counterintuitive. It is hard to imagine that grades could be improved by online courses, but if so, we see new educational possibilities.

(3) "Reverse Causal Relationship"

These causal relationships are unfeasible but feasible in the opposite direction. A causal relationship exists between the total number of operations and the number of participating students (Figure 8, Arrow 4: SUM OP → SUM ATT). However, this causal relationship is clearly reversed. Of course, the total number of operations does not increase the number of students. Rather, the greater the number of students, the greater the number of operations should be, and this is the more natural causal relationship.

RQ2: How different results between the conventional method and observational causal discovery?

Next, to answer RQ2, the SAM was applied to the Dataset X with the same activities as before. In Section 4.2.1, Figure 6 shows the results of the SAM from Dataset X and Table 5 describes each indicator.

(1) "Causality indicating an intervention effect"

The results of the analysis revealed whether there is an intervention effect that we wish to examine. An intervention in education is a series of actions performed by a teacher (human or machine) that produces a change in a student's ability to perform a task. The first day of the study was spent in the classroom. We found that the active reading activity on the second day affected the quality of the summary of the notes (Figure 6, Arrow 3: ARD D2 → SUM SMR). Feedback of these results to the teachers would allow them to repeat what they did on the second day in the next similar class.

(2) "Reverse causation"

The results were similar to those of RQ1. Reversed causality is not unique to the RQ1 dataset. Some time points were also reversed, which also suggests the possibility of a causal relationship. For example, there is a relationship between the number of words per minute in the after-class activity and the amount of notes used on Day 1 (Figure 6, Arrow 1: WPM PST → MEM D1). Given the order of the activities, the direction of causality is opposite. However, if the direction is adjusted, the causal relationship between memo use and WPM is "uninterpretable but feasible causality." It is significant that AI, which can process a large amount of data, can suggest a relationship to something that humans cannot consider a relationship.

(3) "Causation that does not exist"

In addition to the "possibility that a causal relationship exists," there are also indications that a causal relationship may not exist. For example, there are no arrows for active reading dashboard use on Days 1 and 3 ("ARD D1" and "ARD D3"). This indicates that the activities using the dashboards on Days 1 and 3 were not as effective, but the activities on Day 2 may have been very effective. In the future, when designing active reading lessons, we may focus on the activities on the second day. If it is more effective, it will reduce the teacher's workload while contributing to the improvement of learners' learning effectiveness.

Discussion

RQ1: What do we get from the observational causal discovery method adapted to real-world educational data?

These results confirmed that SAM applies to educational data and outputs useful results. However, not all the causal relationship outputs are true, so care must be taken when interpreting them. However, one of the appeals of this method is that it can process large amounts of data and suggest causal relationships that may not have been assumed by humans. Based on the results, it is also possible to select which causal relationships should be examined more deeply. These results suggest that SAM can be useful for educational data analysis. However, to take full advantage of its attractiveness, proper interpretation, and verification of the causal relationships in the output is necessary.

RQ2: How different are the results between hypothesis-driven and data-driven methods?

The auto-comparison method is compared with the causal discovery method SAM. The auto-comparison method can be verified based on the hypothesis, but it is necessary to fix the indices to be compared. However, it also outputs "backward causality," which needs to be improved. In Study 2, we will develop SAM+ that improves this drawback and demonstrate it using RWD.

Study 2

Method

Algorithm developed SAM+

We have developed and used SAM+: Integration of NGLIST with Standard SAM. In this study, we introduce our innovative algorithm, SAM+, which is a step forward by adding NGLIST to SAM to specify the causal direction (Figures 9, 10). In Study 1, we found that SAM suggests many candidates causal relationships, but it also suggests "reverse causation." In Study 2, SAM+ overcomes this shortcoming by adding NGLIST to the SAM; causal directions added to NGLIST are learned as having no causal potential (Table 8). In this study, we compared the original SAM with our newly developed SAM+. This comparative analysis focuses on assessing the ability to handle causal discovery more accurately.

