Study on Exploring Causal Relationships in Educational Big Data Mining

趙, 福政

(Degree)
博士（工学）

(Date of Degree)
2024-03-25

(Date of Publication)
2025-03-01

(Resource Type)
doctoral thesis

(Report Number)
甲第8942号

(URL)
https://hdl.handle.net/20.500.14094/0100490167

博 士 論 文

Study on Exploring Causal Relationships in Educational Big Data Mining

（教育ビッグデータマイニングにおける因果関係分析に関する研究）

2024 年 1 月

神戸大学大学院システム情報学研究科

趙 福政
ZHAO FUZHENG

Table of contents

1. Conceptual and theoretical background of causal relationship exploration

1.1 The concept of causality

A multi-perspective definition of causality from different disciplines and application scenarios. Causality is rooted in different disciplines and application scenarios, which has both the traditional philosophical requirement of its language logic and the statistical emphasis on the necessity of the premise for things to happen. Meanwhile, with the development of experimental hypotheses in psychology and data prediction in economics, the nature of prediction in causality has been embedded, and the concept of counterfactuals based on the analytic assumptions of missing data has gradually become an important concept and a fundamental issue in causal relationship exploration.

First, the traditional philosophical or statistical definition based on the temporal and logical dimensions. Causality is the logical premise of causal relationship exploration, and it is a process of inference based on the understanding of causality. For the definition of causality, traditional philosophical and statistical discussions are centered around necessary and sufficient conditions or time series (Scheines, 1997). In the context of theories based on strict temporal or logical correlations, Hume defined causality as a precondition for the existence of things (Stroud, 1978). Specifically, if a cause is an object and an effect is also an object, then when the former does not exist, the latter does not exist either. Similarly, Burks (1951) attempted to discuss a more precise, explicit, and formal approach to defining causality in terms of the logical structure of causal propositions, causal necessity and sufficiency, the temporal order in which causality occurs, and counterfactual reasoning. As a leading scholar of statistics, Pearson (1920) identifies correlation as a key element of cognitive science. In defining causality, he agrees with the theory of temporal or logical correlation of causality, although there is a dispute with Hume on whether counterfactual factors should be emphasized in causality. At the same time, he blends causation and correlation, arguing that the former is a special case of the latter, emphasizing that causation hinges on how a change in one variable affects the drag of a change in another variable (Pearl, 2009).

Second, the embedding of the nature of prediction in the definition of causation. For a clear and in-depth definition of the concept of causality, Simon (1952) breaks with the traditional and usual approach centered on the temporal or spatio-temporal sequence of events and emphasizes the predictability between events, providing a detailed and insightful perspective that introduces a unique dimension to the understanding of causality (Simon, 1952). He argued that causal relationships are formed by the predictability or anticipation of an event.

Finally, the fundamental problem of causal relationship exploration from the perspective of missing data. Holland (1986) innovatively rethinks the definition and fundamental problem of causality from the perspective of missing data, arguing that causal relationship exploration is essentially a problem of missing data. Holland's analysis revealed that each unit of data could only observe one of many experimental outcomes. He referred to those potentially missing results as counterfactuals. The basic problem of causal analysis then becomes counterfactual relationship exploration, which involves comparing actual outcomes with hypothetical scenarios and inferring causality by understanding the impact of changes or interventions.

1.2 Research status of exploring causal relationships

The maturation of causal relationship exploration cannot be separated from the maturation and development of its concepts, which are manifested in diverse methods and interdisciplinary application scenarios. First, as a research method with close relevance to statistics, causal relationship exploration draws on many approaches from statistics and machine learning while facing the challenge of confusing concepts and methods. Second, as an interdisciplinary application method, the development of causal relationship exploration has obvious traces of multidisciplinary development.

First, statistics is a necessary method for causal relationship exploration. While affirming the role of statistical methods in causal relationship exploration, Holland (1986) distinguishes the difference between causal and correlational relationships. At the same time, he clarifies the impact of confounding variables in causal relationship exploration and analyzes the finding that randomized experiments, or methods that mimic randomized experiments, contribute to causal relationship exploration. Second, the complexity of causal relationship exploration is also reflected in the underlying concepts. Scheines (1997) clarifies the basic concepts in causal relationship exploration, which include the concept of causal diagrams as graphical models to visually depict causal relationships, counterfactual inference to address the problem of missing data, mitigating the effects of confounding and causal attribution on precise causal relationship exploration, and Bayesian or structural equation modeling for the causal representation and analysis methods. Finally, along with the deeper development of causal relationship exploration, causal relationship exploration puts higher demands on the limited computational resources, the efficiency and stability of algorithms, and different data types. In order to increase the adaptability and efficiency of causal relationship exploration, Spirtes (2001) proposed a model for causal relationship exploration that allows relationship exploration methods to provide partial results at any time before the relationship exploration is fully completed, ultimately realizing the efficient use of limited computational resources. At the same time, inference algorithms have continued to evolve. Taking graphical model inference methods as an example, Lauritzen and Jensen (2001) introduced the conditional Gaussian distribution method in order to achieve stable local computation of graphical models as a way to improve the computational efficiency of the algorithms. Different data place various requirements on causal relationship exploration, specifically, when dealing with diverse data characteristics, extending or adapting the algorithm becomes an indispensable approach. Eichler (2012) extends the traditional granger causality approach when performing causal relationship exploration on time series data. He extended the notion of causality to time-continuous processes by defining a graphical model of the labeled time course through local independence, and estimating causality in time-continuous processes with localized causality.

Second, the developmental lineage of causal relationship exploration in an interdisciplinary context. From the Nobel Prize in economics to the Turing Award in computational science, causal relationship exploration has been applied to the fields of sociology, psychology, economics, and, most recently, artificial intelligence, with great potential for causal analysis. As an important tool for understanding social structures, behaviors, and interventions, causal relationship exploration provides an important tool for identifying causal relationships in complex social systems, especially for analyzing the impact of social applications of various policy measures (Keohane, 2009). The contribution of causal relationship exploration in economics lies mainly in econometric approaches to policy evaluation, such as heterogeneous causal effects and breakpoint regression causal relationship exploration, which provide techniques for empirical research in economics (Imbens & Rubin, 2015). The research focus of causal relationship exploration in psychology is more on the understanding of micro-behavior, the treatment of abnormal behavior, and the intervention of mental activities, and causal

relationship exploration establishes causal relationships for researchers in experimental and observational studies (Cook, Campbell & Shadish, 2002). The field of artificial intelligence, which is closely related to the field of statistics, has also received a great deal of attention. Judea pearl was awarded the Turing Award, the highest prize in computer science, in 2011 for her contributions to causality. An article by Savage (2023) in nature discusses the significance of causal relationship exploration for artificial intelligence. Since current AI systems make predictions based only on associations in the data, which are highly susceptible to any changes in the relationships between these variables. So when the statistical distribution of learned relationships changes over time, behavior and other external factors, the structure of AI's analysis changes as well, reducing accuracy. Savage argues that a key benefit of causal relationship exploration is that it makes AI more responsive to changing environments.

1.3 Development trend for exploring causal relationships in Education Big Data mining

Existing research priorities and challenges in causal relationship exploration methods. Spirtes, Glymour and Scheines (2000) systematically introduced the concept of causal relationship exploration in their study, highlighting the basic concepts and the current state of development in statistics and machine learning. In 2017, Peters, Janzing and Schölkopf, in a state-of-the-art review of causal relationship exploration review, the basic principles and recent developments in causal learning algorithms and methods were presented. Based on the analysis of the above findings, it appears that many research challenges persist and evolve, especially in the context of Education Big Data.

Confounding variables, as one of the unobserved consequences of causal relationship exploration, lead to the complexity of causal relationship exploration (Rosenbaum, 2002). The biggest impact of Education Big Data on causal relationship exploration mainly comes from the increase in the amount of data. Traditionally, the identification of learning behaviors mainly relies on measurements such as questionnaires, interviews, and observations to define, characterize, and extract behaviors. However, the large amount of content-rich and complex data raises the dilemma of traditional behavior identification methods and behavior definition based on statistical methods. Muller, Winship and Morgan (2014) suggested that confounding variables have an impact on the observation of learning behaviors, and that correctly identifying and dealing with these factors has become an important causal relationship exploration challenge.

Temporal order building. Causal relationship exploration requires explicit temporal patterns, order, and direction, as opposed to temporally mixed data (Cook, Campbell & Shadish, 2002). With the adoption of online learning platforms, Education Big Data characterized by log data, has being accumulated, which provides both opportunities and challenges for causal relationship exploration. In particular, the dynamic nature of the learning environment increases the complexity of causal relationship exploration.

Complex interactions and nonlinear behaviors characterize digital learning environments. The huge impact of interactions on analyzing complex relationships between data variables is well established in statistics. The rapid development of digital learning and e-learning channels has made it easier to collect learning behavior data from mouse clicks, to eye-tracking attention data, log-type reading manipulation data, teachers' teaching behaviors, students' learning status, and interactions between course materials are recorded with more complex nonlinear features. In the context of Education Big Data, how to extract behaviors from these complex behavioral data and sort out these interactions poses a continuous challenge.

1.4 Conclusions

The concept of causal relationship exploration has been shaped mainly from other fields such as philosophy, psychology, economics, and statistics, with the same influences from interdisciplinary developments behind its development: Definitions are characterized by both linguistic and logical strict qualifications, hypothetical premise-setting, and statistical-mathematical methods of deduction; Diverse methods and complex scenarios. From intuitive relationship exploration of graphs, counterfactual estimation of outcomes, to Bayesian networks and structural equation modeling based on conditional probability, these innovative methods are applied to the estimation of the utility of medical treatments, the evaluation of the effectiveness of the application of economic policies, and the precision of artificial intelligence algorithms.

In the context of Education Big Data mining, causal relationship exploration ushers in new challenges among which confounding variables, temporal order building, complex interactions in digital learning environments and nonlinear behavioral features are challenges that have always existed and continue to evolve, generating new problems. First, the huge amount of data containing confounding variables likewise raises dilemmas for traditional behavior recognition methods and behavior definition based on statistical methods. Second, the dynamic nature of the learning environment highlights the variability of learning behaviors and increases the complexity of causal relationship exploration of educational behaviors. Finally, in the context of Educational Big Data mining, how to extract behaviors from nonlinear and d-interacting data and sort out these interacting behaviors poses a continuous challenge.

2. Practices and challenges of causal relationship exploration in Educational Big Data mining

2.1 Current challenges of causal relationship exploration

2.1.1 Challenges of causal relationship exploration regarding Education Big Data

Challenges arising from the characteristics of Education Big Data itself. The first is confounding variables in the data. Morgan and Winship (2015) systematically explore the core issues of causal relationship exploration, especially the problem of confounding variables, starting from the analysis of the relationship between counterfactuals and causal relationship exploration. The results suggest that the presence of unobserved variables may trigger misleading causal relationship exploration results. Second, a key problem derived from statistics, that is selection bias. Stuart (2010) provides an account of the problem of selection bias in causal relationship exploration and proposes a matching approach to solving the problem. Selection bias is the tendency for data analysis to be biased when analyzing non-random samples or biased data, affecting the accuracy of causal relationship explorations. Simpson's paradox refers to the fact that different ways of combining data can produce different associations, which can be misleading for causal relationship explorations. For this reason, Pearl (2009) specifically discusses the relationship between Simpson's paradox, selection bias and emphasizes the importance of causal relationship exploration under the influence of Simpson's paradox. Finally, another important issue arises from the challenge of causal relationship exploration from temporal order data. With the development of Education Big Data, temporal order data represented by log data, has been collected in large quantities with becoming one of the most important sources of data for causal relationship exploration. Ten Have and Joffe (2012) placed a research perspective on the issue of dealing with temporal order and causal relationship exploration by reviewing the relationship between temporal patterns, sequences, and directions of temporal order data and accurate causal relationship exploration.

Correlation analysis is the mainstream in conventional Education Big Data mining regarding the causal relationship. Focus of causal relationships exploration is not on confirming causal relationships in behavior, so the results of correlation analysis cannot be applied to actual practice. With the development of algorithm techniques, more scalable algorithms that can access an increasing number of computational resources have been proposed (Hashem et al., 2015). Moreover, there is greater availability of large amounts of fine-grained education data than before (Dietze et al., 2016). Despite it being easy to collect detailed learning data from multiple resources and to analyze them in order to provide educational suggestions or recommendations, the analysis results may not be sufficient to understand the critical information (Maldonado-Mahauad et al., 2018). In other words, through the use of some convenient data in a Web-based educational environment such as an e- book system, and the use of many mature analysis techniques such as sequence analysis, some learning behavior patterns can be easily found. However, it is not easy to understand the underlying reasons for behavior patterns. For example, Li et al. (2018) and Yin et al. (2019) found a backtrack reading behavior pattern which showed that certain students often return immediately to previous pages when they read e- textbooks; however, it is difficult to understand the reason for this particular behavior pattern. Therefore, it is very important to understand why the learning patterns or strategies behind behaviors occur, especially in terms of the complex social and cognitive processes involved.

2.1.2 Practical approaches of causal relationship exploration

The practical approaches to causal relationship exploration have been fruitful so far. Depending on the method and purpose of their realization causal relationship exploration has been divided into two categories: one is direct causal estimation and the other is concerned with structural relationships among variables. The former includes traditional statistical methods represented by randomized controlled trials, propensity score matching, instrumental variables, difference-in-differences and regression discontinuity designs (Angrist & Krueger, 2001; Bertrand, Duflo & Mullainathan, 2004; Lee & Lemieux, 2010; Rosenbaum & Rubin, 1983). First, there are proven applications of these traditional methods, which have significant inferential accuracy and focus on direct causal estimation. In addressing the problem of confounding bias, randomized controlled trials and propensity score matching, are considered the gold standard in causal relationship exploration (Rosenbaum & Rubin, 1983; Rubin, 2008). Another difference is that traditional methods are more oriented towards practical problems. Among the newer methods represented by Bayesian networks, structural equation modeling, and neural networks (Bareinboim, Tian & Pearl, 2022; Pearl, 2009; Tabri & Elliott, 2012), methods like Bayesian networks are generally based on directed acyclic graphs to represent causality, which is essentially a probabilistic graphical inference. In addition, it is not only compatible with the scenario of analyzing large datasets with complex interactions, but also has strict requirements for data independence (Pearl, 1995). Therefore, a series of validation assumptions need to be devised on the data when this type of methodology is applied to causal relationship exploration, a limitation that is even more pronounced when it comes to causal relationship exploration in observational studies (Lu, Zheng & Quinn, 2023). In addition, it can be categorized into three types based on the principles of causal relationship exploration, experimental methods, observational methods, and graphical models based on machine learning and deep learning, as shown in Table 1.

Table 1. Three categories of causal relationship exploration methods

Category	Method	Advantages	Disadvantages
Experimental methods	Randomized controlled experiments	most widely used and highly valid	Selection bias
	Regression discontinuity design	Uses natural cutoffs, provides quasi-random assignment to simulate randomization	
	Difference-in-differences	Differential control for unobserved time-varying confounders	
Observation methods	Propensity score matching	Reduces selection	Unable to account for unobserved confounders
	Instrumental variables	Control for unobserved confounders	Relies on strong assumptions, violation of which can lead to biased estimates
Graphical methods	Bayesian networks	Help visualize and understand complex causal structures	Assume there is matching fidelity between data and causal graphs
	Neural networks	Capture complex nonlinear relationships	The result is unstable

	Structural equation modeling	Captures unobservable structure in causal relationships	Selection bias

2.2 Limitations of current causal relationship exploration methods in Educational Big Data mining

Big data applications in education undoubtedly provide new opportunities and challenges for causal relationship exploration. Unlike traditional data characteristics, Education Big Data is characterized by large volume, multiple types, and high formation rate (Manyika et al., 2011). Secondly, with the establishment of digital learning environments and the spread of network communications, student information systems, learning management systems, online assessments, and other interactive platforms provide a diversity of data for learning analytics and educational big data applications (Siemens, 2013). In addition, data dimensions have been enriched. Data dimensions in the context of educational big data are most evident in the fact that contextual data related to learning environments that change over time are recorded, enriching the understanding of educational phenomena, learning behaviors, and characteristics (Liñán & Pérez, 2015).

2.2.1 Limitations of Education Big Data for causal relationship exploration

First, according to existing research, selection bias and confounding variables are among the key factors affecting causal relationship explorations. The diversity of sources and types of Education Big Data increases the difficulty of causal relationship exploration in coping with selection bias and confounding variables. Although, propensity score matching or instrumental variables, can be used to address selection bias and control for unobserved confounders, both of these methods are subject to strict requirements on data distribution and independence, which happen to be difficult to ensure in real data environments. Second, the collection environment of Education Big Data often involves complex interactive learning behaviors, and this nonlinear nature of the data sets limits on the use of traditional causal relationship exploration methods and models (Van de Schoot et al., 2014). This has implications for statistics and machine learning that are predicated on strict experimental assumptions. Thirdly, the growing importance of ethical and privacy issues. Educational Big Data contains sensitive private information about students, while ethics also limit the potential negative impact of different educational interventions on students' learning behavior and psychology.

2.2.2 Limitations of approaches, models and algorithms for causal relationship exploration

Unraveling causal relationships between behaviors is a foundational pursuit, dating back to the early 20th century when randomized controlled trials emerged as the gold standard for causal relationship exploration (Fisher, 1935). However, when randomization becomes unfeasible or unethical, quasi-experimental designs were introduced to address real-world constraints (Angrist & Krueger, 1991). Drawing inspiration from economics, regression discontinuity design was employed to infer causal effects, utilizing natural thresholds or cutoff points for comparison (Campbell & Riecken, 1968). In addition to this historical lineage with econometric approaches, causal relationship exploration has also been facilitated through time-series analysis, panel data models, and the difference-in-difference method (Wooldridge, 2019).

In recent times, there has been a growing interest in machine or deep learning-based causal relationship exploration. Constraint-based approaches, such as the Peter-Claudia algorithm,

have found applications in the field of education. These methods base their causal relationship explorations on clear conditional independence relationships within large observational datasets (Spirtes, Glymour & Scheines, 2000). It's important to note that they assume the faithfulness condition, which may not always hold in real-world data. Structural equation modeling offers another avenue for causal relationship exploration, representing causal relationships among variables through structural equations (Ping & Cunningham, 2013). However, distinguishing causation from correlation is imperative when using this model (Pearl, 2009).

In contrast to traditional hypothesis testing, a combination of conditional probabilistic relationship exploration with directed acyclic graphs has been employed to discover causal relationship exploration in the field of education. Both Bayesian network analysis (Ellis & Wong, 2008) and the Linear Non-Gaussian Acyclic (LiNGAM) model (Shimizu, 2014) have been utilized for this purpose. However, challenges persist, including computational complexity in learning network structures for Bayesian network analysis (Parviainen & Koivisto, 2012) and the potential limitation of the linearity assumption in LiNGAM (Peters et al., 2014).

2.3 Practices and challenges of deep learning in causal relationship exploration

2.3.1 Deep learning techniques offer application potential for causal relationship exploration

First, the biggest advantage of applications based on deep learning techniques is the automatic learning of hierarchical representations of features from data, especially in the face of complex data containing numerous confounding variables, which enables the recognition of complex patterns (LeCun, Bengio & Hinton, 2015). Second, the ability to capture complex nonlinear relationships. Deep learning-based models generally consist of multiple layers of interconnected neural networks, and this multilayer representation can represent nonlinear relationships through nonlinear transformations, and can also be coupled with nonlinear activation functions to model nonlinear relationships (Goodfellow, Bengio & Courville, 2016). Thirdly, deep learning has better scalability and adaptability. Schmidhuber (2015) found that deep learning, with its scalable number of neurons and layers, is more suitable with the massive amount of Education Big Data. Meanwhile, this scalable structure of deep learning can cope with complex noisy data and increase the robustness of data feature mining and analysis (Zhang, et al., 2018).

2.3.2 Challenges in causality exploration based on deep learning

Causal relationship exploration through neural network methods has seen significant progress as a burgeoning educational technique. Traditional methods, exemplified by experiments, entail substantial investments of time and financial resources. Conversely, machine learning methods, particularly Bayesian networks, grapple with limitations related to high-dimensional data, linearity assumptions, conditional independence, and distinguishing causation from correlation. Neural networks offer distinctive advantages in causal relationship exploration. They excel in capturing non-linear data limitations and challenging hypotheses, which can be daunting for other algorithms like LiNGAM. Louizos et al. (2017) introduced a deep latent-variable model built on neural networks to unveil non-linear causal relations within complex structures. Furthermore, one standout advantage is the capacity to directly learn causal relationships from raw data without prior knowledge of underlying causal structures (Nauta et al., 2019).

Generative Adversarial Networks (GANs) bring substantial insights to causal relationship exploration boasting several benefits: the flexibility to capture intricate data distributions, simulating the estimation of causal effects under various interventions (Goodfellow et al., 2014), enhancing the robustness of causal models, and being suited for nonparametric models (Shalit, Johansson & Sontag, 2017). The primary concept behind exploring causal relationships with GANs stems from their ability to estimate causal effects by modeling the distribution of potential outcomes based on observed data (Kocaoglu et al., 2017). Challenges tied to limited sample sizes, which can hinder model training and causal structure discovery, are effectively addressed by GANs. These networks generate synthetic data that retains the statistical properties of the original dataset, mitigating the adverse effects of sample size limitations and imbalances (Gao et al., 2021). The robustness and generalizability of causal relationship exploration with GANs are further bolstered by their capacity for counterfactual reasoning. This enables models to perform counterfactual inference by generating data samples representing hypothetical scenarios under different interventions (Grari et al., 2023).

Deep learning-based causal relationship exploration faces evident challenges, particularly in the case of emerging techniques like GAN-based causal relationship exploration. These challenges stem from the instability of causal results and the associated evaluation metrics. Result instability is a common issue in the deep learning field, often manifesting as mode collapse due to the failure of modes learning within a limited diversity of generated samples (Goodfellow et al., 2014). Some studies posit that GANs may suffer from vanishing and exploding gradients, which impede stability. The choice of loss function for both the generator and discriminator, as a specific structural aspect of the network model, amplifies the volatility of results. Regarding evaluation metrics, commonly employed ones include Inception score, Wasserstein, Frechet inception distance, and 1-Nearest Neighbor. However, these metrics are not suitable for evaluating GAN-based causal relationship explorations because these causal results lack original causal relationships, making it impossible to reference real samples.