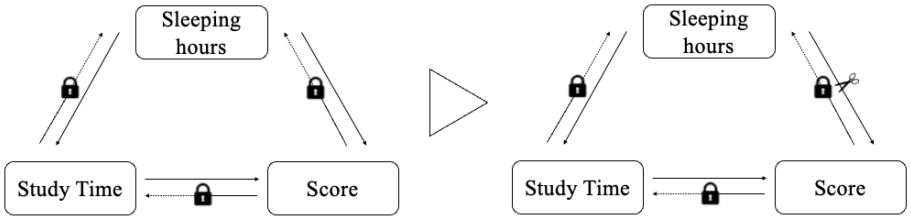

Fig. 9 Image of causal discovery using SAM+

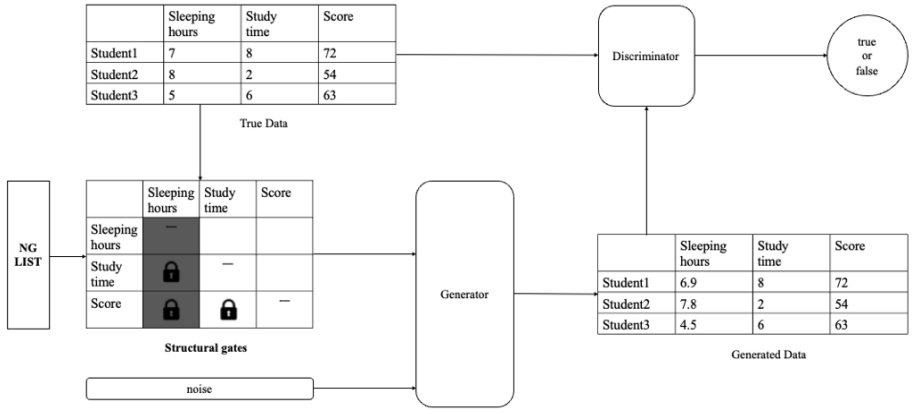

Fig. 10 Overview of SAM+

Table 8 NGLIST

As an example of NGLIST, the NGLIST on the matrix of Dataset Y is shown. The remaining details are provided in Appendix.

NGLIST = [[0,1],[0,2],[0,3],[1,0],[1,3],[2,0],[2,3],[3,0],[3,1],[3,2],[4,0],[4,1],[4,2],[4,3]].

	TPC	MRK	RDT	MDL	CVR
TPC	(0)	0	0	0	
MRK	0	(0)		0	
RDT	0		(0)	0	
MDL	0	0	0	(0)	
CVR	0	0	0	0	(0)

(0): Originally set to 0 because the generated graph is acyclic.

Results

Figure 11-1 shows the result of applying SAM to Dataset X, and Figure 11-2 shows the result of applying SAM+ to the same dataset. In this NGLIST, directions that are impossible on the time axis are set (see attached document for details). The threshold for both figures is 0.8.

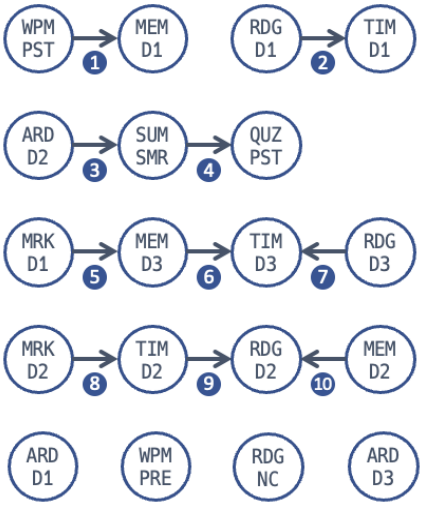

Fig. 11-1 SAM to Dataset X (Same as **Fig. 5**)

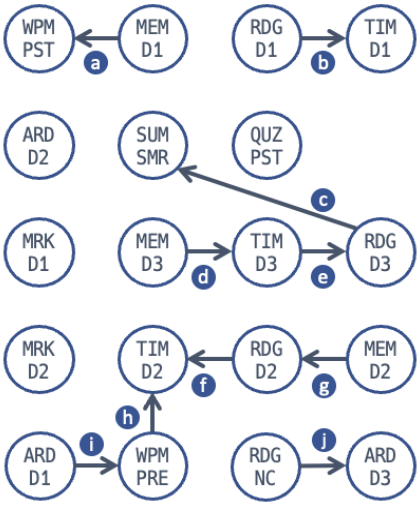

Fig. 11-2 SAM+ to Dataset X

Fig. 11

There are 3 unchanged causations, 4 disappeared causations, 4 newly appeared causations, and 3 reversed causations. Focusing on the reversed causations, first is (1,a). While 1 was "reverse causation," a is "uninterpretable but feasible causation". If the number of notes on the first day affects WPM PST, it could form a good hypothesis since WPM PST is a key indicator in this experiment. The remaining two are (7,e) and (9,f). Does reading operation affect timer use, or is it the other way around? Since "reading operation = timer use + …," either direction could be described as "feasible and interpretable causation." There is also a change in the variable affecting the summary score (SUM SMR), from ARD D2 to RDG D3. This represents a significant change in the hypothesis.

Figure 12-1 shows the result of applying SAM to Dataset Y, and Figure 12-2 shows the result of applying SAM+ to the same dataset. In this case, the NGLIST includes not only

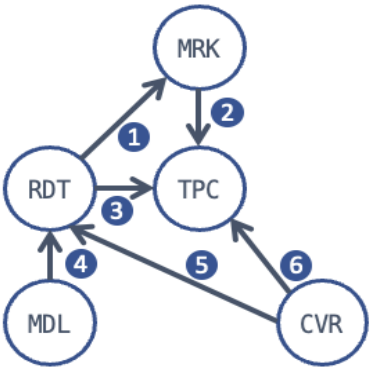

Fig. 12-1 SAM to Dataset Y (Same as **Fig. 6**)

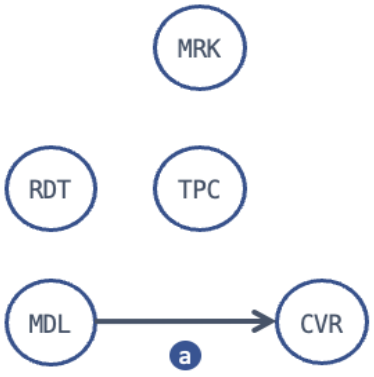

Fig. 12-2 SAM+ to Dataset Y

Fig. 12

Table 9 Structural gate

As an example of Structural Gates, that of SAM+ to Dataset Y is shown.

	TPC	MRK	RDT	MDL	CVR
TPC	0.00	0.00	0.00	0.00	**0.00**
MRK	0.00	0.00	**0.00**	0.00	**0.10**
RDT	0.00	**0.00**	0.00	0.00	**0.00**
MDL	0.00	0.00	0.00	0.00	**0.75**
CVR	0.00	0.00	0.00	0.00	0.00

directions that are impossible on the time axis, but also directions affecting TPC and MDL, as they are deterministically exposed (Table 9).

While SAM has a threshold of 0.8, applying the same threshold for SAM+ caused all causations to disappear, so setting the threshold to 0.75 produced the result shown in Figure 11-2. The sole causation represented by a is the effect originally intended to be observed in this experiment, and the fact that RDT, MRK, and TPC do not affect CVR is a critical point.

Figure 13-1 shows the result of applying SAM to Dataset Z, and Figure 13-2 shows the result of applying SAM+ to the same dataset. In this case, the NGLIST includes directions that are impossible on the time axis, as well as directions where the number of lectures or whether it was online should have no influence (see attached document for details). The threshold for both is set to 0.8.

There are 6 unchanged causations, 2 disappeared causations, 4 newly appeared causations, and 2 reversed causations. Focusing on the reversed causations, first is (3,c). While 3 was "reverse causation," c shows that whether it was online or not influences the final score, making it "uninterpretable but feasible causation" (the distinction between interpretable and uninterpretable might be based on whether it can be expressed mathematically). This seems to be a hypothesis worth testing. Then there's (8,h); 8 was also "reverse causation," but h now represents "uninterpretable but feasible causation."

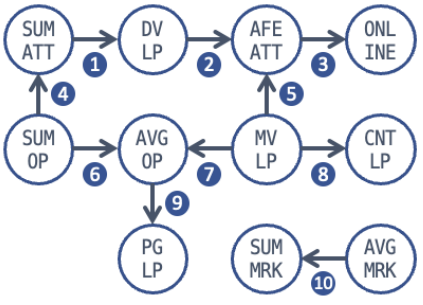

Fig. 13-1 SAM to Dataset Z (Same as **Fig. 7**)

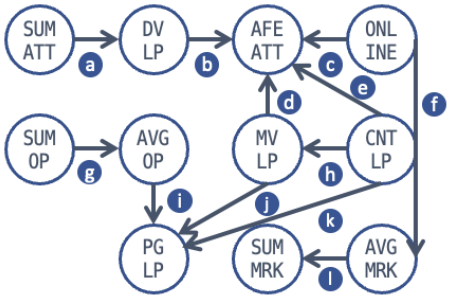

Fig. 13-2 SAM+ to Dataset Z

Fig. 13

Discussion

In this study, we applied two methods, SAM and SAM+, to Dataset X, Dataset Y, and Dataset Z, and compared the results. Observing the changes, disappearances, new occurrences, and reversals in causal relationships from these experiments, we discuss the effectiveness of SAM+.

First, in the results for Dataset X shown in Figure 10, the application of SAM+ led to a notable change where the previous "reverse causation" transformed into "uninterpretable but feasible causation." Specifically, hypothesis (a), which suggests that the number of notes recorded on the first day influences WPM PST, aligns well with the primary purpose of this experiment since WPM PST is a key indicator. Thus, applying SAM+ highlighted more interpretable and feasible causal relationships, producing results that are more aligned with the experimental expectations compared to SAM.