3 Causal relationship exploration process in Education Big Data mining

3.1 Definition of causal relationship exploration process in Education Big Data mining

3.1.1 Understanding causal relationship exploration process

Khalil, Prinsloo and Slade (2022) provide an introduction to the concepts of frameworks, models and architectures in the field of Learning Analytics and Education Big Data, emphasizing that the distinction between the above three definitions depends crucially on contextual information. First, framework refer to the interaction and interdependence between different elements of a given phenomenon. Unlike framework which focus on the relationship between the basic elements, models focus on setting the scope of operation and data to achieve the set goals. Architectures mainly consist of the composition and sources of data to realize the analysis methods of models or models. Based on the understanding of the difference between frameworks, models and architects, causal relationship exploration process for Education Big Data refers to the steps and sequence of implementation for realizing causal relationship exploration in Education Big Data mining.

3.1.2 Analysis of existing causal relationship exploration steps

First, Morrison (2012) divides the exploratory steps for determining causal relationships into four stages: identification of the target causal relationship, constructing the causal relationship, testing the causal relationship, and drawing conclusions about the causal relationship. Also, he summarized the methods of determining causal influence as behavioral research and experimental methods. Second, current approaches to causal relationship exploration are mainly categorized into statistical and design-based models, which rely on traditional experimental and statistical methods, respectively, to design inference steps and construct models of causal relationship exploration. Hammerton and Munafò (2021) summarized current steps of causal relationship exploration, starting with an analysis of the challenges of causal relationship exploration from data, and suggested that incorporating a design-based approach into causal relationship exploration steps is important. When making traditional causal relationship explorations, observational data are subject to confounding variables, selection bias, and nonlinear features that lead to inaccurate causal relationship exploration results. According to Richmond et al. (2014), currently available causal relationship exploration methods mainly attempt to minimize or eliminate sources of bias to increase the accuracy of causal relationship explorations. Design-based causal relationship exploration methods are primarily a complementary approach to statistical methods, such as cross-group analysis and multidimensional comparative analysis. Finally, with the development of Education Big Data and the limitations of traditional methods, Murnane and Willett (2010) provide a systematic approach to statistical and design-based causal relationship exploration modeling by analyzing educational data and the state-of-the-art methods of causal relationship exploration, answering several core questions in the process of causal relationship exploration: what type of data is needed for causal relationship exploration, how does the analytical methodology need to be matched to the data, the interpretation and confirmation of causal relationship exploration results, and the application of causal relationship exploration results.

Causal relationship exploration is part of the analytic models proposed in the field of Education Big Data and Learning Analytics, but there are differences in analytic focus and purpose. As previously discussed, causal relationship exploration aims to identify and analyze the causal

elements and causal effects of educational phenomena, learning behaviors, and intervention strategies. In contrast, current models in Education Big Data and Learning Analytics have vastly different research focuses, purposes, and approaches. According to Yin and Hwang (2018), current learning analytics models focus on prediction, structure discovery, and relationship mining as research objectives. Causal relationship exploration is one of the frequently used methods in relationship mining in learning analytics models.

3.2 Causal relationship exploration process in Education Big Data mining

3.2.1 Current Education Big Data mining process

The term Education Big Data mining first appeared in a workshop in 2005, and then in 2008 in the First International Conference on Education Data Mining was held (Baker & Inventado, 2014). As a sister community to educational data mining, research on learning analytics and its models followed. In 2007 Baker presented a model, "wisdom-knowledge-information-data," calling it a "Knowledge Continuum." This model emphasizes data processing in which knowledge is converted into a meaningful form (Baker, 2007). Compared with the above abstract learning analytic process, Campbell, DeBlois, and Oblinger proposed the "Five Steps of Analytics": capture, report, predict, act and refine, for a more simplified, practice-oriented structure procedure in the same year (Campbell, DeBlois, & Oblinger, 2007). After the basic analytical process had been constructed, coinciding with the rise of network analytic research, Hendricks, Plantz, and Pritchard shifted the emphasis of the learning analytics framework study to the objectives of web analytics in 2008 (Hendricks, Plantz, & Pritchard, 2008). In their perspective, the four operations of defining goals, measuring outcomes, using the resulting data, and sharing data for other stakeholders were identified as four objectives when using web analytics in education. These four objectives commonly constitute a learning analytics framework centered on the analysis objective.

With the recognition of the important role of social software in E-learning, Dron and Anderson noted that one kind of distinct dynamics has emerged in educational settings. It follows that group, network, and collective concepts gradually gained attention. Then, in 2009, they proposed a "Collective Applications Model" framework including five steps in Education Big Data Mining process, that is selecting, capturing, aggregating, processing, and displaying. This mining process was also classified into three cyclical phases: information gathering, information processing, and information presentation (Dron & Anderson, 2009). The cyclic structure with the head and tail connected was first put forward in this framework. In 2011, the definition of learning analytics was established at the 1st International Conference on Learning Analytics and Knowledge, followed by the learning analytics framework development in the theoretical dimension. Meanwhile, Elias provided a comprehensive learning analytics framework by summarizing the existing mining processes consisting of select, capture, aggregate & report, predict, use, refine, and share (Elias, 2011). Despite the fact that the learning analytics frameworks were constructed in relation to the analysis process, there were still questions regarding combining them with the educational theory (Romero & Ventura, 2013). In 2012, Chatti et al. introduced a successful learning theory, namely "Kolb's Experiential Learning Cycle," into the learning analytics. They also described a reference model for learning analytics, in light of an iterative cycle proposed as an older learning theory (Clow, 2012). This framework starts with data collection and processing, then goes into analytics and action, importantly, not ending with post-processing but entering a new cycle in a way that affects the next data collection. Similarly, in 2013, Siemens proposed a learning analytics cycle framework adopting a systems approach that includes seven proceses: collection,

storage, data cleansing, data integration, analysis, visual presentation, and action (Siemens, 2013).

As a subfield of learning analytics, the learning behavior analysis model basically follows the basic ideas of existing learning analytics models. Although there are differences between the above learning analysis process, one thing they have in common is that each process lacks an analysis result confirmation step.

3.2.2 The obvious challenge of causal relationship exploration

In the last decade, technologies of Education Big Data mining and Learning Analytics have made rapid progress (Baker, 2019). With the development of algorithm techniques, more scalable algorithms that can use increasing computational resources have been proposed (Hashem et al., 2015). Moreover, there is greater availability of large amounts of fine-grained education data than before (Dietze, Siemens, & Taibi et al., 2016). Despite it being easy to collect detailed learning data from multiple resources and to analyze them for providing educational suggestions or recommendations, the causal relationship exploration results may not be sufficient to understand the critical information (Maldonado-Mahauad et al., 2018). In other words, through the use of some convenient data in a Web-based educational environment such as an e-book system and the use of many mature analysis techniques such as sequence analysis, some learning behavior patterns can be easily found. However, it is not easy to understand the reason why behavior patterns happen. For example, Yin et al. (2018; 2019) found a backtrack reading behavior pattern which showed that certain students often return immediately to previous pages when they read e-textbooks. However, it is difficult to understand why this particular behavior pattern happens. Therefore, it is very important to understand why the learning patterns or strategies behind behaviors occur, especially in terms of the complex social and cognitive processes involved.

The confirmation of the causal relationship exploration results mainly depends on the judgment of the correct probability of the results obtained by the analysis algorithms. With this help, it can be inferred to what extent the correct analysis result has been obtained, in order to confirm them (Greco, Słowiński, & Szczęch 2016). However, this kind of confirmation emphasizes how relevant the learning behavior is to the analysis results, rather than the reasons underlying the behavior. Kennedy and Terry (2004) pointed out that cognitive components in user log data make it difficult to match a kind of behavioral pattern or strategy with a kind of behavior reason. There may be multiple reasons for the same behavior (Misanchuk & Schwier, 1992).

The main learning analytics framework proposed to date are: “Wisdom-Knowledge-Information-Data” (Baker, 2007), “Five Steps of Analytics” (Campbell, DeBlois, & Oblinger, 2007), “Web Analytics Objectives” (Hendricks, Plantz, & Pritchard, 2008), “Collective Applications Model” (Dron & Anderson, 2009), “Processes of Learning Analytics” (Elias, 2011), and “Learning Analytics Processes” (Chatti, Dyckhoff, & Schroeder et al., 2012). It is evident that the processes in learning analytics frameworks lack the indispensable confirmation step, which is a way to provide a structured process for the learning behavior analysis (Campbell, DeBlois, & Oblinger, 2007). Based on previous research, it is easy to see that all of these models usually consist of data collection, data processes, data analysis, and data application. The learning analytics frameworks have become more sophisticated, but they all still lack the analysis result confirmation step. It is therefore difficult to identify the reasons behind the learners’ behavior patterns for causal relationship exploration and strategies. It also leads to ambiguous guidance for the application of causal relationship exploration.

3.3 Causal relationship exploration process in Education Big Data mining

As shown in Figure 1, a process of causal relationship exploration was proposed, which consists of five main steps: data collection, data processing, data analysis, result confirmation, and result application, has a cyclic and iterative structure. Based on the two frameworks presented by Chatti et al. (2012) and Siemens (2013) and Kolb's Experiential Learning Cycle learning theory, we propose a Result Confirmation-based Learning Behavior Analysis (ReCoLBA) process in this study.

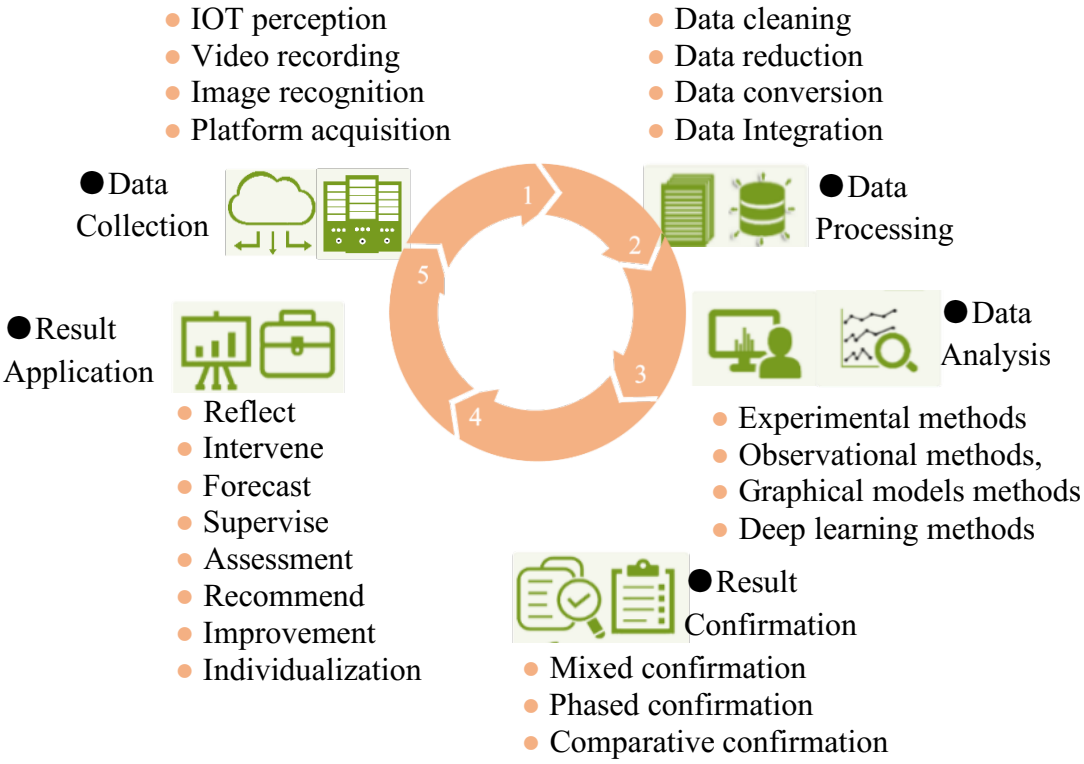

Figure 1. The ReCoLBA process

By comparing and combining the aforementioned models adopted by the ReCoLBA, while clarifying the differences and similarities between them, we highlight the innovative functions of this model in Table 2. The most important functions of identifying and confirming the reasons underlying the learning behavior are obvious. Moreover, an important education theory, Kolb's Experiential Learning Theory, was introduced into the ReCoLBA model. The integration of that theory provides a cyclic and iterative structure for this model. Subsequently, a specific introduction of each step will be provided.

Table 2. The differences and similarities between the reference models and the ReCoLBA model

Name	Learning Analytics model	Processes of Learning Analytics model	ReCoLBA model
Steps	Data collection	Collection	Data collection
		Storage	
	Pre-processing	Data cleaning	Data processing
		Data Integration	

	Analytics	Representation and visualization	Data analysis
	None	None	Result confirmation
	Action and Post-processing	Action	Result application

The first primary function of this model is to collect data from different education environments, with the main data collecting methods, such as IOT perception, video recording, image recognition, and platform acquisition. According to the report "SEDCAR, Standards for Education Data Collection and Reporting" (National Center for Education Statistics, 1991), data in educational settings or systems can be collected and analyzed by means of record extraction, surveys (mail, telephone, face-to-face), observations, experiments, and secondary data analysis. Usually, the collection methods can be divided into two categories: designed experiments and the observational approach. The former means that collectors control the data generation conditions. The latter means that collectors do not participate in the data generation process (Kantardzic, 2011). At present, data collection technologies include the following four main technical categories (Wong, 2017): Internet-of-Things (wearable devices), video recording technology (video broadcasting), learning platform acquisition technology (log data), and image recognition technology (eye-tracking).

Data processing. The second data processing step includes data cleaning, data normalization, data transformation, data missing values imputation, data noise identification, and data integration (Romero & Ventura, 2010; García, Ramírez- Gallego, Luengo, Benítez, & Herrera, 2016). Data processing is a step that transforms raw data into a useful and efficient format (Chakrabarty, Mannan, & Cagin, 2015). The main tasks of data processing are to retrieve inaccurate records in the data set, identify incorrect or irrelevant records in the data set, and manipulate the collected data by deleting, modifying, and replacing (Wu et al., 2013).

Data analysis. The following step is data analysis, which consists of the three analysis goals of prediction, structure discovery, and relationship mining. Data analysis is an operation which is guided by certain research purposes, such as prediction, structure discovery, and relationship mining (Yin & Hwang, 2018). Each particular educational problem has its own specific objectives, so the existing analytical methods and techniques cannot be directly applied to the analysis of such data (Romero & Ventura, 2013). In other words, the analysis methods are not categorized into special research areas in the education field. The data analysis methods are divided into 11 methods according to three categories: prediction, structure discovery, and relationship mining (Yin & Hwang, 2018), as shown in Table 3.

Table 3. Goals and methods of data analysis

Goals	Prediction	Structure Discovery	Relationship mining
Methods	Classification	Clustering	Association rule mining
	Regression	Factor Analysis	Correlation mining
	Latent Knowledge Estimation	Knowledge Inference	Sequential pattern mining
		Network Analysis	Causal data mining

However, causal relationship exploration methods can be categorized into three types based on the principles of causal relationship exploration, experimental methods, observational methods, and graphical models based on machine learning and deep learning, as shown in Table 1.

Table 1. Three categories of causal relationship exploration methods

Category	Method	Advantages	Disadvantages
Experimental methods	Randomized controlled experiments	most widely used and highly valid	Selection bias
	Regression discontinuity design	Uses natural cutoffs, provides quasi-random assignment to simulate randomization	
	Difference-in-differences	Differential control for unobserved time-varying confounders	
Observation methods	Propensity score matching	Reduces selection	Unable to account for unobserved confounders
	Instrumental variables	Control for unobserved confounders	Relies on strong assumptions, violation of which can lead to biased estimates
Graphical methods	Bayesian networks	Help visualize and understand complex causal structures	Assume there is matching fidelity between data and causal graphs
	Neural networks	Capture complex nonlinear relationships	The result is unstable
	Structural equation modeling	Captures unobservable structure in causal relationships	Selection bias

Result confirmation. The obvious function point of the model which differs from any others is the result confirmation, which is made up of mixed confirmation, phased confirmation, and comparative confirmation. This study reviewed previous findings on the learning analytics models, and found that they were lacking a results confirmation step. To address the problems of confirming the analysis results, four representative research papers were identified related to how to conduct a study to confirm the analysis results. For example, to evaluate the online materials in the context of the unit of study, Phillips, Baudains, and Van (2002) confirmed existing results by using learning logs recorded in the WebCT site for students' approaches. In addition to quantitative approaches to confirming learning outcomes using statistical methods, such as comparison of the previous research in early years, an investigation was employed to confirm the situation about completing laboratories, ongoing study, and surface learning. Only relying on the understanding of the fact that the surface analysis using data-based learning analytics is insufficient to identify student learning behavior, a case study of how qualitative data provide rich information to confirm the analysis results was carried out by Phillips (2006), and a comparative confirmation using the data generated from the study subjects in different periods was used to discover the changing laws (Kennedy & Terry, 2004; Li et al., 2018). In response to the insufficient interpretation of data analysis for the learning behavior, they interviewed the learners to confirm the potential interactions after the data analysis. By carrying out investigations, this approach with phased confirmation steps can identify the potential reasons underlying the existing analysis, such as why students repeat the learning behavior.

Based on the specific confirmation methods presented in these four papers, this study summarizes them into the three categories of mixed confirmation, comparative confirmation, and phased confirmation:

Mixed confirmation. This is a method of confirming the analysis results through a combination of quantitative and qualitative studies. It not only includes using evaluation metrics of the analysis algorithm, but also involves some qualitative methods such as documentation, staff interviews, observation of students in laboratory classes, and a student survey (Phillips et al., 2002).

Comparative confirmation. There are two confirmation dimensions of the comparative confirmation method. The first is the time dimension. The comparison of the similarities and differences is used to discover the changing laws, by means of collecting the data generated from the study subjects in different periods. Data collection intervals are chosen, based on semesters or school years. The second is the spatial dimension. For example, students' learning behavior data from different schools are comparatively analyzed for results confirmation (Phillips, 2006).

Phased confirmation. The results confirmation involves two analytic stages. The statistical analysis is usually adopted in the first stage. If the variables of interest had occurred, or the analysis results had abnormal values, then they would move to the second stage. The focuses on interesting variables or abnormal values will be analyzed again. The confirming method includes interviews and questionnaires (Kennedy & Terry, 2004; Li et al., 2018).

Result application. The last one is the result application step which involves four stakeholders: learner, teacher, administrator and course designer. It is noteworthy that the final goal of the ReCoLBA model is to apply the analysis results to educational practice. Different roles can derive different benefits from the confirmed results of learning analytics. For example, the confirmed analysis results could help students share their learning experience, help teachers master students' learning behavior and provide them with timely support, help manager administrators to evaluate teachers and students, and help course designers to improve the course content and instructional materials according to the confirmed analysis results (Yin & Hwang, 2018).

3.4 Application of causal relationship exploration process in Education Big Data Mining: an example analysis of repeated learning behaviors

The last one is the result application step which involves four stakeholders: learner, teacher, administrator and course designer. It is noteworthy that the final goal of the ReCoLBA process is to apply the analysis results to educational practice. Different roles can derive different benefits from the confirmed results of learning analytics. For example, the confirmed analysis results could help students share their learning experience, help teachers master students' learning behavior and provide them with timely support, help manager administrators to evaluate teachers and students, and help course designers to improve the course content and instructional materials according to the confirmed analysis results (Yin & Hwang, 2018).

3.4.1 Experimental design for verifying the usability of the ReCoLBA process

Using the ReCoLBA process, we designed a case study including two experiments, one aiming to examine the usability of the model, and the other focusing on verifying the effectiveness of a learning strategy proposed by the confirmed analysis results.

To examine the usability of the proposed model, a ReCoLBA-based case study was conducted. To this end, 48 participants were recruited and assigned to read academic papers using an e-book system. There are five parts in this experiment, namely using the e-book system to collect students' reading behavior, performing five steps of data processing, adopting behavioral sequential analysis for data analysis, utilizing a mixed confirmation method to confirm the analysis results, and applying a learning strategy by an intervention.

Data collection in the case study. An e-book system was developed corresponding with the functions of the ReCoLBA process, to evaluate the model usability. As shown in Figure 3, this system can collect the students' reading behavior when reading learning materials, such as (1) page-turning, (2) zoom+/-, (3) writing a memo, (4) adding or deleting an underline, and (5) adding or deleting a highlight. Most importantly, all reading behavior actions were recorded in the form of reading action logs. Moreover, the e-book system can provide automatically coded action logs as one of the convenient functions to reduce subsequent data processing. In this way, the usual manual coding process completed by programmers can be avoided.