Next, in Figure 11 for Dataset Y, adjusting the threshold to 0.75 in SAM+ revealed the originally anticipated causation. This suggests that SAM+ can more accurately identify causal relationships that indicate intervention effects. Additionally, it was confirmed that RDT, MRK, and TPC do not influence CVR, providing significant insights for this experiment.

Finally, in Figure 12 for Dataset Z, causal relationships previously identified as "reverse causation" were reinterpreted by SAM+ as "uninterpretable but feasible causation." In particular, (c) revealed that whether the session was online or not affected the final score, indicating a hypothesis worth further investigation.

In summary, SAM+ reduces "reverse causation" and increases both "feasible and interpretable causation" as well as "uninterpretable but feasible causation," making it a method that reveals meaningful causal relationships aligned with the experiment's goals. Additionally, relationships previously deemed non-existent in SAM are shown as intervention effects in SAM+, further substantiating the utility of SAM+. This, in turn, enhances the reliability of the experiments and contributes to the formulation of more effective intervention strategies.

General discussion

Key findings

1. Discovering and improving the limitations and potential of data-driven approaches

In this study, we developed and validated a data-driven approach for automated evidence extraction. This data-driven approach demonstrated the ability to generate unexpected

causal hypotheses. However, it also has the drawback of generating unfeasible causal hypotheses. This drawback was addressed by improving the model.

2. Issues in automating causal analysis in education

It is important to note, however, that the current process requires the teacher to manually add causality of 0, which is a major barrier to automation. Future work will focus on overcoming this challenge through pattern recognition and automatic integration of temporal data from other sources. This advancement is critical to fully realize the potential of automated evidence extraction in educational settings.

3. Importance of estimating intervention effects in educational evidence extraction

In addition, estimating the effects of these interventions is essential for providing meaningful feedback to the field. This helps to assess whether the outcomes justify the costs involved.

Implication

Using educational data to understand causal relationships could contribute to the development of a more equitable and effective education system. Data-based insights could inform policies aimed at improving the quality of education and ensuring equal opportunities for all learners. Incorporating individual learner variables could also lead to optimized and personalized education, providing a more effective learning experience tailored to the needs of each student.

Impact on education

1. Evidence-based decision support

This is a transformative approach that significantly enhances evidence-based practice in education. This method harnesses the power of data analytics and machine learning to extract actionable insights from vast amounts of educational data, including routine. In doing so, educators can develop, refine, and implement instructional strategies that are not only theoretically correct, but empirically proven to be effective.

2. Utilization by teachers, learners, and researchers

Based on the provided causal graphs, the system can offer practical advice to teachers, learners, and researchers. A tremendous amount of data is now being collected in educational settings from a variety of Educational Technology tools and services. This

creates opportunities to leverage methods from Artificial Intelligence (AI) and Learning Analytics (LA) to improve both learning and the environments in which it occurs.

However, the analytics results produced by these methods often fail to connect with theoretical concepts from the learning sciences, making it difficult for educators to trust, interpret, and apply them. At the same time, many educational theories are challenging to formalize into testable models that connect to educational data. Causal modeling can help bridge this gap by formalizing the connection between big data and educational theory (Kitto et al., 2023).

Furthermore, graphical causal models can help bridge the disciplinary divide, offering a new tool that assists educators in understanding and, potentially, challenging the technical models developed by LA practitioners (Hicks et al., 2022).

It is essential that this advice be presented in a way that is comprehensible and acceptable to educators. Large language models (LLMs), such as ChatGPT, can also be utilized to support this process.

3. Possibility of evidence-based educational policy and curriculum design

RWE can be used in educational policy and curriculum design. In the example of Dataset Y, since the reasons for the recommendation are effective, it may be recommended that the reasons for the recommendation be given as a policy or curriculum.

Social impact

1. Contribution to a fair and effective educational system

Using educational data to understand causal relationships could contribute to the development of a more equitable and effective education system. Data-based insights could inform policies aimed at improving the quality of education and ensuring equal opportunities for all learners.

2. Prospects for personalized education

Incorporating individual learner variables could result in optimized and personalized education, providing a more effective learning experience tailored to each student's needs.

Limitation

1. RELIANCE on user input for unfeasible causal relationship

Current processes require manual intervention by the user to add unfeasible causal hypotheses (NGLIST). This reliance is a major barrier to achieving full automation and poses challenges to scalability and efficiency. However, the potential use of large language models (LLMs) presents a promising avenue for overcoming this limitation. LLMs can be

trained to identify patterns in causal relationships and detect unfeasible causalities by analyzing large datasets. They could assist in automatically generating NGLIST by recognizing potential contradictions or inconsistencies in the causal directions based on contextual understanding. This would reduce the dependence on user input and improve scalability, efficiency, and the overall automation of the process.

2. Contextual limitations of the data

The effectiveness of automated case extraction methods is highly dependent on the context and quality of the data used. Misinterpretation of the context of the data can lead to erroneous conclusions, underscoring the need for careful data management and understanding of the context.

3. Limitations of performing causal discovery for every analysis

It may not be practical to conduct causal discovery for every analysis, and in such cases, methods like propensity score adjustment to account for background information could be a useful alternative. However, this would shift the approach to a hypothesis-driven framework, which differs from the data-driven approach used in this study. Moving forward, it will be important to consider how these two methodologies—hypothesis-driven and data-driven—can be effectively integrated to optimize the analysis process.

Future works

1. Improved hypothesis generation

In future developments, a key focus will be on enhancing the system's ability to generate more precise and meaningful hypotheses. This will involve incorporating more sophisticated statistical and machine learning techniques to better differentiate between plausible and implausible hypotheses. Additionally, integrating external datasets and prior knowledge from educational theories could improve the accuracy of hypothesis generation. By doing so, we can ensure that the system not only responds to patterns in the data but also aligns with educational insights and pedagogical frameworks. Moreover, we aim to develop mechanisms that enable the system to prioritize hypotheses that have higher potential impact on educational outcomes, guiding educators towards the most effective interventions and strategies.

2. Advances in automation

Emphasis will be placed on overcoming a key barrier to full automation: the current need for teachers to manually input backward causality. To automate this aspect, we will explore the integration of pattern recognition techniques and temporal data from a variety of

sources. Additionally, leveraging large language models (LLMs) could help automate the identification of unfeasible causality and other aspects that currently rely on manual input. LLMs can be trained to assist in predicting and refining causal relationships based on the vast amount of educational data, potentially improving the scalability and efficiency of the system.

3. Intervention effect estimation and feedback

An important area for future research is to estimate the effects of these causal hypotheses. This will provide essential feedback to educators and help determine the utility and cost-effectiveness of the results. Strengthening this evaluative aspect is key to ensuring the applicability and value of the extracted real-world evidence (RWE). As Smith suggests (Smith, 2023), combining counterfactual analysis may also enhance the understanding of intervention effects. However, attention must be paid to the potential issue of "double dipping" when using such methodologies (Gradu et al., 2024). Moreover, the emergence of datasets such as CausalEdu (Gong et al., 2023) offers an opportunity to further develop automated evidence extraction algorithms. Using these datasets, we aim to upgrade the system to better support educational decision-making based on robust evidence.

Conclusion

In conclusion, as an educational data research, this study is an important advance in the field of data-driven educational research, encompassing two very important studies.

Study 1 is a comparative analysis of the SAM algorithm and conventional methods. Study 1, the first part of this study, successfully demonstrated the ability of the SAM (Structural Agnostic Model) algorithm to generate innovative causal hypotheses in the educational domain. This study not only demonstrated the potential of the SAM algorithm by identifying differences from traditional hypothesis-driven methods, but also highlighted its challenges, such as generating causal relationships without regard to practical constraints. Study 2 is the development and demonstration of SAM+. To address the challenges identified in Study 1, Study 2 developed and introduced an improved version of the SAM algorithm, called SAM+. This model was developed specifically to improve the accuracy and reliability of the model in identifying plausible causal relationships; the introduction of SAM+ showed significant improvements in observational causal discovery and came close to maximizing the potential of data-driven approaches in education.

This study highlights the transformative potential of data-driven automated evidence extraction in education. The development and validation of this method opens new avenues for generating causal hypotheses that do not rely solely on traditional human reasoning. Despite challenges such as the generation of unfeasible hypotheses, continuous improvement of the model is paving the way for more accurate and practical applications.

As educational data researchers, we are excited about the prospects this research brings to the field. Continued refinement of these methods will not only deepen our understanding of the dynamics of education but will revolutionize approaches to teaching and learning with data-driven insights. It is hoped that this data-driven methodology will result in more effective and evidence-based educational practices.

Appendix

Dataset Y

NGLIST=

[[0,10],[1,0],[1,3],[1,8],[1,10],[2,0],[2,3],[2,8],[2,10],[3,0],[3,10],[4,0],[4,8],[4,10],[5,0],[5,10],[6,0],[6,8],[6,10],[7,0],[7,8],[7,10],[8,0],[8,10],[9,0],[9,1],[9,2],[9,3],[9,4],[9,5],[9,6],[9,7],[9,8],[9,10],[10,0]]

The array is difficult to read, so it is converted into a table for better readability.
The same approach will be applied to the Dataset Z section as well.