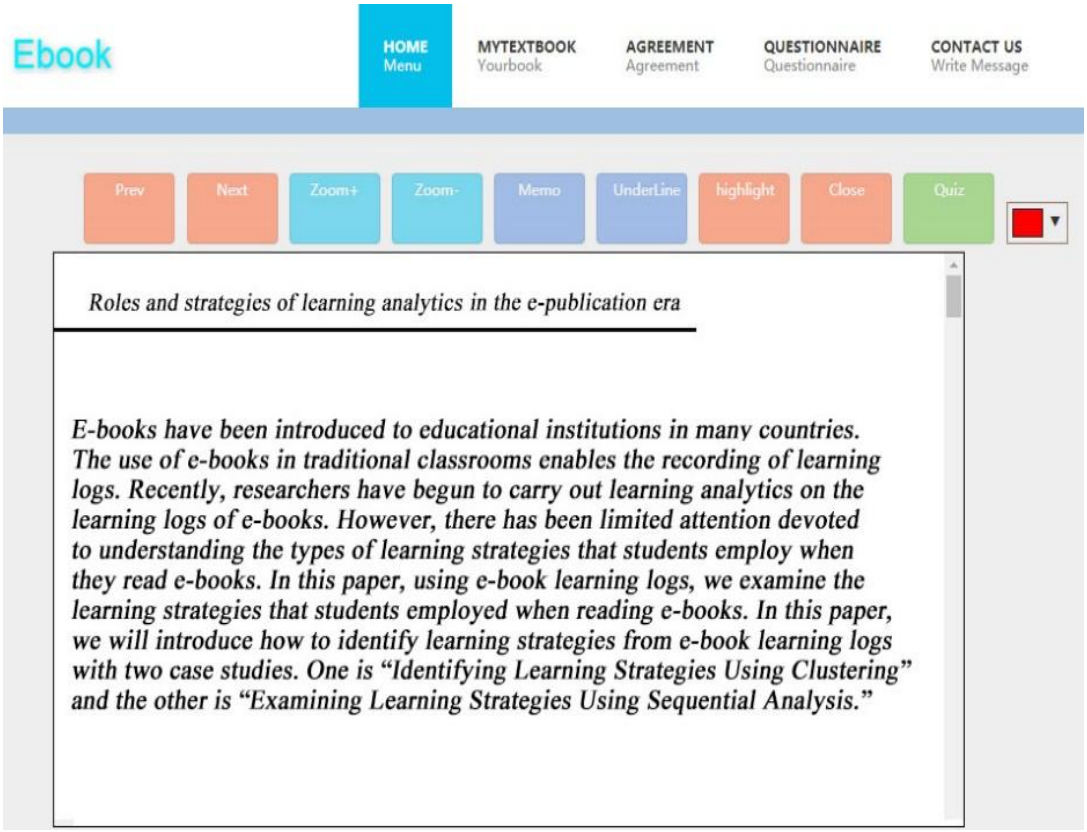

Figure 3. E-book system interface

A total of 4,748 records were collected from 60 graduate students, who were asked to complete reading academic papers in the e-book system within 90 mins in this experiment. Their reading behavior data, which consisted of user ID, operation name, page number, and action time, were stored in the database for analyzing their learning behavior patterns or strategies, as shown in Table 4.

Table 4. Sample action log

User ID	Operation name	Page number	Action time
Student 1	NEXT page:0	19	2020/7/10 12:05:18
Student 2	PREV page:8	16	2020/7/10 12:07:31

Data processing in the case study. This study summarizes data processing as the following four points. The first is data transformation. Through the statistical processing, the sum of the students' reading behavior including adding or deleting underlines, adding or deleting highlights, and adding or deleting memos, was respectively counted. The second point is data cleaning. If the following learning phenomenon or behavior occurred during the learning activity, it was viewed as an invalid record. For example, if the longest duration between two learning behaviors exceeded 20 minutes, then the record was invalid as it indicated that no reading activities had occurred because the student did not conduct any learning behavior within 20 minutes. Besides, incomplete records were filtered. The third point is missing values imputation and noise identification. The valid sample size changed from 60 to 47 by identifying and discarding missing values data and noisy data. The fourth point is data normalization. There were some data completed by those who had an invalid preview, for example, some students who completed the preview of the lesson (read the learning content before class) less than 3 minutes before the class. In that case, the preview of the lesson was viewed as invalid. The final point is data integration. The objective of data processing is mainly from the e-book system, so that the sole data generating source avoids the need to integrate data from different data sources, which makes it unnecessary to adopt data transformation and integration in the data processing.

Data analysis in the case study. In the case study, behavioral sequential analysis was adopted to gain a detailed understanding of progressive learning behavioral patterns (Bakeman & Gottman, 1997; Hou, 2012). The progressive learning behavior patterns were obtained by sequential analysis, as shown in Figure 3. The squares represent learning behavior, and the arrow-lines and numbers represent the direction and extent to which the behavior is associated with other behaviors. Rounded curves represent the association of the behavior with itself. The result is significant at the $< .05$ level when the z-value is greater than 1.96 (Bakeman & Gottman, 1997; Hou, 2012).

In Figure 4, it is obvious that "HIGHLIGHT" is mutually associated with "DEL HIGHLIGHT" and "DEL UNDERLINE," and "UNDERLINE" is mutually associated with "DEL HIGHLIGHT" and "UNDERLINE." Meanwhile, "HIGHLIGHT," "DEL HIGHLIGHT," "DEL UNDERLINE," and "UNDERLINE" are respectively associated with themselves. "BOOKMARKER" is mutually associated with "DEL BOOKMARKER." Finally, "BOOKMARKER" has a one-way association with "UNDERLINE" and "MEMO." The association between this repetitive behavior was confirmed to be significant, using the Z-score binomial test parameters in the sequential analysis method. However, this is not sufficient to understand why these patterns and strategies occurred, and especially to understand why some students exhibit some repetitive behavior of the same operation. For example, some students deleted UNDERLINE after adding an UNDERLINE, and some deleted the HIGHLIGHT after adding a HIGHLIGHT.

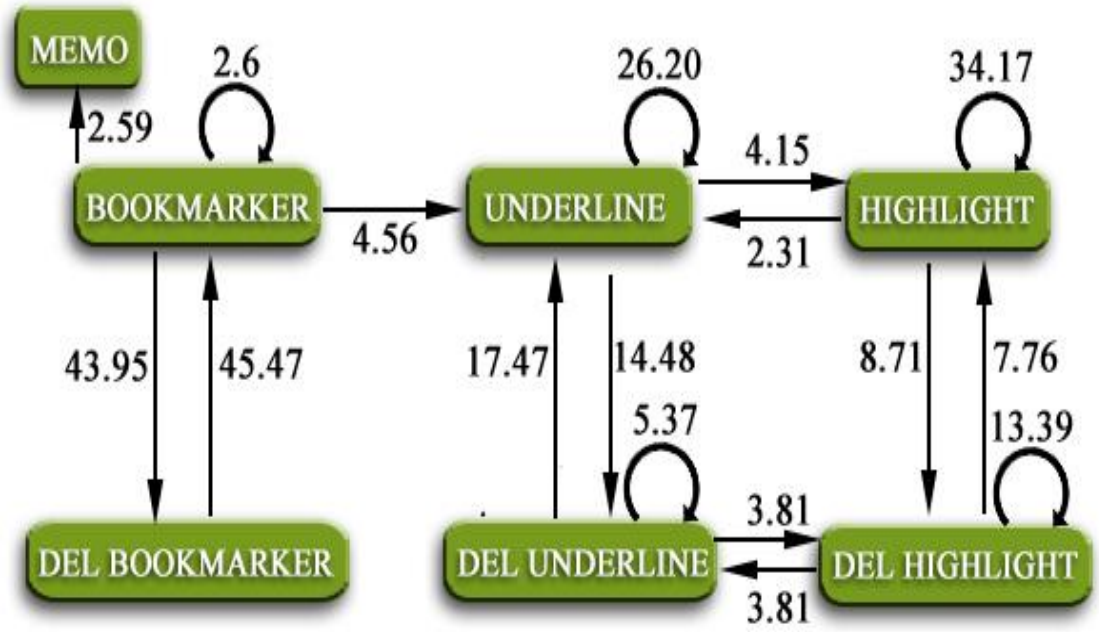

Figure 4. The students' progressive learning behavioral patterns

Resulting confirmation in the case study. We identified the repetitive learning behavior patterns as "Deleting a HIGHLIGHT after adding the HIGHLIGHT," "After adding an UNDERLINE, delete the UNDERLINE." However, it was difficult to understand the reason why these behavior patterns happened. Therefore, it was necessary to conduct the result confirmation step. To this end, an investigation system was used to investigate the participants who had repetitive learning behavior patterns, as shown in Figure 4. The participants were investigated to answer the questions that matched their behavior such as, "Why did you adopt "After adding a HIGHLIGHT, delete the HIGHLIGHT," and why did you adopt "After adding an UNDERLINE, delete the UNDERLINE." Samples of extracts from their answers are shown in Table 5.

Table 5. Samples of interviewees' answer content

Interviewee	Answer
Student 1	When I first read the paper, I highlighted the content that I thought was important; however, going back to read it a second time, I realized that it was not the most important content anymore, so I decided to delete it.
Student 2	It is useful to deepen my understanding of the paper.
Student 3	When I read the sentences, from my perspective, it is more suitable for the question.
Student 4	After adding the highlight, I subsequently added an underline, which is my way of distinguishing the important content.
Student 5	That is a mistake in the highlight position owing to not being proficient in operating the platform.

In this investigation, only if common results were gained by two researchers was the coding of the students' answers accepted. The notion for the data analysis of up to 24 participants was extracted, and the reference basis is strong. After the investigation, it was confirmed that it was difficult for some students to identify which were the keywords while reading. Other students often read repeatedly to understand the main idea of the papers because they had difficulty

finding the most important content. It was concluded that the repetitive operation of markers happened because they could not correctly identify the learning emphasis. Besides, the marking method, such as underline and highlight, is beneficial for students to mark the important content, and these choices are mainly related to personal interest.

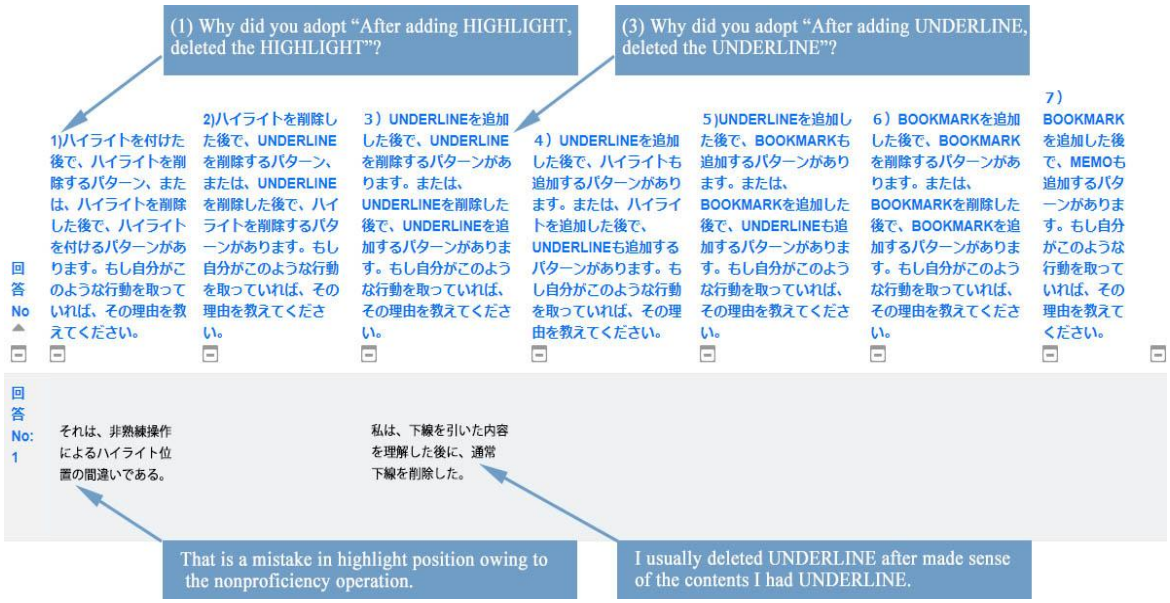

Figure 5. The investigation system interface

3.4.2 Results application in the case study

Through the ReColBA-based case study, it was found that the reason why students exhibit the repetitive behavior of the same operation is that they could not find the learning emphasis. The e-textbooks lack any marking of the learning emphasis, which is not conducive to students' understanding. Therefore, we proposed a learning strategy that marked the learning emphasis in the e-textbooks using underlines and highlights, as shown in Figure 6.

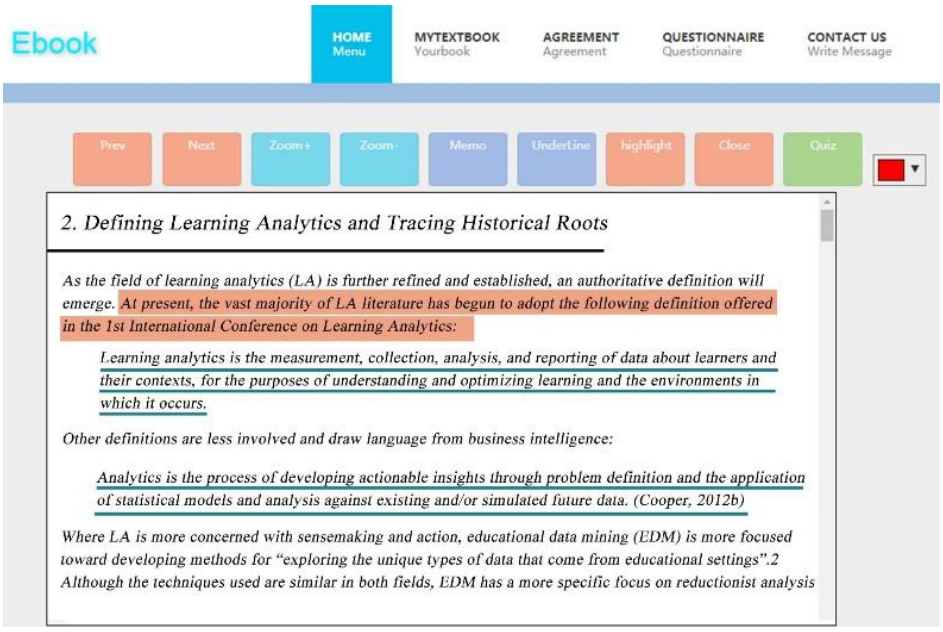

Figure 6. The e-textbook with marking of the learning emphasis

3.4.2.1 Experiment for verifying the learning effectiveness of the ReCoLBA process

To verify the learning effectiveness of the learning strategy discovered by the model, we conducted a verification experiment. The specific steps are as follows.

Participants. A total of 80 graduate students from a university's graduate school were recruited to participate in this experiment. As the experimental group, 43 students read an e-textbook via the e-book system, using the marking learning emphasis method. Another 37 students in the control group adopted the same experiment conditions, except that they did not use the marking learning emphasis technique.

Measuring methods. The measuring methods consisted of a pre-test, a post-test, and an interview. Besides, the frequency and duration of learning behavior which were automatically recorded by the e-book system were used to evaluate the effectiveness of the learning strategy. Firstly, the pre-test was to evaluate whether both groups had equivalent prior knowledge of the upcoming learning content. It included 10 multiple-choice items. Secondly, the post-test was to measure whether the marking learning emphasis method was helpful for students' learning achievement. It also included 10 multiple-choice items, which were related to the core learning emphasis of the article. The total score was 100 in both the pre-test and post-test. Thirdly, a 30-minute semi-structured interview was conducted for the experimental group which was recorded verbatim. The following questions were asked: What do you think of this learning method? What is the difference between this way of marking the key points of the paper and your previous learning method? Two researchers were invited to analyze the interview data, and to determine the participants' attitudes towards the proposed learning strategies through the extracted core keywords from the interview content.

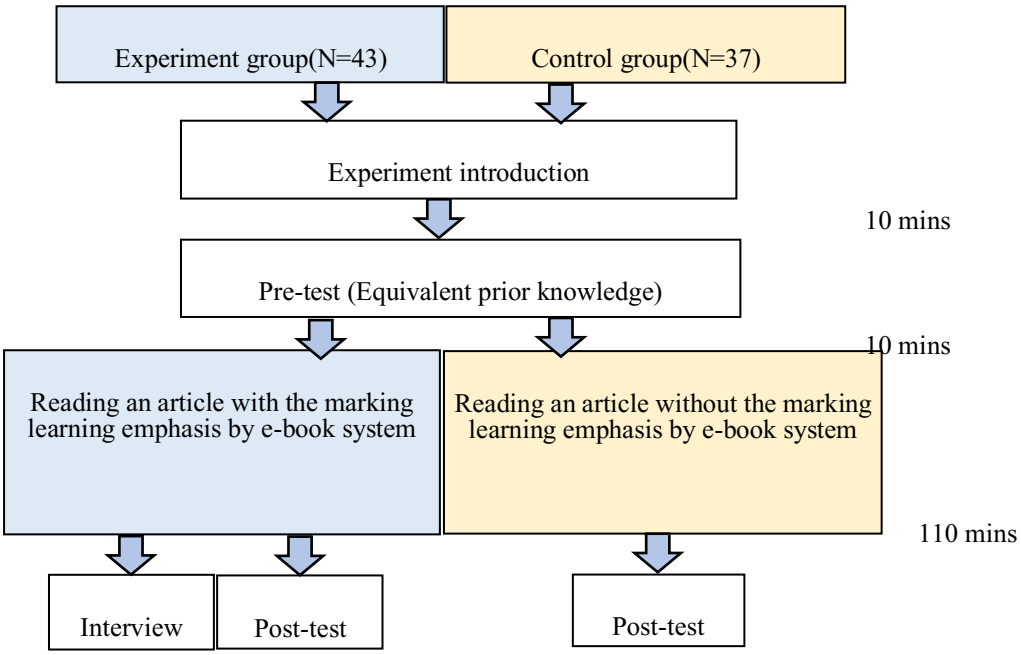

Figure 7 Experimental procedure

Experimental procedure. The learning emphasis aimed to help students understand the definition, historical development, and application fields of Learning Analytics, which is a unit of an Education Technology course in the university.

As shown in Figure 7, before the experiment, the students were given a pre-test to test their prior knowledge of the learning emphasis. Subsequently, we introduced the learning emphasis and e-book system, then conducted a 110-minute learning activity. Both the experimental group and the control group were required to use the e-book system functions at will, such as page-turning, underlining, highlighting, and making memos. Unlike the control group, the e-textbook for the experiment group was already underlined and highlighted, to indicate the learning emphasis. After the learning activity, a post-test was conducted, and the experimental group students were asked to take part in an interview.

3.4.2.2 Experiment results

First, analysis of learning achievement. The effectiveness of the proposed method was examined to ensure whether this approach could improve the students' learning performance. Based on the pre-test, the standard deviations and mean values were 20.471 and 60.00 for the control group, and 15.596 and 78.92 for the experimental group. According to the *t*-test result (F=2.897, p>0.05), as shown in Table 6, there was no significant difference between the two groups. It was therefore found that they had equivalent prior knowledge before the learning activity.

Table 6. Descriptive data and t-test result of the pre-test

Variable	Group	N	Mean	Std. Deviation	*F*
Pre-test	Experimental group	43	60.00	20.471	2.897
	Control group	37	78.92	15.596	

The post-test was taken after the learning activity. We used the one-way analysis of covariance (ANCOVA) to evaluate the learning achievement of the two groups, setting the groups as a fixed factor, the pre-test scores as the covariate, and the post-test scores as the dependent variable. The Levene's test of equality of error variances (F=1.409, p>0.05) indicated that the assumption of regression homogeneity was followed. The results of the ANCOVA (F=18.424, p<0.01) in Table 7 determined that the experimental group achieved better results. It was concluded that the marking learning emphasis method was beneficial for the students to understand the learning emphasis.

Table 7. Descriptive data and one-way ANCOVA result of the post-test

Variable	Group	N	Mean	Std. Deviation	*F*
Post-test	Experimental group	43	88.60	13.378	18.424*
	Control group	37	77.62	11.526	

*p < .01.

Secondly was the analysis of the system reading time. To better understand the effectiveness of this method, the ANCOVA was also used to analyze the reading time of the two groups. The Levene's test of equality of error variances (F=1.409, p>0.05) was not violated. From the ANCOVA results in Table 8, it can be seen that the means and standard deviations of the experimental group are 0:41:06 and 0:21:21, and for the control group they are 1:12:27 and 0:26:26. There was a significant difference (F=41.731, p<0.01) between the two groups. It was clear that the experimental learners spent less system reading time finishing the learning task than those who used the conventional method. It was found that it was helpful to save time compared with the students who did not use the marking learning emphasis method.

Table 8. Descriptive data and one-way ANCOVA result of reading time

Variable	Group	N	Mean	Std. Deviation	*F*
Reading time	Experimental group	43	0:41:06	0:21:21	41.731*
	Control group	37	1:12:27	0:26:26	

*$p < .01$.

Thirdly was the analysis of the interviews. After the experiment, an interview was conducted on the topic of how to evaluate the learning methods, and 43 interview transcripts were obtained. It was concluded that 42 participants gave positive evaluations and 1 participant had a neutral attitude. Two dimensions constitute a positive evaluation, including Knowledge comprehensibility (27 participants) and Reading convenience (32 participants). In all, 27 participants used the key items "Good for memorizing content" (14 participants), "Strong visual guidance" (4 participants), "Helps me understand the content" (6 participants) and "Convenient for future work" (3 participants) to describe what benefits it brought. In terms of reading convenience, 32 participants shared their positive evaluations. Four participants mentioned that this method could increase reading convenience, 18 argued that they could read better with the help of the reading emphasis marked by underlining or highlighting, and 10 maintained that the efficiency gains and time saving were the most important benefits of the method.

3.4.2.3 Conclusions

The existing research clearly indicates that the analysis results have the hidden probability of creating misunderstandings of students' behavior in practice, especially in the case of lacking result confirmation. Owing to the cognitive element and multiple interpretations of learning behavior, result confirmation requires not only providing the patterns or strategies for learning behavior, but also identifying the reasons behind their occurrence, and only by doing so will the analysis results be better applied in practice.

This study proposed the ReCoLBA process for learning behavior analysis. Compared to the previous models, the ReCoLBA process added a result confirmation step including three main methods: mixed confirmation, phased confirmation, and comparative confirmation.

A case study was then conducted to verify the feasibility of the ReCoLBA process. Based on the findings concluded from the application experiment, the ReCoLBA process has been successfully verified. It can help identify why certain learning behaviors occur and the strategies adopted, and avoids applying analysis results without confirmation. The interview in the result confirmation step was different from that usually employed in the domain of human science. This interview method is combined with the analysis technique, and the subjects that needed to be studied were based on the primary analysis results, which was deduced from the data.

The confirming function designed in the ReCoLBA process enables researchers to have an opportunity to access the reasons underlying the learning behavior. Through the application of the model, it was found that the reason for the students repeatedly adding and deleting underlines or highlights was the lack of learning emphasis. Based on the findings, an e-textbook with the learning emphasis marked with underlines and highlights was designed and developed. To verify the effectiveness of the revised learning method of previously giving the learning emphasis, an experiment was conducted. From the results of the *t*-test, one-way ANCOVA, and interviews, it was obvious that the marking learning emphasis method could help students improve their learning achievement and save time when using the revised e-textbook.

4. Data prerequisites for causal relationship exploration: behavior identification by multidimensional scales

As mentioned earlier, the complex confounding variables and non-linear features in Educationak Big Data pose many challenges to causal relationship exploration. First, facing data with confounding variables and non-linear features increases the complexity of identifying and extracting learning behaviors in the context of Education Big Data mining (Breiman, 2001), which can easily lead to inaccurate subsequent causal relationship explorations. Hernán and Robins (2020) argued that it is very difficult to differentiate between true behavioral patterns and behavioral patterns influenced by confounding variables. Second, data with nonlinear characteristics puts tremendous analytical pressure on behavior identification methods and causal relationship exploration models that rely on traditional linear models. This pressure manifests itself in the fact that complex nonlinear data make the behavioral patterns contained in the data itself extremely complex (Hastie et al., 2009). At the same time, this increases the need for more sophisticated methods of causal relationship exploration to separate true behavioral information from confounding information (Pearl, 2009). Third, in order to mitigate the challenges of behavior identification and causal relationship exploration caused by confounding variables and nonlinear features, Yao et al. (2020) after surveying current causal relationship exploration methods, suggest robust methods that can address confounding effects and capture nonlinear relationships, such as machine learning or deep learning methods.

4.1 Methods for identifying learning behaviors in Education Big Data mining

4.1.1 Understanding behavioral identification in Education Big Data mining

First, the understanding of behavioral identification can gain different insights from the medium in which the behavior occurs, the supporting technology, and the way it is presented in the context of Education Big Data mining. Baker, Martin & Rossi (2016) refer to the scope of quantifying learning behaviors in their discussion of models and methodologies when mining educational data. They include observable activities in the data, patterns of student engagement in activities, and interactions with learning materials in the scope of quantifying behaviors. Second, regarding the types of behavioral identification methods, Romer and Ventura (2010) suggest the analysis of data sources and metrics for quantifying behaviors include quantitative methods such as time of engagement, mouse-click manipulation, learning participation, social interaction, and assessment of learning performance.

Third, the importance of behavioral identification is mainly reflected in the fact that it provides data support for educational data mining and determines the accuracy of analysis results (Slavin et al., 2011). As the information input of data analysis, behavioral identification objectively describes the behavioral characteristics and patterns of learning subjects in the learning environment and determines the accuracy of data mining. In addition, behavioral identification plays a mediating role in causal analysis. Hayes (2017) argues that quantified learning behavior variables play a mediating role in causal relationship exploration models, from which insights can be gained into the characteristics, patterns, and mechanisms by which interventions affect learning outcomes. Without quantifying learning behaviors, although we can obtain causal relationship exploration results, to that point it is not possible to identify the detailed information needed to identify the specific behaviors or interactions that influence educational outcomes. Quantified learning behaviors are the ones that allow us to identify and measure the exact relationship between student behavioral characteristics and patterns and outcomes.

4.1.2 Four methods to identify behavior

There are four main methods to quantify behavior in the context of Education Big Data mining: learning engagement, learning performance, collaborative interaction, and assessment of cognitive processes.

First, the identification of learning engagement. Siemens and Baker (2012) proposed learning engagement metrics to quantify behaviors, which are defined as the frequency, duration, and intensity of students' interactions with learning vehicles, learning tasks, and learning resources. Engagement is an important indicator of the learning process because it reflects students' affective, attentional, and cognitive engagement in the learning process (Fredricks et al., 2004). At the same time, learning engagement can also have a positive impact on learning behaviors, as well as being influenced by the reverse from learning achievement and learning satisfaction. Currently, specific methods to quantify learning behavior based on learning engagement include, firstly, relying on established learning management systems, such as moodle, to record information about students' logins and actions on learning materials. In addition, the mouse, keyboard and other sensors to obtain the click rate, access frequency, and line-of-sight information distribution data.

Second, identification of learning performance. The most commonly used quantitative indicators and methods for assessing student performance are test scores and completion rates of test procedures. Romero and Ventura (2010), after analyzing the latest techniques in Education Big Data mining, concluded that performance indicators are the most common and effective objective measures of learning comprehension, mastery, and application of learning tasks and objectives. The specific forms of performance are very rich, ranging from traditional subjective and objective test forms to tests using the latest audio-visual technologies.

Third, the identification of collaborative interaction. Collaborative learning is one of the most important components of the learning methodology, and is a means for students to obtain emotional identification, distribution of attention, and self-assessment of cognitive processes in the learning process (Cabrera et al., 2002). In the context of Education Big Data, there is an increasing variety of ways to realize collaborative learning. For example, Google's Collaboration Space allows students to discuss, edit and share documents (Wang et al., 2012). Meanwhile, forums, blogs, and communication tools linked to social media provide a humane way to discuss, feedback, and share collaborative information (Junco, Heiberger, & Loken, 2011). In addition, new methods of collaboration are emerging in the form of virtual learning and artificial intelligence-assisted tools. Examples include role-playing and virtual interaction tools for students in virtual teachers, which take full advantage of immersive environments to the fullest extent (Dalgarno & Lee, 2010). Dillenbourg and Traum (2006) suggest that forum posts, interactive comment texts, and group discussions in social media can be used as indicators for the identification of collaborative interactions.

Fourth, the identification of cognitive processes. The assessment of cognitive processes has long been an important research topic, including how to assess cognitive changes in students' learning processes, regulation, and metacognitive assessment, which influences the analysis of educational activities and learning behaviors. Winne (2010) emphasizes the importance of quantifying cognitive processes, and identifies relationships between quantitative metrics of cognition and learning outcomes and thinking processes, as well as self-monitoring skills. In terms of realization methods, Winne utilized self-assessment to monitor the use of cognitive

strategies and to quantify cognitive processes. He found that Likert scales were most commonly applied with self-report surveys. Meanwhile, video recording became one of the most convenient and efficient means when conducting monitoring cognitive strategies. Tracking changes in cognitive strategies is more challenging to quantify than the first two. And, Winne found that using a learning management system or other learning software to obtain information about students' logs, such as click-through rates or tracking system logs, can be used in conjunction with statistical or machine learning algorithms. Based on this information, along with statistical or machine learning algorithms, it is possible to understand how students interact with learning environments, tasks, and materials, and how their cognitive strategies change.

4.2 Requirements for causal relationship exploration to identify behavior

There are some challenges for identifying behavior: 1) EBD provides rich objective data, expanding the scope and methods of identifying behavior. Although current researchers use objective data such as log data to identify objective behaviors such as learning engagement, learning performance, and collaborative interaction. However, there are also significant limitations in current methods of identifying behavior, namely the inability to use objective data to identify subjective behaviors, such as reflecting student emotions, attention, and cognitive development. 2) when exploring causal relationships in the context of EBD, identifying subjective behavior through methods such as questionnaires can easily lead to low identification accuracy. Besides, subjective behavior identification not only needs to solve the challenges of low accuracy caused by traditional methods such as questionnaires, interviews, and observations, but also requires the use of natural language processing, image recognition, and eye tracking technologies to capture more accurate subjective data. 3) The most important thing is identifying reading patterns from individual actions. Therefore, as a data prerequisite for exploring causal relationships, behavior identification needs to improve the accuracy of subjective behavior identification.

4.2.1 Growing challenge of behavioral features in the digital environment

Due to the widespread use of online technologies and e-publishing standards, digital environment has continuously contributed to the transformative shift in reading behavior (Birkerts, 1994), in which an increasing amount of time spent on keyword spotting, non-linear reading, and shallow reading instead of in-depth reading. The extensive convergence of video, audio, and images (Yin et al., 2019) as substitutes for print-based textbooks (Yin & Hwang, 2018) has a great potential impact on reading behavior (Zhao et al., 2021). This change, along with hypertext fragmentation and website navigation (Fitzsimmons et al., 2020), threatens favorable reading behavior. It leads to shallow reading (Liu, 2005) and sustained attention decline, whereas attention is more easily focused in the printed environment characterized by linear reading (Walsh, 2016). It is particularly challenging to consider the lower comprehension and critical reflection skills, which are characteristics of shallow reading (Sarris, 2020).

4.2.2 Key issues in identifying learning behaviors in causal relationship exploration

There are several key issues that need to be addressed in the identification of learning behaviors during causal relationship exploration in the context of Education Big Data. First, the precision of identification of learning behaviors. Data granularity refers to the degree of how much detail is contained in the data. Han, Kamber and Mining (2006) understood data granularity in the context of Education Big Data as the detail or scale at which the data represent behavior. In other words, the extent to which the data is able to break down the learning behavior into smaller,

more specific or observed outcomes. This kind of behavioral identification based on high granularity data can maximally summarize the characteristics and patterns of learned behaviors and provide a high precision understanding of causal relationship explorations. Therefore, only high-precision behavioral identification can ensure high-precision causal relationship exploration.

Second, the difference between subjective and objective behavior identification. Learning behaviors can be broadly classified as objective and subjective behaviors, which show similarities and differences in data accuracy, collection ease, and analysis needs. As shown in Table 9.

Table 9. Differences in subjective and objective data

Data categories	Data accuracy	Collection difficulty	Analysis of requirements
Subjective data	Relying on independent reporting, easily mixed with bias, memory bias,	Long cycle, manual operation	Obvious fusion trend
Objective data	Observable, measurable	Automatically collected, non-interference in learning	

Objective behavior refers to an observable and measurable behavior, which can be manifested in physical activity, manipulation of objects, and interaction with the learning environment and materials. The identification of this objective behavior relies heavily on tracking the behavior. According to Ferguson (2012), when quantifying objective behavior, methods based on logs and sensor data have a greater match with objective learning behaviors in an Education Big Data environment. Among them, click-through rates, time-stream based logs can be used to track and record as well as quantify learning behaviors. Quantifying and identifying subjective behavior are more complex than objective behaviors. Winne (2010) used self-reports, Likert scales, and structured discourse questionnaires to collect students' mental activities, affective changes, and cognitive development during the learning process when addressing the problem of measuring self-regulated learning.

Third, the historical bias in the identification of objective behavior and the trend of integrating subjective and objective behavior. Behaviors quantified and identified during causal relationship exploration have historically tended to be objective behaviors, dominating Education Big Data mining. This is because objective behavior is inherently observable and measurable (Ferguson, 2012). Then, relying on quantifying objective behaviors alone cannot adequately explain increasingly complex learning behaviors, especially when causal relationship explorations are confronted with data with nonlinear characteristics that contain confounding variables. According to Winne's work (2010), as online learning or self-directed learning methods continue to be incorporated into students' everyday learning, a range of subjective learning, such as self-regulation, motivation discovery, optimizing self-efficacy, and facilitating metacognitive strategies, have attracted research attention. This shift in focus indicates that scholars are increasingly aware of the importance of subjective behaviors in shaping learning outcomes. On the other hand, this is also due to the development of natural language processing, image recognition and eye movement technology to capture emotional changes. Dawson and Siemens (2014) found that integrating subjective and objective behaviors into Education Big Data makes it easier to gain a comprehensive understanding of learning behaviors, and this idea has also gained more and more attention from scholars.

Fourth, the identification of learning behaviors supported by information technology needs to consider the impact of technology on learning behaviors. Taking reading behavior as an example, the development of online reading platforms and e-book resources has brought about a shift in learning behavior. Chen and Catrambone (2015) discuss how e-book technology affects changes in reading behavior. They argue that the shift from physical textbooks to e-books not only changes how students interact with the material, but also how behaviors are captured. Lai and Hwang (2016) found that the shift from manual page-turning to digital interactions through the mouse and keyboard provided new opportunities to quantify reading behaviors. In addition, digital interaction methods such as scrolling, clicking, highlighting, and note-taking are emerging as important behavioral characteristics that describe online reading behavior (Yin et al., 2019). Liu (2005) analyzed the recent developmental changes in reading behavior and argued that digital contexts have a significant impact on shaping digital learning behaviors. Specifically, different media contexts produce different reading patterns and influence different learning interactions and engagement behaviors.

4.3 Behavior identification for causal relationship exploration in Education Big Data mining: an example analysis of shallow reading behavior identification

4.3.1 Behavior identification model for shallow reading behavior

4.3.1.1 Definition, characteristics, and identification of shallow reading behavior

Definition of shallow reading. The current definition of shallow reading is established through an analysis of reading forms concerning the methods and mediums used, as well as the reader's perceptual experience during the reading process. It has been referred to as power browsing, bouncing, and squirreling. Similarly, shallow reading is described as a graphic, skipping style of reading (Kucirkova, 2019). Key attributes of shallow reading encompass browsing and scanning, keyword spotting, and one-time reading (Liu, 2005). These reading methods differ from browsing newspapers and skimming books, which have evolved to become the ultimate objectives of reading rather than mere intermediate strategies (Adler & Van Doren, 2014).

Second, shallow reading is not an isolated behavior and is integrated with computer-assisted reading, active reading, and reading comprehension metacognition in the online digital context. Computer-assisted reading incorporates sensory-rich materials like animations, sounds, and documents, mimicking the manipulative habits of paper-based reading, such as underlining and tabbing, to enhance reading functions (Zouaghi, 2016). However, this integration also introduces reading attention distractions, contributing to the vulnerability of poor comprehension monitoring during reading, leading to shallow reading (Baker & Beall, 2014). From a metacognitive standpoint, poor comprehension monitoring manifests as insufficient intrinsic motivation and ineffective self-regulation of reading plans and strategies (Pintrich & Zusho, 2002). Computer-based systems offer contextual reading cues guided by assistive technology and sensory stimulation (Morineau et al., 2005). Over time, decreased reading self-regulation leads to shallow reading behavior such as rapid scanning, reduced contemplation, and memory consolidation (Loh & Kanai, 2016). This transformation in reading results in a shift in the purpose of reading, where decoding of text supersedes the process of meaning acquisition (Myers & Paris, 1978). Therefore, the read no longer consciously engages with the text and thinks about it during reading, as observed in active reading. Instead, the focus remains on primary reading and information checking, encompassing the basic and key elements of texts, as well as assessing the cover, title, and other text components (Porter-O'Donnell, 2004).

Behavioral features of shallow reading. Distinct from the abstract definition of shallow reading, behavioral features represent externally observable and quantifiable manifestations. The behavioral features of shallow reading can be characterized by three dimensions consisting of frequency, duration, and quantity (Martin & Pear, 2019). Especially, the behavioral features of shallow reading can be characterized by three dimensions: reading speed (frequency: how often a learner can process information), reading dwell time (duration: for how long can they do it), and keyword reading (quantity: how much information can they process). The most obvious feature of shallow reading is that reading speed tends to encourage learners to acquire large amounts of information extensively (Faisal et al., 2012), but only at a superficial level (Miller, Kargin & Guldenoglu, 2014). Reading speed enhances the ability to avoid information overload without compromising deep comprehension and retention of content and allows one to keep up with the amount of content available (Alrizq et al., 2021). Another behavioral feature related to shallow reading is keyword reading, where learners tend to locate the required information with the least amount of reading (Shapiro & Waters, 2005). As a result, keyword reading is a reading-biased choice. The most obvious feature of shallow reading in the reading dwell time dimension is page jumping. Page jumping is a movement operation for extensive reading during a short dwell time, by clicking a hyperlink and scrolling the redirected page up or down (Hauger et al., 2011). Carroll (1971) demonstrates that the level of detail in reading is proportional to the jump dwell time and affects the level of gaze stability. In other words, the longer the jump dwell time, the more focused the gaze is on a specific location and the easier it is to read carefully.

4.3.1.2 Identifying methods and challenges for shallow reading

Several studies have been conducted to identify shallow reading behavioral features, as shown in Table 10. Typically, collecting data is the primary step in quantifying the behavioral features of shallow reading. For example, Liu (2005) explored reading behavior using a survey questionnaire to examine the change in participants' reading behavior, such as increasing reliance on electronic reading material, a tendency toward skimming and fragmented reading, and difficulty concentrating. The results confirmed the proposed hypothesis that shallow reading is a feature of hypertext-centered reading. Notably, in contrast to small-scale surveys, survey instruments of learning behavior have been gradually adopted in large-scale settings, mainly including global questionnaire tests and reading diaries to document reading activities (Locher & Philipp, 2023). Unlike Questionnaire-based Data Collection (QDC), Carr (2011) determines the inevitable consequences of the digital environment on modern psychology by employing a literature review. Two notable features of shallow reading behavior are fragmented reading and page jumping. Despite similar data collection, Delgado et al. (2018) preferred meta-analysis. Results showed a significant reduction in reading comprehension due to participants' propensity for shallow reading.

Table 10. Data collection and behavior identification of shallow reading

Research	Data collection			Behavior identification	
	Category	Method	Feature	Method	Subject
Liu (2005)	QDC	Questionnaire	Subjective	Content analysis	Behavioral factors
Carr (2011)	MDC	Literature database	Subjective	Thematic analysis	Behavioral features
Delgado et al. (2018)	MDC	Literature database	Subjective	Meta-analysis	Learning impact

Two issues exist in the definition of shallow reading in the aforementioned research. There are abundant subjective data, such as non-numerical and textual data. Although there are data collection categories, including QDC, Manual Data Collection (MDC) using professional tools, and Automatic Data Collection (ADC) represented by the e-book system (Yin & Hwang, 2018), objective data quantifying is lacking. When probing into the causative factors of shallow reading, questionnaires designed with behavior-oriented judgment or literature reviews have usually become the primary choice. Therefore, the definitions of the behavioral features could vary significantly due to the differences in self-reported questionnaires and objective behavior, potentially leading to misleading identifications (Kormos & Gifford, 2014). The qualitative analysis includes content, thematic, and meta-analysis. Such open-ended methods may make reaching objective conclusions difficult in contrast to mathematical models.

In summary, the challenges encountered in defining shallow reading are presented by clearly defining the behavioral features in three dimensions (Cooper et al., 2007) and collecting objective rather than subjective data, such as motivation, thoughts, or feelings (Leshner & Pfaff, 2011). To compensate for the drawbacks of defining shallow reading in existing research, a behavior identification model incorporating three dimensions was designed. A major contribution of this model is the exploration of the relationship between shallow reading and learning performance, particularly with respect to accurate evidence supported by subjective and large-scale behavior log data.

The research questions (RQs) to be addressed in this study are as follows:
RQ1. How to design a behavior identification model for precisely identifying shallow reading behavior, especially about the three behavioral dimensions: Reading Speed (RS), Reading Dwell Time (RDT), and Reading Word Count (RWC)?
RQ2. Based on objective data, what kind of relationship exists between shallow reading and learning performance? What are its implications?
RQ3. What learning strategies can be used to precisely mitigate shallow reading behaviors and consequently improve learning performance?

4.3.1.3 The dilemma of identifying shallow reading behaviors

The information to which learners are exposed has boomed exponentially, most of which is available online. Online learning has grown with the widespread adoption of internet hardware devices, growing sophistication of learning environments, and continuous related pedagogical attempts (Yin & Hwang, 2018). Challenges due to behavioral shifts in online learning, such as distraction (Rose, 2010), interrupted comprehension (Coiro, 2014), memory deficits (Simamora, 2020), and cognitive stress caused by information overload (Schommer-Aikins & Easter, 2018), have emerged. Accordingly, many studies have assessed online learning performance. That is, they calculate the effect size of online learning in contrast to face-to-face learning (US Department of Education, 2009). Notably, online and face-to-face learners differ mainly in behavioral features—preference for shallow vs. deep reading (Carr, 2011).

Shallow reading has become a core trend in online learning. It is related to learning strategies and performance. For example, it may reduce text comprehension (Miller, Kargin & Guldenoglu, 2014) and make it harder to find the thesis (Shapiro & Waters, 2005), compared with deep reading. Nevertheless, the question of how to accurately define shallow reading in terms of its behavioral features, remains. First, the behavioral features of shallow reading may be differently correlated with complex behavioral factors, such as dwell time, reading speed, and keyword reading (McLean, 2019), which increase the difficulty of defining it and analyzing

its impact on learning. Second, the representational bias of behavioral modeling for identifying shallow reading behaviors depends on the definitions' accuracy (Maldonado-Mahauad et al., 2018). Numerous studies on shallow reading have relied solely on conceptual descriptions from individual opinions, questionnaires, and literature reviews, owing to the absence of a clear definition (Carr, 2011). Notably, Riley-Tillman et al. (2009) found that the accuracy of disruptive behavior assessment was low and affected by behavior wording.

Clear definitions of shallow reading and its behavioral features are still lacking, particularly for large-scale online learning with massive data of measurable complex behavioral features (Locher & Philipp, 2023). To address the lack of a clear definition of shallow reading in the context of attendant confusion, it is necessary to extract measurable data on the behavioral features based on accurate definitions (Cooper et al., 2007). In particular, Martin and Pear (2019) suggest three behavioral features: duration, rate, and intensity. In other words, the definition describes the frequency and duration of actions, and the quantity of information involved. A behavior identification model is also crucial (Leshner & Pfaff, 2011). Khalil, Prinsloo and Slade (2022) report that it is necessary to define the problem the model is to solve and operation sequence required to solve it. Without accurate and effective models, it is difficult to determine how much shallow reading is associated with learning performance, class engagement, and behavioral patterns in online learning.

To this end, a behavior identification model is proposed that highlights measurable reading behavioral features using three dimensions consisting of frequency, duration and quantity and large-scale data. Next, to explore the relationship between shallow reading characteristics and learning performance, a combined approach of cluster analysis and correlation analysis is adopted. This approach examines data from both user-based and log-based perspectives and employs probability density visualization to accurately analyze the impact of behavioral characteristics. Based on this, a strategy to enhance shallow reading behavior by targeting and reducing reading speed. Finally, a real educational environment was utilized to conduct a comparative experiment, evaluating the learning effect, cognitive load, occurrences of shallow reading, and changes in reading speed.