	CNT LP	SUM OP	AVG OP	SUM ATT	PG LP	DV LP	SUM MRK	AVG MRK	MV LP	AFE ATT	ONLINE
CNT LP	(0)										0
SUM OP	0	(0)		0					0		0
AVG OP	0		(0)	0					0		0
SUM ATT	0			(0)							0
PG LP	0				(0)				0		0
DV LP	0					(0)					0
SUM MRK	0						(0)		0		0
AVG MRK	0							(0)	0		0
MV LP	0								(0)		0
AFE ATT	0	0	0	0	0	0	0	0	0	(0)	0
ONLINE	0										(0)

Dataset Z

NGLIST=

[[0,16]
,[1,0],[1,3],[1,5],[1,6],[1,9],[1,10],[1,13],[1,16],[1,18]
,[2,0],[2,1],[2,3],[2,4],[2,5],[2,6],[2,7],[2,9],[2,10],[2,11],[2,13],[2,14],[2,16],[2,18]
,[3,16]
,[4,0],[4,3],[4,5],[4,6],[4,9],[4,10],[4,13],[4,16],[4,18]
,[5,16]
,[6,0],[6,3],[6,5],[6,9],[6,13],[6,16],[6,18]
,[7,0],[7,3],[7,5],[7,6],[7,9],[7,10],[7,13],[7,16],[7,18]

,[8,0],[8,1],[8,3],[8,4],[8,5],[8,6],[8,7],[8,9],[8,10],[8,11],[8,13],[8,14],[8,16],[8,18]
,[9,16]
,[10,0],[10,3],[10,5],[10,9],[10,13],[10,16],[10,18]
,[11,0],[11,3],[11,5],[11,6],[11,9],[11,10],[11,13],[11,16],[11,18]
,[12,0],[12,1],[12,3],[12,4],[12,5],[12,6],[12,7],[12,9],[12,10],[12,11],[12,13],[12,14],[12,16],[12,18]
,[13,16]
,[14,0],[14,3],[14,5],[14,6],[14,9],[14,10],[14,13],[14,16],[14,18]
,[15,0],[15,1],[15,3],[15,4],[15,5],[15,6],[15,7],[15,9],[15,10],[15,11],[15,13],[15,14],[15,16],[15,18]
,[17,0],[17,1],[17,2],[17,3],[17,4],[17,5],[17,6],[17,7],[17,8],[17,9],[17,10],[17,11],[17,12],[17,13],[17,14],[17,15],[17,16],[17,18],[17,19]
,[18,0],[18,3],[18,5],[18,9],[18,13],[18,16]
,[19,0],[19,1],[19,2],[19,3],[19,4],[19,5],[19,6],[19,7],[19,8],[19,9],[19,10],[19,11],[19,12],[19,13],[19,14],[19,15],[19,16],[19,18]
,[20,0],[20,1],[20,2],[20,3],[20,4],[20,5],[20,6],[20,7],[20,8],[20,9],[20,10],[20,11],[20,12],[20,13],[20,14],[20,15],[20,16],[20,17],[20,18],[20,19]]

Abbreviations

AI: Artificial Intelligence; DAG: Directed Acyclic Graph; EDE: Council for Evidence-Driven Education Research; EEF: Education Endowment Foundation; EHR: Electronic Health Records; FCI: Fast Causal Inference; GAN: Generative Adversarial Network; GES: Greedy Equivalence Search; GOLEM: Gradient-based Optimization algorithm for Learning causal Models; LA: Learning Analytics; LEAF: Learning Evidence Analytics Framework; LiNGAM: Linear Non-Gaussian Acyclic Model; LLM: Large Language Model; LRS: Learning Record Store; PC: Peter-Clark Algorithm; RCT: Randomized Controlled Trial; RQ: Research Question; RWD: Real-World Data; RWE: Real-World Evidence; SAM: Structural Agnostic Modeling; SAM+: Enhanced Version of SAM with NGLIST Integration; WWC: What Works Clearinghouse.

Acknowledgements

We would like to express our sincere gratitude to the students and teachers of the junior high school, high school, and university who kindly provided the data for this study. We also thank the members of our research laboratory for their valuable support and insightful discussions throughout the course of this research.

Authors' contributions

Koki Okumura is the main author of this manuscript. Kento Nishioka contributed as the system developer. Kento Koike, Izumi Horikoshi, and Hiroaki Ogata supervised the research and provided critical revisions to the manuscript. All authors read and approved the final manuscript.