4.3.1.4 Design and application of behavior identification model for shallow reading

The shallow reading behavior-identifying model. A behavior identification model for shallow reading behavior was developed. It incorporates three dimensions: RS, RDT, and RWC. Figure 8 shows its structure, functions, and components. Specifically, the model involves three steps: collecting reading behavior data via an e-book system, identifying the three dimensions, and identifying behavior.

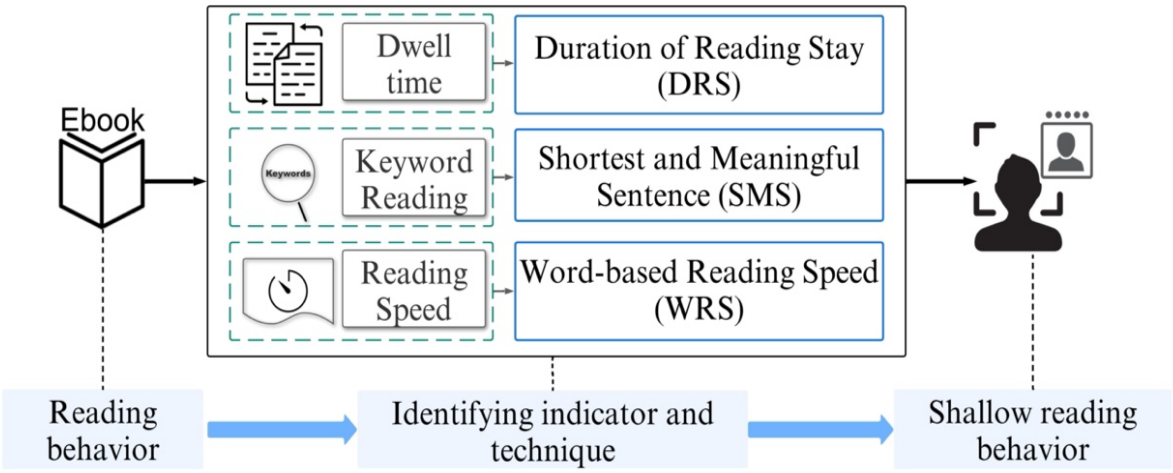

Figure 8. Shallow reading behavior identification model

Collecting reading behavior log data via an e-book system**.** An e-book learning system was integrated into the model to provide the necessary functionality for a large-scale collection of objective data on shallow reading behavior. It extends the range of reading behavioral features by providing numerous system functions (Zhao et al., 2021), such as (1) Page-turning (Prev, Next), (2) Zoom+/-, (3) Underline, (4) Highlight, and (5) Memo, as shown in Figure 9. The results of data collection are shown in Table 2, including User ID, Learning Materials (LM), Page, Reading Operation (RO), Start and End Times (ST, ET) at which the underline or highlight and the content were marked. These data provide the necessary support for subsequent measurement of shallow reading behavior. In addition, the data collection strategy requires underlining or highlighting to mark the read content. In other words, learners are trained to proficiently underline or highlight relevant content as a prerequisite, as shown in Figure 10.

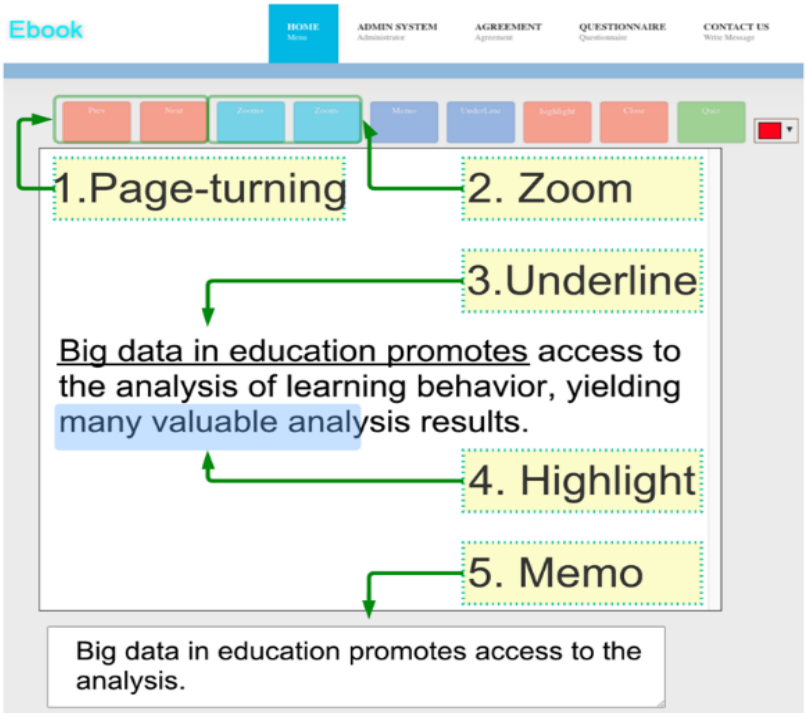

Figure 9. E-book system interface

Participants were instructed to utilize the underline or highlight function at the beginning of reading a designated text, and again at the conclusion of the reading session, where the effects

persisted. This approach enabled the recording of start and end times, dwell time, and specific text read by each participant, facilitating subsequent measurement of shallow reading behavior. The reading duration encompasses the time spent using the underline or highlight buttons, in addition to the time devoted to reading the text.

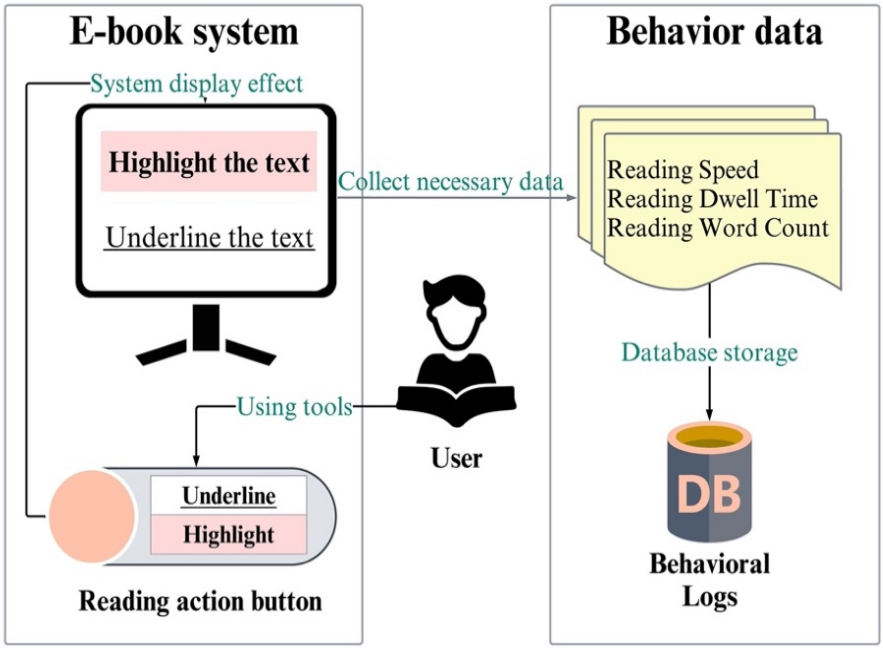

Figure 10. Three-dimensional data collection strategy for shallow reading behavior

Based on the data collection strategy, the focus is primarily on gathering the start and end times, dwell time, and specific text when students read a particular text. Other data captured by the e-book system, such as students taking memos, and clicking Zoom, have been excluded from the collection process. The data collection strategy primarily focuses on obtaining start and end times, dwell time, and specific text read by students during their engagement with a particular text. Data related to other activities within the e-book system, such as memo-taking or zooming, has been excluded from the collection process.

Table 11. Samples of reading behavior data collected by the e-book system

ID	LM	Page	RO	ST	ET	Content
1	Game	0	Underline	2022:01:01 22:22:22	2022:01:01 22:32:22	"add an underline"
2	Game	1	Next	2022:01:01 22:34:22	2022:01:01 22:45:22	

The three dimensions of reading behavior**.** According to previous research, reading speed (frequency: how fast), reading dwell time (duration: how long), and keyword reading (quantity: how much) are three fundamental dimensions of reading behavior. These are three basic indicators of shallow reading behavior, with measurable and observable reading data. First, reading speed value is the most prominent response to reading state (McGeown et al., 2015). Accordingly, measuring shallow reading behavior primarily relies on the reading speed indicator, through which an accurate correspondence between reading state and speed could be found. Currently, the common method for calculating reading speed is to divide the time it takes a learner to read a textbook page by the total number of words on that page (Liang & Huang, 2014). This method assumes that learners read all words on a page. By contrast, a word-based

method for calculating reading speed was developed. In this method, the time spent on reading words using highlights or underlines is employed, rather than assuming that all words on the page have been read, thus avoiding counting time for unread content. The mathematical formula is as follows:

$$RS = \frac{N}{T}$$

where RS refers to the reading speed value; N is the number of words read, and T is the time spent on reading. It was found that 400 words per minute is the threshold for shallow reading (Liang & Huang, 2014). In other words, if the RSV exceeds this value, the reading behavior is identified as Shallow Reading Speed (SSR).

Another standout indicator is RDT by measuring Duration of Reading Staying (DRS). In shallow reading, DRS is characterized by whether learners jump pages. Specifically, learners spend a short time scanning a text and then reading the next page (Fitzsimmons et al., 2020). Therefore, it is critical to determine the minimum DRS required for effective learning. McSpadden (2015) proved that the minimum time required for effective memorization when reading material is 8 seconds. Moreover, the performance of RDT dimension in shallow reading is Page Jumping (PJ). This is the DRS threshold for the PJ behavior. Accordingly, DRS at or below this value characterizes a PJ behavior in accordance with the mathematical expression shown below.

$$\{d \in PJ | d \leq DRS\}$$

where PJ refers to a student's page jump behaviors and d represents the DRS of one page jump behavior, which is less than or equal to the DRS threshold.

Keyword Reading (KWR), known as fragmented reading, is the third dimension. In contrast to reading complete sentences, it refers to reading fragmented words, phrases, and short sentences. The key to quantifying this behavior is to determine whether incomplete sentences are being read. An incomplete sentence refers to the shortest meaningful sentence. A meaningful sentence represents the textual content excluding symbols, punctuation, and graphics.

$$\{n \in KWR | n \leq SMS\}$$

where KWR represents a student's keyword reading behavior; n refers to the Reading Word Count (RKC) associated with this behavior; SMS is the number of words in the shortest and most meaningful sentence. When n is less than or equal to the SMS threshold, it can be determined as a KWR.

In addition, the Content Importance of Keyword Reading (CIKR) is calculated by the TF-IDF algorithm, as shown below. It is recorded along with behavior identification. Additionally, the Natural Language Toolkit (NLTK), a leading Python-based natural language processing platform, was used to implement the TF-IDF algorithm.

$$TF - IDF = TF \times IDF$$
$$TF = \frac{f_{td}}{N_j} \times IDF = \log\left(\frac{N}{C_t + 1}\right)$$

where TF and IDF are term frequency and inverse document frequency, respectively. f_{td} is the frequency of word t in document d; N_j is the total number of words in document d; N refers to the total number of documents in the corpus; and C_t represents the number of documents that contain word t.

4.3.1.5 Identifying shallow reading behavior

Figure 11 shows a flow chart of behavior identification. We input the data associated with the aforementioned dimensions. Then we identified the shallow reading behavioral features as follows. For every student, we first determined whether the DRS was below the specified threshold for shallow reading behavior; otherwise, a PJ behavior was indicated. If no PJ behavior was identified, we then determined whether the corresponding sentences were the shortest and most meaningful. Given the reading speed, we determined if the number of words read was below the specified threshold for shallow reading behavior; otherwise, a KWR was indicated. Lastly, if the sentences were not the shortest and most meaningful and the reading speed was above the set threshold, shallow reading behavior is characterized.

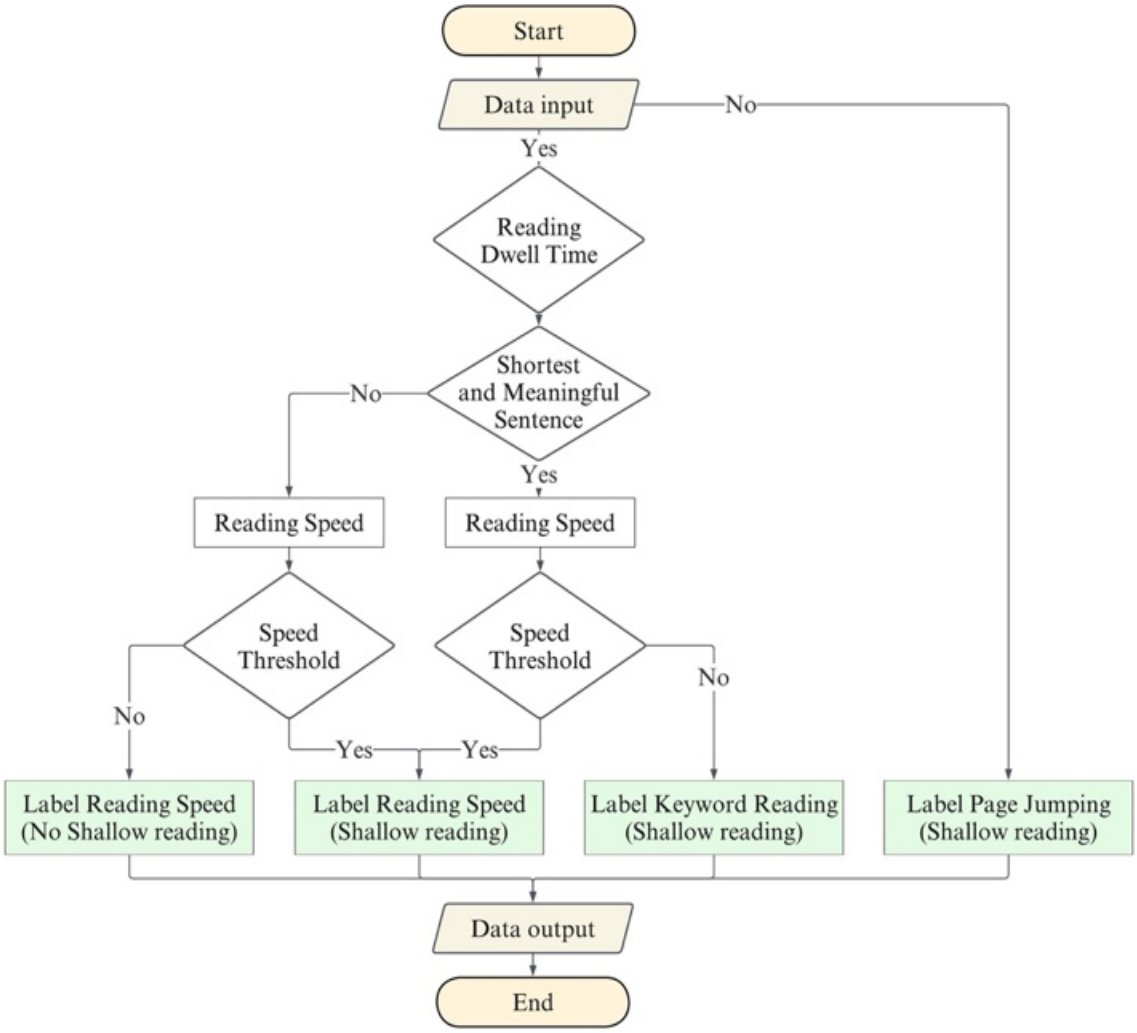

Figure 11. Identifying shallow reading behaviors

Based on Figure 11, the reading behaviors and learning performance were first collected through e-book system and in-class tests, which were used as Input Data. After that, the Input Data was then calculated using the indicators designed in the model for measuring the three

behavioral dimensions, and the behaviors that meet the indicators were saved as Output Data and assigned the corresponding labels. During the process, the User ID and Page were retained; Learning Material and Reading Operation were deleted, Reading Speed (RS) was calculated by Starting Time, End Time, and Content, RDT is obtained by subtracting Starting Time from End Time, and Content generated RWC and CIKR. Behaviors in the output data that met the model identification metrics were labeled. Notably, the Score is the result of participants' learning performance assessment after reading materials. It is measured using an in-class test consisting of multiple-choice and short-answer questions, with a total score of 100 points.

4.3.2 The analysis of shallow reading behavior

4.3.2.1 The shallow reading behavior identification

According to the model, the steps consist of collecting data, measuring the three dimensions, and identifying behavior-identifying outcomes.

Data collection. The e-book system mentioned in Figure 2 was utilized to record the reading behavior of 51 graduate participants in a university for a week. The experimental activities were derived from a one-week graduate course on educational technology. The reading material focused on learning behavior measurement and analysis, spanning a total of 19 pages. During the learning process, students were required to complete the reading within 90 minutes based on the system operation requirements from previous training. At the end of the course, an in-class test including multiple-choice tests for basic knowledge and short-answer tests for application of knowledge was performed to evaluate learning performance within 30 minutes. After removing missing and inconsistent data through the Python-based Pandas library, 10,694 valid behavioral logs remained.

Identifying shallow reading behavior using the proposed model. Initially, reading behavior data of the 51 participants were calculated using three dimensions of RS, RDT, and RWC. In addition, the CIKR and learning performance scores in-class tests were recorded. Finally, the behavior that triggered the judgment criteria would be identified as SSR, PJ, and KWR, as shown in Table 12.

Table 12. Examples of identifying results of shallow reading behavior

ID	Page	RS	Label	RDT	Label	RWC	Label	CIKR	Score
1	1	1.342	SSR	7 s	PJ	5	KWR	0.098	80
1	2	1.817		123 s		20		0.263	80

4.3.2.2 Analysis of the impact of shallow reading on learning

A two-step analysis was adopted to determine the behavioral patterns of shallow reading and their correlations with learning performance. Firstly, the k-means clustering algorithm, combined with a probability density visual analysis, provided insight into the latent behavioral patterns that represent the critical components of shallow reading. In particular, sole reliance on user-based behavior identification data analysis might lead to bias towards users with extensive log data contributions. To mitigate this concern, we subsequently conducted clustering analysis at the user level, followed by correlation analysis at the log data unit level. Correlation analysis focused on exploring the kind and extent of the relationship between

learning performance and shallow reading. These approaches minimized potential errors arising from uneven behavior distribution and offered diverse analytical perspectives.

Analysis of behavioral patterns of shallow reading. K-means clustering model. To obtain an accurate understanding of various behavioral features of shallow reading, the k-means clustering algorithm was applied to divide shallow reading into optimal clusters. The error square sum (SSE) is the objective function for dissimilarity among the clusters. The number of clusters (hyperparameter k) is determined using the elbow method. Additionally, the silhouette coefficient, an evaluation method ranging from -1 to 1, assesses the clustering effect. Based on a rule of thumb and existing research practices (Lovmar et al., 2005), clustering results with a Silhouette score greater than 0.65 exhibits the best performance. Results with a Silhouette score below 0.25 are not accepted, while scores falling between 0.25 and 0.65 require confirmation through visual inspection.

$$\mathrm{SSE} = \sum_{i=1}^{\mathrm{k}} \sum_{\mathrm{x} \in \mathrm{C}_i} |\mathrm{x} - \bar{x}_i|^2$$

C represents one category of the clustering results, $\bar{x}_i$ is the average of one category. The k-means algorithm consists of three main steps. Initially, data samples are assigned to the nearest cluster center to minimize the objective function value. Subsequently, the cluster means are updated accordingly. The optimal clustering outcome is achieved when the cluster centroids attain the lowest values of the objective function. Euclidean distances serve as the measurement for data samples within the k-means algorithm.

$$d(x_i, x_j) = \sqrt{\sum_i (x_i - x_j)^2}$$

For data samples x_i, x_j, the Euclidean distances are computed by squaring and summing the differences of their coordinates and then taking the square root.

The k-means model embedded in the Scikit-learn package in Python was applied. By combining the elbow method, k was set to 2. An acceptable accuracy of the clustering effect can be indicated by a silhouette coefficient of 0.615, which meets the reliability requirements.

Analysis of Clustering results and probability density distribution. The clustering results were evaluated using two criteria: silhouette values and features-by-clusters distribution graphics. As shown in Table 13, the two clusters can be distinguished by three dimensions, learning performance (Score and Number of People (NP)), number of behaviors (Number of Occurrences of Shallow Reading (NOSR)), and behavior elements (SSR, PJ, KWR, and CIKR). The findings in Table 13 unveil the substantial disparities between the two clusters in relation to the clustering centroid values for the five behavioral features. Notably, Cluster 1 exhibits markedly higher scores in all five features when compared to Cluster 2. Specifically, Cluster 1 showcases a score twice that of Cluster 2 for the PJ feature, more than twice the value for CIKR, over three times the magnitude of NOSR, and merely one-fifth the frequency of KWR, thus signifying the evident divergence in the behavioral profiles of these two clusters.

Table 13. Clustering results of shallow reading behavior variables

	Clustering centroid value						
Cluster	Score	SSR	NOSR	PJ	KWR	CIKR	NP
1	8.15	16.94	153.9	91.6	29.05	0.00076	40
2	4.0	36.16	410.5	128	104.50	0.00031	11

For a comprehensive investigation into the behavioral features of both clusters, we employed probability density distribution plots, visually depicted in Figure 12. These graphical representations capture the ranges of values observed for the five behavioral features on the horizontal axis, while the corresponding probability density values are represented on the vertical axis. Moreover, the more occurrences of the behavior, the higher density with a higher magnitude. Fewer occurrences, and lower density with a flat curve. The area formed by the product of these horizontal and vertical coordinates effectively signifies the probability of occurrence for the specific behaviors.

By meticulously analyzing the probability density distribution depicted in Figure 12, we were able to discern the intricate behavioral patterns characterizing Cluster 1. Within this cluster, the feature of NOSR exhibited values ranging from 100 to 200 times, with the Score concentrated predominantly in the range of 7 to 9 points. The PJ feature demonstrated occurrences ranging from 50 to 100 times, while the KWR featured frequencies spanning from 10 to 30 times. Furthermore, the CIKR values ranged from 0.0003 to 0.0008, signifying a wide spectrum of occurrences for this behavior. Contrastingly, in the case of Cluster 2, a distinct behavioral profile emerged, as evident from the probability density distribution. The NOSR exhibited values ranging from 300 to 500 times, with the Score predominantly falling within the range of 3.5 to 4.5 points. Moreover, the PJ feature demonstrated frequencies spanning from 200 to 400 times, while the KWR manifested frequencies ranging from 70 to 120 times. Notably, the CIKR values ranged from 0.00002 to 0.0004, indicating a narrower scope of occurrences for this behavior within Cluster 2.

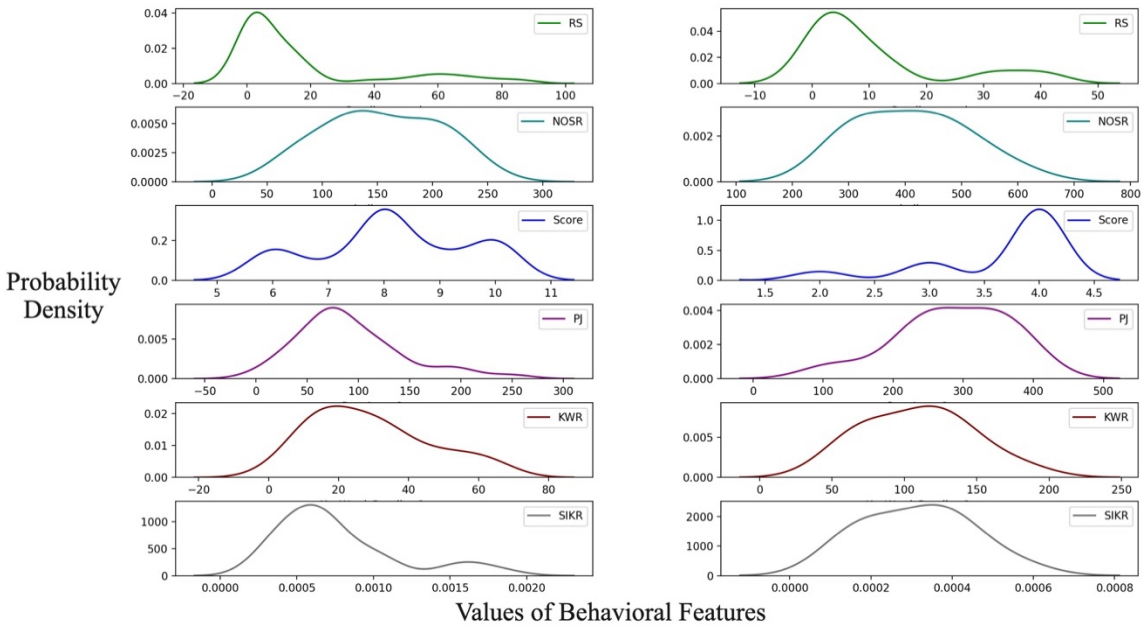

Figure 12. Probability density distribution of Reading Behavior Clusters

In summary, the Score-based reference benchmark reveals that Cluster 1 performs learning effects with significantly fewer occurrences (100 to 200 times) of shallow reading behaviors. Besides, Cluster 1 also shows fewer in terms of behavior elements, PJ, SSR, and KWR, while it has better performance on CIKR two times than Cluster 2. Based on the user-based analysis dimension, it can be inferred that varying occurrences of shallow reading behaviors exhibit distinct learning performances.

Correlation analysis between shallow reading behavior and learning performance. The results of the analysis suggest that the frequency of shallow reading behavior is negatively related to learning performance. Considering the accurate extent to which shallow reading affects learning performance, a subsequent correlation analysis was performed using Pearson's correlation coefficient (*r*). The correlation coefficient ranges from -1 for negative association to 1 for positive association, which statistically indicates the strength and direction of the associations among variables. Table 14 shows eight types of correlation strength: Positively Very Strong (PSR), Positively Strong (PR), Positively Moderate (PM), Positively Weak (PW), None (N), Negatively Weak (NW), Negatively Moderate (NM), Negatively Strong (NS), and Negatively Very Strong (NVS).

$$r = \frac{\sum(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\,\sqrt{\sum(y_i - \bar{y})^2}}$$

r denotes the correlation coefficient, while x_i, y_i signify the values of the x-variable, y-variable, respectively, within a given sample. Besides, $\bar{x}$ and $\bar{y}$ represent the means of the x-variable and y-variable values.

Table 14. Interpretation of correlation coefficient value

Strength	PSR	PR	PM	PW	N	NW	NM	NS	NVS
Value	$0.7 < r \leq -1$	$0.5 < r \leq 0.7$	$0.3 < r \leq 0.5$	$0 < r \leq 0.3$	0	$-0.3 \leq r < 0$	$-0.5 \leq r < -0.3$	$-0.7 \leq r < -0.5$	$-1 \leq r < -0.7$

A z-test was also performed to test for normality, skewness, and kurtosis. The results confirm that the absolute z-scores of the five variables are less than 1.96, satisfying the assumption of normality and the premise of using Pearson's correlation (Kim, 2013). Pearson's correlation analysis shows a very strong correlation between shallow reading variable frequency and learning performance. According to Table 15, the SSR and CIKR were higher than the other two variables, PJ and KWR, for correlation strength. In particular, the coefficient of CIKR was 0.768 ($p < .01$), which indicates that learning performance improves with CIKR.

Similarly, a high degree of correlation of -0.765 ($p < .01$) between SSR and learning performance scores using an in-class test was also observed; learning performance is negatively and strongly dependent on reading speed. The students' scores were inversely proportional to their reading speed. Moreover, the correlation coefficient value of KWR is -0.520 ($p < .05$), indicating a strong negative correlation with learning achievement. However, this magnitude is less than that of SSR and CIKR. Distinct from these three variables, PJ is weakly correlated with learning performance; their correlation coefficient is -0.297 ($p < .01$). correlations between the three behavioral characteristics (SSR, PJ, and KWR) of shallow reading were analyzed. Table 16 demonstrates PSR correlation with 0.725 ($p < .01$) between SSR and KWR, and a PM correlation with 0.413 ($p < .05$) between SSR and PJ. This indicates that slower reading speed corresponds to fewer occurrences of PJ and KWR behaviors, and vice versa.

Table 15. Correlation analysis between shallow reading behaviors and learning performance

	SSR	PJ	KWR	CIKR
Score	-0.765**	-0.297*	-0.520**	0.768**

*$p < .05$, **$p < .01$.

Table 16. Correlation analysis of shallow reading behavior features

	PJ	KWR
SSR	0.413*	0.725**

*$p < .05$, **$p < .01$.

A learning strategy to precisely mitigate shallow reading behavior. The formulation of reading strategies prioritizes acceptability and practicality. The acceptance of new reading strategies by students varies significantly due to diverse factors, including motivation, individual differences, vocabulary processing, prior knowledge, and areas of interest (Coiro & Dobler, 2007). To increase acceptance and suitability for school-based interventions, we introduced familiar and easily mastered reading strategies into the behavioral intervention approach (Gresham, 2004). Through clustering and correlation analysis, we identified a correlation between shallow reading behavior features and learning performance, with the reading speed feature exhibiting the strongest correlation and contributing the most to academic achievement. Consequently, we developed a learning behavior intervention strategy with a core objective of reducing reading speed.

To address how to precisely mitigate shallow reading, the study adopts an easily implementable strategy, transferring effective paper-based methods to the digital environment. Specifically, reading aloud was identified by McCallum et al. (2004) as a technique to achieve low-speed reading and improve memory. Furthermore, computer-assisted learning emulates paper reading habits by providing reading manipulators like underlining and highlighting, with the pen replaced by a mouse. In this study, Dots Per Inch (DPI) serves as the measure of mouse sensitivity, with a setting of 500 Hz shown by Microsoft research to effectively slow down reading speed while maintaining an acceptable mouse operation sensation. Based on these findings, a learning strategy known as Dpi and Read Loud (DRL) was proposed. The DRL strategy, which is shown in Figure 13, includes elements that inhibit excessive reading speed, such as reading loud and low DPI, while also balancing a satisfactory mouse-driven reading tool experience at the 500 Hz parameter setting.

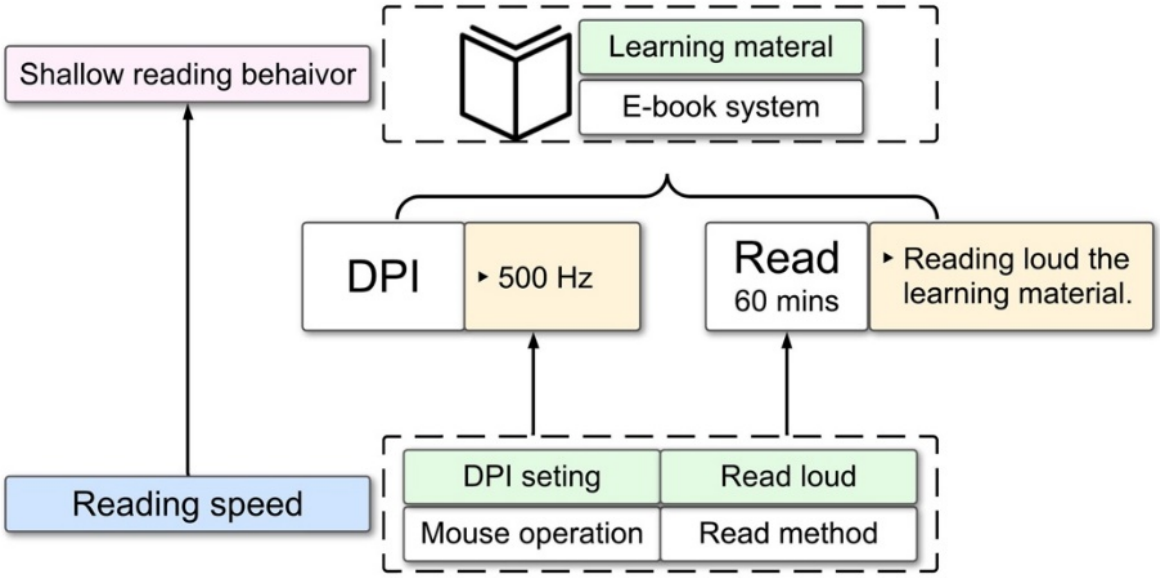

Figure 13. DRL learning strategy.

4.3.2.2 Evaluating the contribution of the DRL strategy

This study aimed to evaluate the effectiveness of the DRL strategy in promoting learning performance and reducing the occurrence of shallow reading behaviors. So, another experiment designed to evaluate the contribution of the DRL strategy was conducted. As its main function is based on RS reduction and its ability to provide a low-speed reading method, its effect in terms of RS reduction must be determined.

Experiment design. *Participants.* A total of 71 graduate students participated in the experiment and were randomly assigned to the experimental and control groups. The experimental group consisted of 37 participants who were asked to read the learning content of educational technology in the "Large Scale Intelligent Systems Theory" course using the DRL strategy. In contrast, the 34 participants in the control group were not assigned a learning strategy and were allowed to adopt a learning strategy that suited their preferences in the learning task.

Instruments. A pre-test measured prior knowledge background, post-test assessed learning performance, and cognitive load questionnaire examined the impact of learning strategies were used in the experiment. Because the learning content selected in the experiment was part of a professional course at a university, the pre-test and post-test were modified based on the in-class quiz. The pre-test consisted of ten items on a 10-point scale, which was mainly related to the learning emphasis in the subsequent learning content. Unlike the pre-test, the post-test had six items and four narrative questions. Based on Sweller's (1988) measurement of cognitive load, the questionnaire was used to determine cognitive responses when new learning strategies were introduced in learning, which included eight questions related to mental load and mental effects. It used a 5-point rating scheme ranging from 5 (*strongly agree*) to 1 (*strongly disagree*). The question's design primarily employs dual-task paradigms, subjective ratings, and physiological measuring techniques to assess the cognitive resources expended during learning.

Experiment procedure. To investigate the effectiveness of the proposed DRL strategy in reducing shallow reading behavior and promoting better learning performance, a two-week experiment was conducted at a university. Figure 14 shows the experimental procedure,

including the introduction of the experiment, pre-test and post-test, training of the learning system for all participants, and learning activities with two different learning strategies.

Before the experiment, participants were introduced to the experiment requirements and procedures, the use of the e-book system, and exercises, and a pre-test was given within 30 minutes to measure their prior knowledge of learning content. After a week, the participants in the experimental group were trained on the DRL strategy within 60 minutes. A week later, both groups performed two 60-minute learning activities one day apart. During the activity, both groups used the same e-book system but different learning strategies. Finally, a post-test on learning performance and a post-questionnaire were conducted within 30 minutes.

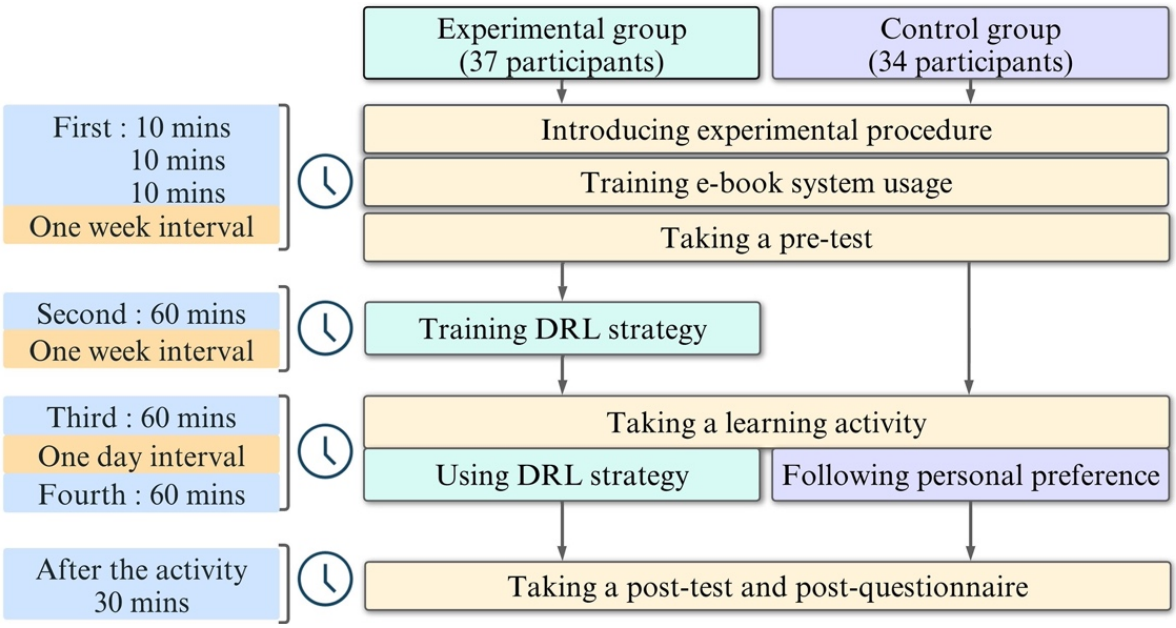

Figure 14. Experimental procedure

4.3.3.3 Experimental results

Analysis of learning achievement. To determine whether participants had equivalent basic knowledge about the learning content, a pre-test was conducted before the learning activity. The assumption of homogeneity of variance was not violated ($F = 1.327$, $p > .05$). Accordingly, all statistics obtained by the *t*-test support no significant differences ($t = -1.176$, $p > .05$) between the independent groups. Specifically, the mean and standard deviation (SD) of the pre-test were 2.22 and 1.03 for the experimental group and 2.53 and 1.21 for the control group, respectively, as shown in Table 17.

Table 17. Descriptive statistics and t-test results of the learning achievement in the pre-test

Variable	Group	N	Mean	SD	*t*
Pre-test	Experiment group	37	2.22	1.03	-1.176*
	Control group	34	2.53	1.21	

Note. *$p > .05$.

A post-test applied in both groups was employed to delve into the effectiveness of the DRL strategy in learning performance. A *t*-test was conducted to confirm whether any statistically significant differences existed between the two groups after the learning activities. Prior to the Analysis of Variance (ANOVA), Levene's test for equality of variances was initially performed,

satisfying the assumption of homogeneity of variance ($F = 0.490$, $p > .05$). The t-test results ($t = 4.103$, $p < .01$) in Table 18 illustrate that the learning performance of the experimental group was superior to that of the control group. Statistically, a higher degree was achieved by participants in the experimental group. It was concluded that the DRL strategy facilitates learning performance.

Table 18. Descriptive statistics and t-test results of the learning achievement in the post-test

Variable	Group	N	Mean	SD	t
Pre-test	Experiment group	37	8.03	1.69	4.103**
	Control group	34	6.29	1.89	

Note. $^{**}p < .01$.

Analysis of occurrence and RS of shallow reading behavior. The most distinguishing feature of the DRL strategy lies in the constraint of shallow reading behavior, which consists of a reduction in occurrence and RS. According to Levene's test for the equality of variances ($F = 3.362$, $p > .05$), the use of a t-test is appropriate. First, the t-test results ($t = 6.510$, $p < .05$) show a significant difference in the occurrence of shallow reading behavior between the two groups. As shown in Table 19, the mean and SD were 123.21 and 44.29 for the experimental group and 228.18 and 86.55 for the control group, respectively. It was found that the number of occurrences of shallow reading behavior in the experimental group was nearly twice that of the control group, which implies that the DRL strategy could be conducive to reducing the occurrence of shallow reading behavior. Considering further investigation of the intervention effect on shallow reading behavior, RS was also considered. Second, after Levene's test for equality of variances ($F = 66.769$, $p > .05$), the one-way ANOVA analysis results ($F = 53.285$, $p < .01$) in Table 20 show that an obvious decrease in reading speed was found by students in the experimental group compared to those in the control group. The students in the experimental group had a mean RS of 3.03 WPM, which was less than the mean of 5.20 WPM in the control group. This revealed that the DRL strategy could significantly facilitate a decline in RS.

Table 19. Descriptive statistics and t-test results of occurrence of shallow reading behavior

Variable	Group	N	Mean	S.D.	t
Pre-test	Experiment group	37	123.21	44.29	6.510*
	Control group	34	228.18	86.55	

Note. $^{*}p < .05$.

Table 20. Descriptive statistics and ANOVA results of reading speed of shallow reading behavior

Variable	Group	N	Mean	S.D.	F
Pre-test	Experiment group	37	3.03	7.52	53.285**
	Control group	34	5.20	18.56	

Note. $^{**}p < .01$.

Analysis of cognitive load. Cognitive load, composed of mental load and mental effort, potentially affects learning performance and the extent to which learners allocate and accommodate resource demands when processing uncertain information (Shuell, 1986). Mental load refers to the amount of stress caused by the expected cognitive capacity. By contrast, mental effort signifies the extent to which limited cognitive capacity is occupied (Zhao et al., 2023). Consequently, a post-cognitive load questionnaire was administered to measure the cognitive condition of the two groups.

Table 21. Descriptive statistics and t-test results for cognitive load

Variable	Group	N	Mean	S.D.	*t*
Mental load	Experiment group	37	11.94	3.438	-1.920*
	Control group	34	13.46	3.156	
Mental effort	Experiment group	37	11.14	3.117	-1.713*
	Control group	34	12.66	4.200	

Note. *$p > .05$.

The variance homogeneity was shown in Levene's test results ($F = 0.238$, $p > .05$). As shown in Table 21, no significant differences were found between the two groups. Concerning mental load, the t-test result ($t = -1.920$, $p > .05$) demonstrated that all participants had an approximate mental load regardless of group. Another conclusion also follows from the t-test result ($t = -1.713$) of the mental effort—results from the experimental group did not differ from the control group, indicating that the DRL strategy did not increase the cognitive load. This conclusion corresponds to that obtained on the basis of mental load.

4.3.4 Discussion

Avoiding the effects of various emotional tendencies on behavior identification for RQ1. Questionnaire-centric subjective data included self-assessments of behavioral characteristics, motivations, thoughts, and feelings. Kormos & Gifford (2014) found that differences in behavioral assessments have been shown to arise from various emotional tendencies in behavior wording in subjective assessment instruments, which emphasized a dilemma in behavioral identification based on subjective data is emphasized: most truthful self-reporting and unconscious self-avoidance. To avoid the behavior identification bias induced by subjective data, the objective behavioral data obtained through the e-book learning system is a continuous record of the behavioral activities of participants unconsciously, reflecting the state of their learning behavior (Yin & Hwang, 2018). The identifying model indicators that measure behavior are based on the three dimensions (RS, RDT, and RWC), which encompasses behavioral determination thresholds. So, emotional tendencies impeding behavior identification are substituted with unconscious behavioral features and corresponding formulas and values.

The accurate formulation of learning strategies hinges on precise behavior analysis for RQ2 and RQ3. Previous studies lack such detailed intervention strategies due to the complexity of shallow reading behavior and the absence of a comprehensive analysis of its core characteristics. For instance, Chen and Chen (2014) attempted a collaborative approach to address poor comprehension in a digital learning environment characterized by shallow reading. However, they failed to identify shallow reading behavior and did not extract its core characteristics from complex reading behavior analysis. To design precise behavioral intervention strategies, a thorough understanding of shallow reading's core characteristics is essential. This study adopts two analysis dimensions: individual students as units (cluster analysis) and single shallow reading behaviors as units (correlation analysis). This approach addresses differences in the number of shallow reading behaviors among students and mitigates potential analysis biases towards those with shallow reading behaviors. It ensures the analysis accurately explores the relationship between shallow reading characteristics and learning performance while maximizing the impact of each behavioral characteristic.