Authors' information

Koki Okumura is a doctoral student at the Graduate School of Informatics, Kyoto University, Japan, specializing in learning analytics and educational data science.

Kento Nishioka is a master's student at the Graduate School of Informatics, Kyoto University, Japan, with expertise in algorithm and system development.

Kento Koike is an Assistant Professor at Faculty of Engineering, Tokyo University of Science, Japan. He received a Ph.D. degree in Engineering from Graduate School of Engineering, Tokyo Polytechnic University. His research interests include Intelligent Tutoring Systems, Cognitive Science and Knowledge Engineering.

Izumi Horikoshi is an Assistant Professor at the Academic Center for Computing and Media Studies and the Graduate School of Informatics, Kyoto University, Japan. Her research interests include learning analytics and classroom visualization for formative assessment and reflection. She is a member of APSCE and SoLAR.

Hiroaki Ogata is a full Professor at the Academic Center for Computing and Media Studies, Kyoto University, Japan. His research interests include Learning Analytics, Evidence-Based Education, Educational Data Mining, Educational Data Science, Computer Supported Ubiquitous and Mobile Learning, and CSCL.

Funding
This study was supported by NEDO JPNP20006 and JSPS KAKENHI JP23H00505.

Availability of data and materials
The data used in this study were provided specifically for the purpose of this research and will not be shared due to confidentiality agreements with the participating institutions.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Graduate School of Informatics, Kyoto University, Japan

[2] Faculty of Engineering, Tokyo University of Science, Japan

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

Received: 1 March 2024 Accepted: 5 March 2025
Published online: 1 January 2026 (Online First: 9 July 2025)

References

Brokenshire, D., & Kumar, V. (2009). Discovering causal models of Self-Regulated Learning. In V. Dimitrova, R. Mizoguchi, B. du Boulay & A. Graesser (Eds.), *Proceedings of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling* (pp. 257–264). IOS Press.

Burns, N., & Grove, S. K. (2010). *Understanding nursing research: Building an evidence-based practice* (5th ed.). Saunders.

Buysse, V., & Wesley, P. W. (2006). Evidence-based practice: How did it emerge and what does it mean for the early childhood field? *Zero to Three*, *27*(2), 50–55.

Cook, D. J., Sackett, D. L., & Spitzer, W. O. (1995). Methodologic guidelines for systematic reviews of randomized control trials in health care from the Potsdam Consultation on Meta-Analysis. *Journal of Clinical Epidemiology*, *48*(1), 167–171.

Davies, H., & Nutley, S. (2000). Healthcare: Evidence to the fore. In H. T. O. Davies, S. M. Nutley & P. C. Smith (Eds.), *What works?: Evidence-based policy and practice in public services* (pp. 43–68). Policy Press. https://doi.org/10.2307/j.ctt1t892t3.9

Davies, P. (1999). What is evidence-based education? *British Journal of Educational Studies*, *47*(2), 108–121. https://doi.org/10.1111/1467-8527.00106

de Lusignan, S., Crawford, L., & Munro, N. (2015). Creating and using real-world evidence to answer questions about clinical effectiveness. *Journal of Innovation in Health Informatics*, *22*(3), 368–373. https://doi.org/10.14236/jhi.v22i3.177

EDE: Council for Evidence-Driven Education Research. (2022, September). *Educational data analysis contest. EDE Data Challenge 2023*. https://sites.google.com/view/ede-datachallenge-23/

Evans, D. (2003). Hierarchy of evidence: A framework for ranking evidence evaluating healthcare interventions. *Journal of Clinical Nursing*, *12*(1), 77–84. https://doi.org/10.1046/j.1365-2702.2003.00662.x

Fancsali, S. (2014). Causal discovery with models: Behavior, affect, and learning in cognitive tutor algebra. In J. Stamper, Z. Pardos, M. Mavrikis & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (pp. 28–35). International Educational Data Mining Society. https://educationaldatamining.org/EDM2014/uploads/procs2014/long%20papers/28_EDM-2014-Full.pdf

Gong, W., Smith, D., Wang, Z., Barton, C., Woodhead, S., Pawlowski, N., Jennings, J., & Zhang, C. (2023). CausalEdu: A real-world education dataset for temporal causal discovery and inference. In *CLeaR 2023 Datasets Track 1–9*. https://www.cclear.cc/2023/AcceptedDatasets/gong23a.pdf

Gradu, P., Zrnic, T., Wang, Y., & Jordan, M. I. (2024). Valid inference after causal discovery. *Journal of the American Statistical Association*, 1–12. https://doi.org/10.1080/01621459.2024.2402089

Greenhalgh, T. (2004). Effectiveness and efficiency: Random reflections on health services. *BMJ*, *328*(7438), 529.