Data collection accuracy plays a crucial role in identifying shallow reading when dealing with RQ1, and we found that RS is a major factor for RQ2. Despite the wide range of methods for calculating it, the calculation accuracy is not high (Zhao & Yin, 2022). The main reason for this is the lack of high-precision behavioral capture tools to record various data (e.g., time spent reading a sentence, phrase, and specific texts and the total number of words read). In a study by Liang & Huang studies (2014), their behavior identification is reliant on RS determination, assuming linear homogeneous reading with average speed describing reading a page of text. However, this approach fails to correspond to the real non-linear reading state. To address this limitation, this study collects high-precision data using the underlining and highlighting functions of the e-book system to calculate RS, incorporating the speed at which a few words are read.

4.3.5 Conclusions

A behavior identification model has been proposed that highlights measurable and observable reading behavior based on three behavioral dimensions: RS, RDT, and RWC. To avoid the shortcomings of the accuracy of self-assessment and the small amount of subjective data captured by the questionnaire for RQ1, the objective data automatically recorded by the learning system using the e-book was used to ensure the accuracy of results. Furthermore, shallow reading behavior was defined using three key dimensions that constitute behavior: duration, rate, and intensity. Finally, combined with the existing descriptive definitions of shallow reading, this study obtained three behavioral features: SSR, PJ, and KWR.

To accurately understand the relationship between shallow reading behavior and learning performance, the k-means clustering analysis algorithm and Pearson's correlation analysis were performed successively for RQ2. First, the results of the clustering analysis revealed that the difference between high and low learning performance was inversely proportional to the occurrence of shallow reading in terms of the three behavioral characteristics of RS, RDT, and RWC. In other words, the higher the learning performance, the lower the three behavioral characteristics of shallow reading. Second, Pearson's correlation analysis results suggest that the negative correlation between SSR and learning performance is higher than that of RDT and RWC. Finally, it was found that students with poor performance had a higher number of KWR behavior but few CIKR calculated by the TF-IDF algorithm.

The analysis revealed that differences in RS had the greatest impact on learning performance when exploring RQ1. For this reason, a learning strategy called DRL has been developed that interferes with shallow reading by slowing down reading speed. There are two inhibiting factors: DPI parameter settings for reading tool operation and reading loud for reading content. The experimental results show that DRL strategy based on reducing reading speed can improve learning performance, prevent shallow reading, and reduce RS without causing additional cognitive load. To attain a higher-precision behavior identification method, the eye-tracking-based approach is anticipated to mitigate the interference of reading tools involved in the learning process, and be more integrated with the real educational environment and students' usage habits.

5 Deep learning-based causal relationship exploration: causal mechanisms extraction

The relationship between causal relationship exploration and causal mechanisms. Pearl (2000) believes that the main purpose of causal relationship exploration is to reveal causal relationships between data variables, while causal mechanisms are to explore how causal relationships occur and the driving factors behind their occurrence. The explanation of causal mechanisms helps to understand how causal relationships are formed. As discussed earlier, deep learning technology provides enormous potential for causal relationship exploration, mainly including: firstly, automatically learning hierarchical representations of features from data to achieve recognition of learning mechanisms; Secondly, the ability to capture complex nonlinear relationships; Thirdly, deep learning has good scalability and adaptability.

The impact of artificial intelligence technology on causal mechanism exploration. Generative artificial intelligence technology, represented by GANs, has great potential for exploring causal mechanisms with its unique data generation methods. The biggest feature of generating adversarial networks is that this technology can generate generated data that is extremely similar to the distribution of real data (Goodfellow et al., 2014). The characteristic of generating data that reflects the distribution of the real world provides technical support for solving the key problem of causal relationship exploration: counterfactual inference. Specifically, generating adversarial networks creates counterfactual instances that reflect real-world distributions, simulating hypothetical scenarios to understand the formation of causal mechanisms and the underlying reasons (Shalit, Johansson & Sontag, 2017). However, when using generative artificial intelligence networks for causal relationship exploration, it is necessary to keep the generated data consistent with the real causal structure. Otherwise, relationship exploration based on mismatched causal structures will lead to inaccurate causal relationship exploration. For this issue, Pellet and Elisseeff (2008) proposed and validated the effectiveness of using Markov blanks to learn causal structures.

5.1 Application of deep learning into causal relationship exploration

5.1.1 Deep learning-based causality relationship exploration methods

Causal relationship exploration through neural network methods has seen significant progress as a burgeoning educational technique. Traditional methods, exemplified by experiments, entail substantial investments of time and financial resources. Conversely, machine learning methods, particularly Bayesian networks, grapple with limitations related to high-dimensional data, linearity assumptions, conditional independence, and distinguishing causation from correlation. Neural networks offer distinctive advantages in causal relationship exploration. They excel in capturing non-linear data limitations and challenging hypotheses, which can be daunting for other algorithms like LiNGAM. Louizos et al. (2017) introduced a deep latent-variable model built on neural networks to unveil non-linear causal relations within complex structures. Furthermore, one standout advantage is the capacity to directly learn causal relationships from raw data without prior knowledge of underlying causal structures (Nauta et al., 2019).

GANs bring substantial insights to causal relationship exploration, boasting several benefits: the flexibility to capture intricate data distributions, simulating the estimation of causal effects under various interventions (Goodfellow et al., 2014), enhancing the robustness of causal models, and being suited for nonparametric models (Shalit, Johansson & Sontag, 2017). The primary concept behind exploring causal relationships with GANs stems from their ability to

estimate causal effects by modeling the distribution of potential outcomes based on observed data (Kocaoglu et al., 2017). Challenges tied to limited sample sizes, which can hinder model training and causal structure discovery, are effectively addressed by GANs. These networks generate synthetic data that retains the statistical properties of the original dataset, mitigating the adverse effects of sample size limitations and imbalances (Gao et al., 2021). The robustness and generalizability of causal relationship exploration with GANs are further bolstered by their capacity for counterfactual reasoning. This enables models to perform counterfactual inference by generating data samples representing hypothetical scenarios under different interventions (Grari et al., 2023).

5.1.2 Challenges of deep learning-based causal relationship exploration

Deep learning-based causal relationship exploration faces evident challenges, particularly in the case of emerging techniques like GAN-based causal relationship exploration. These challenges stem from the instability of causal results and the associated evaluation metrics. Result instability is a common issue in the deep learning field, often manifesting as mode collapse due to the failure of modes learning within a limited diversity of generated samples (Goodfellow et al., 2014). Some studies posit that GANs may suffer from vanishing and exploding gradients, which impede stability. The choice of loss function for both the generator and discriminator, as a specific structural aspect of the network model, amplifies the volatility of results. Regarding evaluation metrics, commonly employed ones include Inception score, Wasserstein, Frechet inception distance, and 1-Nearest Neighbor. However, these metrics are not suitable for evaluating GAN-based causal relationship explorations because these causal results lack original causal relationships, making it impossible to reference real samples.

5.2 Principles and development of GANs-based causal relationship exploration

5.2.1 Principle of GANs-based causal relationship explorations

The core principles of causal relationship exploration using GANs are outlined. Establishing Markov chains and causal mechanisms. Janzing & Schölkopf (2010) integrate Markov kernels and causality mechanisms, learning d Markov kernels. Each j-th Markov kernel q_j, represents the conditional density of x_j, with candidate parents denoted as Figure $\hat{g}$. The learning process involves optimizing data likelihood based on the conditional distribution of $q_j\left(x_j \middle| x_{\mathrm{pa}(j;\hat{g})}\right)$, where $\mathrm{pa}(j;\hat{g})$ is an estimated cause of x_j, represented by a binary vector set. Additionally, $\hat{g}$ is specified to enforce sparsity and acyclicity. Janzing and Schölkopf (2010) designate each Markov kernel as a causal mechanism $\widehat{f_j}$, defined as follows:

$$\widehat{\boldsymbol{X}}_{\boldsymbol{j}} = \widehat{f}_j\left(\left[a_j \odot X, E_j\right], \theta_j\right)$$

Where X represents a distinct sample comprising d variables, that is $X = (x_1, \cdots, x_d)$. The j-th Markov kernel, represented as $a_j = (a_1, j, \cdots, a_d, j)$ with a binary vector. If the coefficient $a_i$in a_i, j is 1, it signifies that variable x_i in sample X can generate x_j, and edges $x_i \to x_j$ belong to the graph $\hat{g}$. θ_j denotes the parameters used in calculating $\widehat{f_j}$, such as the weights in neural networks. E_j encompasses all noise variables that lack observed causal relationships with $x_j$and are independent.

Derived from the causal mechanism $\widehat{f_j}$, the estimated cause based on x_j is defined as $q_j\left(x_j \middle| x_{-\mathrm{j}}, a_j, \theta_j\right)$, and it can be simplified to $q_j\left(x_j \middle| x_{\mathrm{pa}(j;\hat{g})}, \theta_j\right)$.

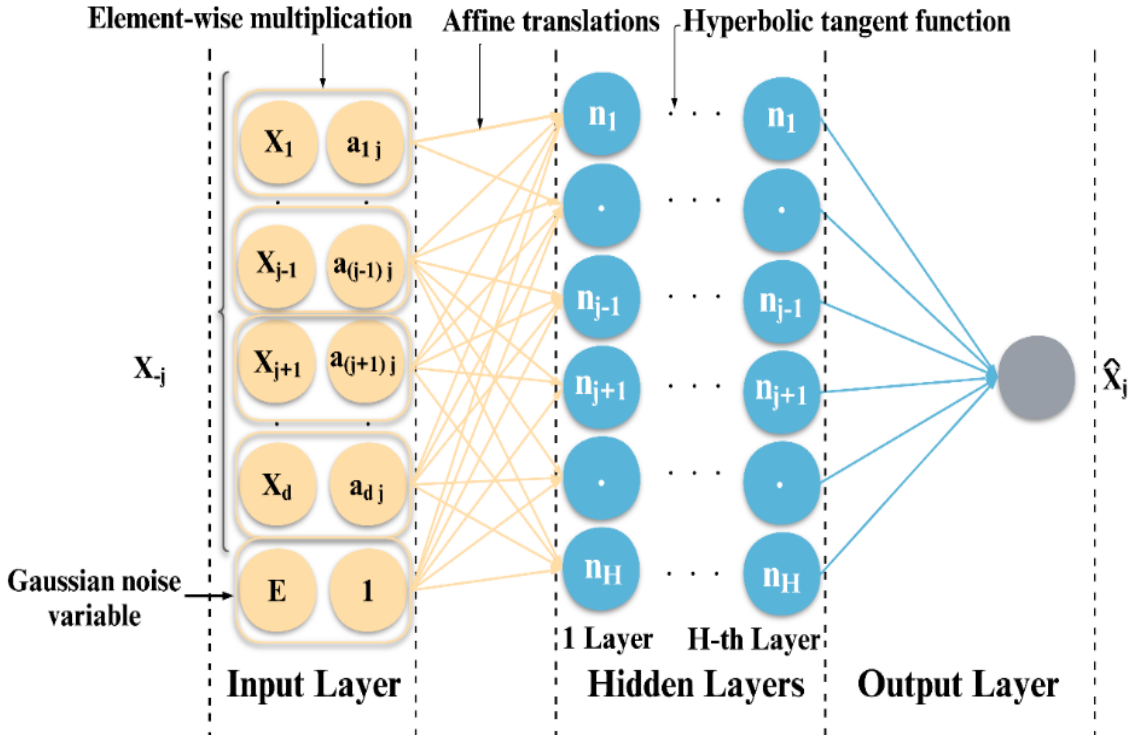

Figure 15. Generative neural network modeling by causal mechanism

Learning Markov kernels. According to Brown et al. (2012), the learning of the causal mechanism $\widehat{f}_j$ entails both learning $\widehat{f}_j$ itself and the minimal subset of $\mathrm{pa}(j;\hat{g})$. This learning process is ultimately achieved by minimizing the conditional log-likelihood of the data. The specific learning equation is as follows.

$$S_j^n(a_j,\theta_j,X) = \hat{I}^n\left(x_j, x_{\overline{pa}(j;\hat{g})} \middle| x_{pa(j;\hat{g})}\right) - \frac{1}{n}\sum_{\ell=1}^{n} log\, q_j\left(x_j^{(\ell)} \middle| x_{pa(j;\hat{g})}^{(\ell)}, \theta_j\right)$$

After decomposition,

$$S_j^n(a_j,\theta_j,X) = \hat{I}^n\left(x_j, x_{\overline{pa}(j;\hat{g})} \middle| x_{pa(j;\hat{g})}\right) + \frac{1}{n}\sum_{\ell=1}^{n} log \frac{p\left(x_j^{(\ell)} \middle| x_{pa(j;\hat{g})}^{(\ell)}\right)}{q_j\left(x_j^{(\ell)} \middle| x_{pa(j;\hat{g})}^{(\ell)}, \theta_j\right)} + cst$$

$\hat{I}^n\left(x_j, x_{\overline{\mathrm{pa}}(j;\hat{g})} \middle| x_{\mathrm{pa}(j;\hat{g})}\right)$ is utilized for identifying the Markov equivalence class of the true g. It quantifies the discrepancy between the true distribution and the one inferred by the model, where $\frac{1}{n}\sum_{\ell=1}^{n} \log \frac{p\left(x_j^{(\ell)} \middle| x_{\mathrm{pa}(j;\hat{g})}^{(\ell)}\right)}{q_j\left(x_j^{(\ell)} \middle| x_{\mathrm{pa}(j;\hat{g})}^{(\ell)}, \theta_j\right)}$. Its primary purpose is to distinguish between graphs within the Markov equivalence class of g, specifically assessing how well x_j aligns with the conditional distribution based on its parents $x_{\mathrm{pa}}(j;\hat{g})$. In this equation, "cst" denotes the operation of converting the input into a constant with the same value but zero gradient.

Assessing the structural loss of the generated causal relationship diagram. When generating the structural loss of the causal structure graph, identifying the minimum subset of $\mathrm{pa}(j;\hat{g})$ corresponds to a feature selection problem. The challenge lies in the Markov Blanket of x_j

converging to the true causal graph under the large sample limit (Yu et al., 2021). Kalainathan et al. (2022) attempt to optimize the data log-likelihood with a regularization term $\lambda_S|\mathrm{pa}(j;\hat{g})|$, where $|\mathrm{pa}(j;\hat{g})|$ represents the number of parents of x_j in $\hat{g}$. The sparsity of the Markov Blanket selection is controlled by the hyper-parameter λ_S >0. Consequently, without considering the acyclicity constraint, the following equation (4) is used to identify the moral graph associated with the true causal graph g.

$$\lambda_S^n(\hat{g}, D) = \sum_{j=1}^{d} \hat{I}^n\left(x_j, x_{\overline{pa}(j;\hat{g})} \middle| x_{pa(j;\hat{g})}\right) + \lambda_S|\hat{g}|$$

Evaluating the parameter loss of the causal relationship result diagram conforming to the Markov equivalence class. To achieve model identification within the Markov equivalence class of the true directed acyclic graphs, Neyshabur et al. (2017) controlled the complexity of the causal mechanisms by utilizing the Frobenius norm in the $\widehat{f_j}$ parameter as the regularization term, and the regularization weight as λ_F. Finally, Kalainathan et al. (2022) defined the parametric loss equation by combining the data fitting terms with the regularization term.

$$\lambda_F^n(\hat{g}, \theta, D) = \sum_{j=1}^{d}\left[\frac{1}{n}\sum_{\ell=1}^{n} log\frac{p\left(x_j^{(\ell)}\middle|x_{pa(j;\hat{g})}^{(\ell)}\right)}{q_j\left(x_j^{(\ell)}\middle|x_{pa(j;\hat{g})}^{(\ell)}, \theta_j\right)} + \lambda_F \left\|\theta_j\right\|F\right]$$

5.2.2 Implementation of GANs-based causal relationship explorations

Kalainathan et al. (2022) introduced a structural agnostic model (SAM) based on the aforementioned theory. To minimize the global score of the formula above, they employed three steps. First, each Markov kernel is represented as a conditional generative neural network. Second, an adversarial generative neural network is employed to optimize the conditional likelihood score with a Markov kernel. Finally, while introducing an acyclicity constraint to address the DAG combination optimization problem, stochastic gradient descent is used to optimize the model structure and causal mechanisms.

Establishing Markov chains and causal mechanisms. The Markov kernel $q_j\left(x_j|x, a_j, \theta_j\right)$ and causal mechanism $\widehat{f_j}$ are both modeled using a deep neural network with H hidden layers, where each layer consists of $n = (n_1, \cdots, n_H)$ nodes. The input dimension is $n = (n_1, \cdots, n_H)$, and the output dimension is $n_{H+1} = 1$. The mathematical expression can be represented as:

$$\hat{X}_j = \hat{f}_j\left(X, E_j\right) = L_{j,H+1} \circ \sigma \circ L_{j,H} \circ \cdots \circ \sigma \circ L_{j,1}\left(\mathrm{a_j} \odot \mathrm{X}, E_j\right)$$

This formula explains how the neural network $\hat{f}_j$ with multiple layers, takes the input X and combines it with Gaussian noise variable E_j. It processes this input through multiple hidden layers involving affine translations and activation functions to generate the output $\hat{X}_j$ This output is used to model the Markov kernel $q_j\left(x_j|x, a_j, \theta_j\right)$). Here, $a_j$represents an affine linear map that transforms the input X, and $\theta_j$encompasses all the weights and biases of the neural network $\hat{f}_j$, which can be expressed as $\theta_j = \left\{\left(w_{j,1}, b_i\right), \left(w_{j,1}, b_i\right), \cdots, \left(w_{j,H+1}, b_{H+1}\right)\right\}$.

In this context, $L_{j,H}$ represents the H-th layer of the neural network for the causal mechanism $\widehat{f_j}$. It takes the result of the previous step $(\mathrm{a_j}\odot \mathrm{X}, E_j)$ and applies an affine linear transformation to it. Here, $(\mathrm{a_j}\odot \mathrm{X}, E_j)$ signifies that the input X is element-wise multiplied by the vector a_j. This transformation involves a weight matrix a_j, a bias vector b_i, and the result is combined with the Gaussian noise variable E_j. The activation function, denoted as σ, is typically the hyperbolic tangent function, defined as $\sigma(z) = \left(tanh(z_1), \cdots, tanh\left(z_{n_h}\right)\right)^T$. The symbol $\odot$ represents function composition.

Optimizing the trade-off the structural and parametric regularization terms in the following equation when assessing the candidate DAG $\hat{g}$.

$$s^n(\hat{g}, \theta, D) = \sum_{j=1}^{d} \hat{I}^n\left(x_j, x_{\overline{\mathrm{pa}}(j;\hat{g})} \middle| x_{\mathrm{pa}(j;\hat{g})}\right) + \lambda_S |\hat{g}| + \sum_{j=1}^{d}\left[\frac{1}{n}\sum_{\ell=1}^{n} \log \frac{p\left(x_j^{(\ell)} \middle| x_{\mathrm{pa}(j;\hat{g})}^{(\ell)}\right)}{q_j\left(x_j^{(\ell)} \middle| x_{\mathrm{pa}(j;\hat{g})}^{(\ell)}, \theta_j\right)} + \lambda_F \left\|\theta_j\right\| F\right]$$

Firstly, the structural loss in the SAM algorithm encompasses the complexity of each mechanism $\hat{f}_j$, characterized by the sum of structural complexity and function complexity, regulated by respective weights $\lambda_S > 0$ and $\lambda_F > 0$. This algorithm gauges model complexity by evaluating the L_0 norm of a_j, where the L_0 norm essentially tallies the number of non-zero elements in a vector or matrix. In this context, a_j denotes a structural facet of the model and signifies the count of parents of x_j. More parents indicate a more complex structure for a variable. Another facet of model complexity pertains to functional complexity, which relates to how the causal mechanisms operate. To measure this complexity, $\left\|\theta_j\right\| F$ is employed, representing the Frobenius norm of the weight matrix $\theta_j = \{(w_{j,1}, b_i), (w_{j,1}, b_i), \cdots, (w_{j,H+1}, b_{H+1})\}$,⋯,(w_(j,H+1),b_(H+1))} associated with the causal mechanism. Specifically, the Frobenius norm quantifies the magnitude of a matrix and serves as a means to assess the complexity of neural network models. The specific formula for this calculation is as follows:

$$\left\|\theta_j\right\| F = \sum_{h=1}^{H+1} \left\|W_{j,h}\right\| F + \sum_{h=1}^{H+1} \left\|b_{j,h}\right\| F$$

$$\left\|W_{j,h}\right\| F = \sqrt{\sum_{\substack{1\le k\le n_h \\ 1\le l\le n_{h-1}}} \left|w_k^j \quad {}_l^h\right|^2} \text{ and } \left\|b_{j,h}\right\| F = \sqrt{\sum_{1\le k\le n_h} \left|b_k^j \quad {}^h\right|^2}$$

Secondly, data fitting loss. The formula for the data fitting loss is based on the hypothesis that as the number of samples, denoted as n, approaches infinity, the data fitting approximation approaches the expected log-likelihood of data under a generative distribution.

$$\lim_{n\to\infty} \frac{1}{n}\sum_{\ell=1}^{n} \log q_j\left(x_j^{(\ell)} \middle| x_{-\mathrm{j}}^{(\ell)}, a_j, \theta_j\right) = E_{p(x)} \log q_j\left(x_j \middle| x_{-j}, a_j, \theta_j\right)$$

The key challenge in the data fitting loss lies in quantifying the disparity and uncertainty between the real data distribution and the generated data distribution. Define $\tilde{X} = [X_1, \cdots X_{j-1}, \hat{X}_j, X_{j+1,\ldots}X_d]$, where true observed
Variables and $\hat{X}_j$ data generated by $\widehat{f_j}$. Consequently, $E_{p(x)} \log q_j(x_j|x_{-j}, a_j, \theta_j)$ is transformed into:

$$-E_{p(x)} \log \frac{p(x)}{\tilde{q}_j(x_j, x_{-j}, a_j, \theta_j)} + E_{p(x)} \log p(x_j|x_{-j})$$

where $\tilde{q}_j(\tilde{X})$ represents the joint distribution, $\tilde{q}_j(x_j, x_{-j}, a_j, \theta_j) = p(x_{-j})q_j(x_j|x_{-j}, a_j, \theta_j)$. To quantify the disparity in probability between the real data distribution and the generated data distribution, $E_{p(x)} \log \frac{p(x)}{\tilde{q}_j(x_j, x_{-j,,} a_j, \theta_j)}$ represents the Kullback-Leibler divergence between $p(X)$ and $\tilde{q}_j(\tilde{X})$, which is transferred to $D_{KL}[p\|\tilde{q}_j]$. Simultaneously, to gauge the predictability of x_j considering the context provided by x_{-j}, $H(X_j|X_{-j})$ constant, domain-dependent entropy of x_j conditionally to x_{-j} was employed. Consequently, the data fitting loss becomes an optimization task involving $D_{KL}[p\|\tilde{q}_j]$ and $H(X_j|X_{-j})$.