Hicks, B., Kitto, K., Payne, L., & Buckingham Shum, S. (2022). Thinking with causal models: A visual formalism for collaboratively crafting assumptions. In A. F. Wise, R. Martinez-Maldonado & I. Hilliger (Eds.), *Proceedings of the*

12th International Learning Analytics and Knowledge Conference (pp. 250–259). ACM. https://doi.org/10.1145/3506860.3506899

Kalainathan, D., Goudet, O., Guyon, I., Lopez-Paz, D., & Sebag, M. (2022). Structural Agnostic Modeling: Adversarial learning of causal graphs. *Journal of Machine Learning Research*, *23*(219), 1–62.

Kitto, K., Hicks, B., & Buckingham Shum, S. (2023). Using causal models to bridge the divide between big data and educational theory. *British Journal of Educational Technology*, *54*(5), 1095–1124.

Kuromiya, H. (2023). *Development of a learning analytics platform for supporting evidence-based teaching* [Doctoral Dissertation (Informatics)]. Kyoto University, Japan. https://doi.org/10.14989/doctor.k24735

Mahajan, R. (2015). Real world data: Additional source for making clinical decisions. *International Journal of Applied & Basic Medical Research*, *5*(2), 82. https://doi.org/10.4103/2229-516X.157148

McMillan, J. H., & Schumacher, S. (2010). *Research in education: Evidence-based inquiry* (7th edition). Pearson.

Nakanishi, T. (2021). *Extracting evidence on lesson design using real world educational data* (Unpublished master's thesis). Kyoto University. (In Japanese)

Ogata, H., Majumdar, R., Akçapinar, G., Hasnine, M. N., & Flanagan, B. (2018). Beyond learning analytics: Framework for technology-enhanced evidence-based education and learning. In J. C. Yang et al. (Eds.), *Proceedings of the 26th International Conference on Computers in Education* (pp. 493–496). Asia-Pacific Society for Computers in Education.

Ouaadi, I., & Ibourk, A. (2023). Causal discovery and features importance analysis: What can be inferred about at-risk students? *Business Intelligence*. In R. El Ayachi, M. Fakir & M. Baslam (Eds.), *Business Intelligence. CBI 2023. Lecture Notes in Business Information Processing, vol 484* (pp. 134–145). Springer, Cham. https://doi.org/10.1007/978-3-031-37872-0_10

Oyinlola, J. O., Campbell, J., & Kousoulis, A. A. (2016). Is real world evidence influencing practice? A systematic review of CPRD research in NICE guidances. *BMC Health Services Research*, *16*, 299. https://doi.org/10.1186/s12913-016-1562-8

Rycroft-Malone, J., Seers, K., Titchen, A., Harvey, G., Kitson, A., & McCormack, B. (2004). What counts as evidence in evidence-based practice? *Journal of Advanced Nursing*, *47*(1), 81–90. https://doi.org/10.1111/j.1365-2648.2004.03068.x

Sackett, D. L., Rosenberg, W. M., Gray, J. A., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: What it is and what it isn't. *BMJ*, *312*(7023), 71–72. https://doi.org/10.1136/bmj.312.7023.71

Sherman, R. E., Anderson, S. A., Dal Pan, G. J., Gray, G. W., Gross, T., Hunter, N. L., LaVange, L., Marinac-Dabic, D., Marks, P. W., Robb, M. A., Shuren, J., Temple, R., Woodcock, J., Yue, L. Q., & Califf, R. M. (2016). Real-world evidence - What is it and what can it tell us? *The New England Journal of Medicine*, *375*(23), 2293–2297. https://doi.org/10.1056/NEJMsb1609216

Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *The American Behavioral Scientist*, *57*(10), 1510–1529. https://doi.org/10.1177/0002764213479366

Smith, B. I. (2023). *Causal discovery and counterfactual explanations for personalized student learning*. http://arxiv.org/abs/2309.13066

Takami, K., Dai, Y., Flanagan, B., & Ogata, H. (2022). Educational explainable recommender usage and its effectiveness in high school summer vacation assignment. In A. F. Wise, R. Martinez-Maldonado & I. Hilliger (Eds.), *Proceedings of the 12th International Learning Analytics and Knowledge Conference* (pp. 458–464). ACM. https://doi.org/10.1145/3506860.3506882

Toyokawa, Y., Majumdar, R., Kondo, T., Horikoshi, I., & Ogata, H. (2024). Active reading dashboard in a learning analytics enhanced language-learning environment: Effects on learning behavior and performance. *Journal of Computers in Education*, *11*(2), 495–522.

Publisher's Note