To tackle the challenges associated with estimating the quantity of each $D_{KL}[p\|\tilde{q}_j]$, especially when dealing with continuous data, the SAM employs a variational dual representation. This approach utilizes F-divergence as an alternative way to express the KL divergence, providing a lower bound on the KL divergence. A class of functions $T_\omega : R^d \longrightarrow R$ is defined, where $\omega \in \Omega$ represents a parameter, and T_ω is a family of functions parameterized by a deep neural network. These functions T_ω aim to minimize the lower bound on $D_{KL}[p\|\tilde{q}_j]$.

$$D_{KL[p\|\tilde{q}_j]} \geq \underset{\omega \in \Omega}{sup} E_{p(x)}[T_\omega(X)] - E_{q(x)}\left[e^{T_\omega(X)-1}\right]$$

During GAN training, the discriminator networks T_ω are employed to distinguish between samples drawn from the original data distribution $p(X)$ and the pseudo distribution $\tilde{q}_j(\tilde{X})$. For the pseudo-sample $\tilde{X}_j^{(\ell)} : \ell = 1, \cdots, d \; and \; j = 1, \cdots, d$ in dataset $\widetilde{D}_j$, it was defined from $x^{(\ell)}$ with j-th variable generated based on $\widehat{f_j}$, in which $e_j^{(\ell)}$ was drawn from Gaussian $N(0,1)$. Finally, the minimization of the empirical approximation of the lower bound on $\sum_{j=1}^{d} D_{KL[p\|\tilde{q}_j]}$ was modified as below.

$$\underset{\omega \in \Omega}{sup} \left(\frac{d}{n} \sum_{\ell=1}^{n} T_\omega\left(X^{(\ell)}\right) + \frac{1}{n} \sum_{j=0}^{d} \sum_{\ell=0}^{n} \left[-exp\left(T_\omega\left(\tilde{X}_j^{(\ell)}\right) - 1\right)\right] \right)$$

Finally, In the SAM, the introduction of the acyclicity constraint is crucial to ensure a Directed Acyclic Graph (DAG). This constraint couples feature selection problems by limiting the number of edges between pairs of variables. Consequently, this restriction leads to the removal of explanatory variables from the set of parents of any variable, followed by the deletion of spouse variables. The SAM defines the acyclicity constraint based on a learning criterion with an acyclicity term (Zheng et al., 2018) as follows:

$$\sum_{k=1}^{d} \frac{trA^k}{k!} = 0$$

Combing with the acyclicity constraint, the learning criterion is modified to:

$$L^n(\hat{g}^*, \theta^*, D) = \underset{A, \theta\, \omega \in \Omega}{min\ max} \left(\frac{1}{n} \sum_{\ell=1}^{n} \sum_{j=1}^{d} \left[T_\omega\left(X^{(\ell)}\right) - exp\left(T_\omega\left(\hat{f}_j^{\theta_j, a_j}\left(X^{(\ell)}, e_j^{(\ell)}\right), X_{-j}^{(\ell)} \right) - 1 \right) \right] \right.$$
$$\left. + \lambda_S \sum_{ij} a_{i,j} + \lambda_S \sum_j \left\| \theta_j \right\|_F + \lambda_D \sum_{k=1}^{d} \frac{trA^k}{k!} \right)$$

5.2.3 Structure, loss function and learning of GANs-based causal relationship exploration model

The GANs comprises a discriminator and a generator. The generator's role is to construct a causal graph between variables by simulating the causal relationship generation process. The discriminator, on the other hand, is responsible for distinguishing between generated data and real data. Unlike image-oriented adversarial generative neural networks, when exploring causal structures, one variable is transformed into a noisy variable while the others remain as real variables, as depicted in Figure 16.

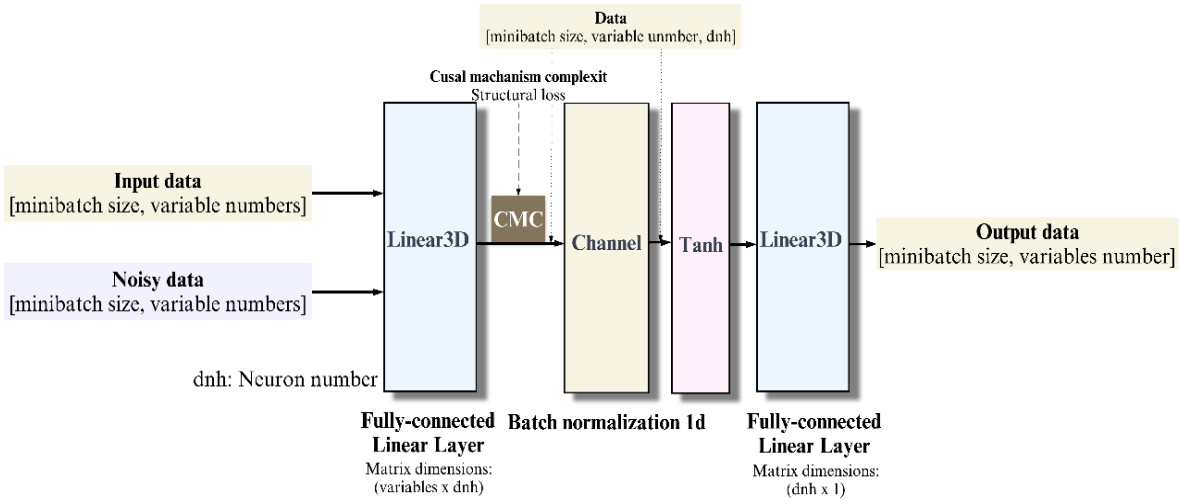

Figure 16. Generator structure of GANs model

The generator network includes a Linear 3D model, ChannelBatchNorm 1d, the Tanh activation function, and another Linear 3D model. The Linear 3D model represents a fully-connected linear layer, incorporating both input data and noisy data for the calculation of linear combinations. The output data is three-dimensional [minibatch size, variable numbers, dnh]. This layer has dnh neurons, each representing a $\widehat{f_j}$, with the weight values being a_j and θ_j, respectively. Subsequently, the complexity of causal mechanisms is calculated to measure structural loss. Batch-Channel Normalization is employed to normalize the data, ensuring a mean of 0 and a variance of 1. This normalization not only performs batch normalization but also uses batch information to prevent models from approaching "elimination singularities" (Qiao et al., 2019). The Tanh activation function introduces non-linearity, and finally, the

Linear 3D model performs a linear transformation to obtain the final data [minibatch size, variable number]. To maintain training stability, Batch-Channel Normalization, Tanh, and LeakyReLU are applied to both the generator and discriminator.

The discriminator. is composed of a network designed to handle mixed data, which includes both generated and real observation data. Taking the real data as an example of 5 variables, the mixed data is a generated fake in which one of the 5 variables in the real data is replaced. data. The network structure is depicted in Figures 17.

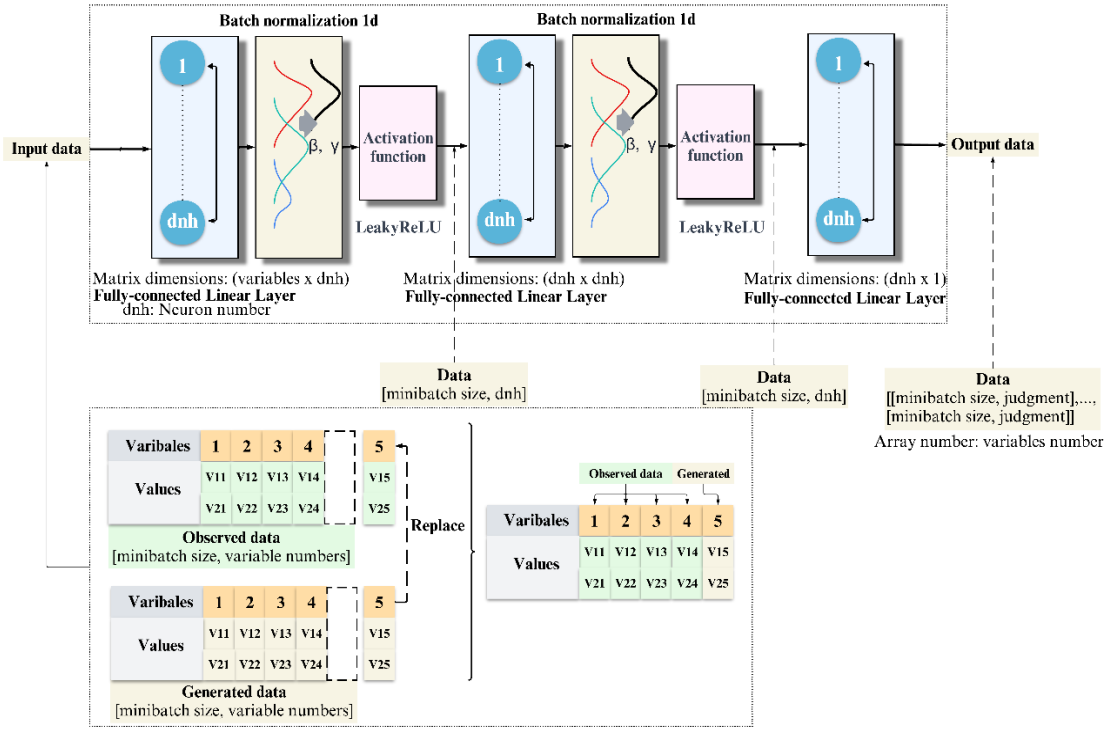

Figure 17. Discriminator structure of GANs model

The discriminator network consists of a sequence of components from left to right: a fully-connected linear layer, batch normalization 1d, and LeakyReLU. This sequence is executed twice before reaching the final output layer, which is another fully-connected linear layer. The neural network takes [minibatch sizes, variable numbers] as input data, and its output is an array with several elements equal to the variable numbers. Each element has a shape of [minibatch size, judgment], where judgment represents the discriminator's evaluation. This evaluation corresponds to the probability value indicating whether a sample is true or false. The fully-connected linear layer contains dnh neurons, and its matrix dimension is variables × dnh. To enhance the training process, batch normalization 1d is applied, which regularizes the data by zero-mean convent and facilitates more stable gradient propagation, thereby accelerating convergence. LeakyReLU, serving as an activation function, ensures a gradient of 1 when the input is greater than 0 and 0 otherwise, effectively preventing gradient vanishing or diminishing. Through this iterative process, data with dimensions [minibatch sizes, dnh] are transformed into the output data by the final fully-connected linear layer, resulting in data with dimensions dnh × 1.

According to the SAM algorithm and network structure, the SAM algorithm uses the global penalized min-max optimization, as defined as below, when calculating loss optimization.

$$L^n(\hat{g}^*, \theta^*, D) = \min_{A,\theta} \left(\lambda_S \sum_{i=1,j=1}^{d} a_{i,j} + \lambda_S \sum_{j=1}^{j=1} \|\theta_j\|_F + \sup_{\omega \in \Omega} \left(\frac{d}{n} \sum_{\ell=1}^{n} T_\omega \left(X^{(\ell)} \right) + \frac{1}{n} \sum_{j=0}^{d} \sum_{\ell=0}^{n} \left[-exp\left(T_\omega \left(\hat{f}_j^{\theta_j, a_j} \left(X^{(\ell)}, e_j^{(\ell)} \right), e_{-j}^{(\ell)} \right) - 1 \right) \right] \right) \right)$$

Where generator $\hat{f}_j$ is related on both parameter θ_j and a_j, this minimization is performed by the two parameters. In the equation, the maximization is carried out by the discriminator T_ω using the estimated fit terms $D_{KL}[p\|\tilde{q}_j]$ for variables, thereby which is to evaluate the global fit loss.

Model learning. When learning the model, not only the global penalized min-max loss optimization problem must be considered, but also the acyclicity of causal graph, as shown in the formula. To avoid the computational problems caused by the need to retrain each new network graph, the SAM algorithm introduces stochasit gradient descent and Binary Concrete relaxation approach to solve a learned dropout of edges of the neural network.

5.3 An application of GANs-based causal relationship explorations: an example analysis of the case of shallow reading behavior

5.3.1 Identifying shallow reading behavior

5.3.1.1 Identifying model of shallow reading behavior

Shallow reading behavior can be identified based on three primary behavioral characteristics that established by previous research shown in figure 18: Reading speed, shallow readers tend to read content at an accelerated pace. Liang and Huang (2014) set a benchmark of 400 words per minute. If exceed this benchmark, behavior is considered shallow reading. Page skipping is characterized by rapid page turning without focusing on effective memorization or in-depth understanding. Ms. Spadden (2015) found that it takes a minimum of 8 seconds to achieve effective memory. This benchmark helps assess whether readers are turning pages without the intention of retaining information effectively. Non sentence reading, the third distinguishing feature is fragmented reading characterized by a tendence to focus on phrases and individual words rather than complete sentences. Additionally, the importance of the content in fragmented reading is generally low. In this context, TF-IDF algorithm was used to gauge the significance of reading material.

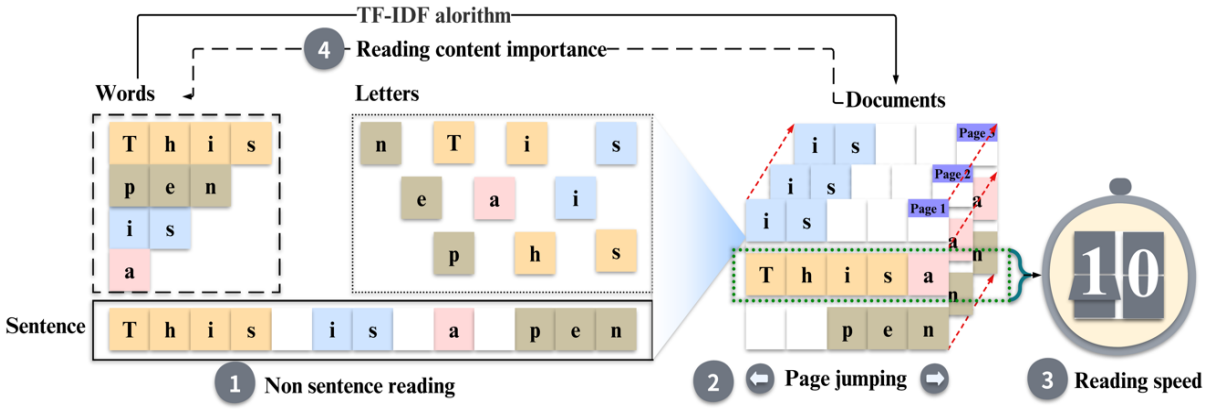

Figure 18. Shallow reading behavior identifying model.

5.3.1.2 Identification process of shallow reading behavior

Figure 19 depicts the process of identifying shallow reading behavior, with a crucial 8-second threshold to detect page jumping. If this criterion isn't met, the assessment shifts to sentence-level reading. Within this, reading behavior focusing on words and phrases with a speed of 400 words per minute or more, is considered fast speed in shallow reading. Slower speeds are categorized as non-sentence reading in shallow reading. Speed exceeding speed threshold at the sentence level are classified as fast reading in the context of shallow reading, while lower speeds are non-shallow reading.

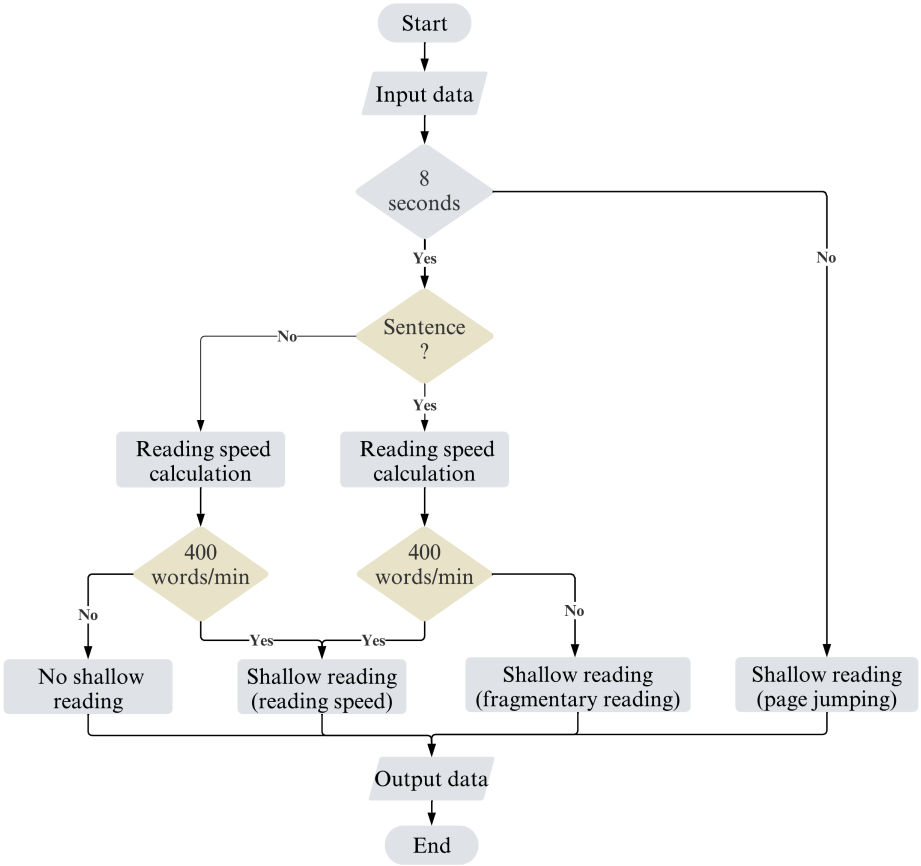

Figure 19. Identifying process of shallow reading

5.3.1.3 Data collection tool and strategy for shallow reading behavior

This study utilizes an e-book system as the data collection tool shown in figure 20.

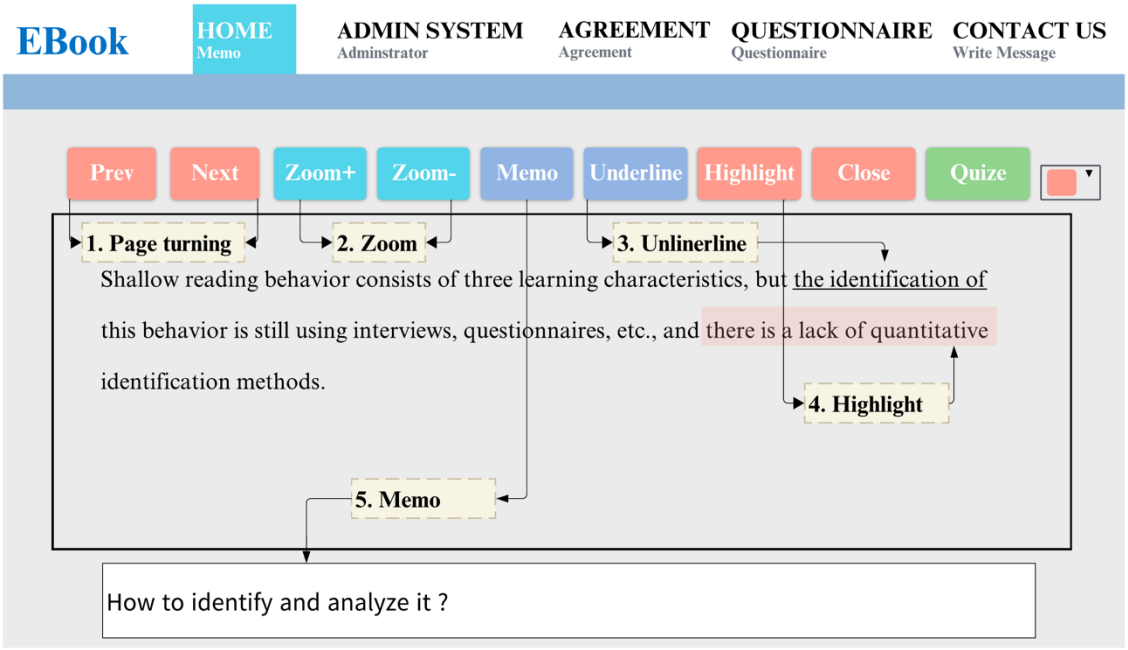

Figure 20. *Ebook system interface*

It offers various reading functionalities such as prev, next, underline, highlight, memo, and records these reading actions in log format. The e-book system serves a dual purpose: facilitating quantitative assessment of shallow reading by extracting specific reading content including text and word count, and measuring reading duration, which includes both start and end times.

Table 22. Data examples of shallow reading behavior

Id	Page	Reading speed	Label	Reading time	dwell	Label	Word number	Label	TF-IDF	Label	Score
1	1	1.23	RS	7s		PJ	57	WNNS	0.097	INS	80
2	1	13.54		20s			27		0.15		70

5.3.2 Evaluation and data analysis of GANs-based causal relationship explorations

5.3.2.1 GANs-based causal relationship exploration implementation

PyTorch served as the foundational model for GANs implementation. For enhanced computational power during model training, we leveraged GPU resources generously provided by Google Colaboratory. The primary concern addressed here is the instability observed in the results of the GANs. The figure 21 illustrates the distribution of discriminator loss, generator loss, and structural loss within the GAN model. Notably, the GAN model exhibits a consistent trend in structural loss, indicating that the generated causal relationships align with the DAG and accurately represent the influence of causal links between elements.

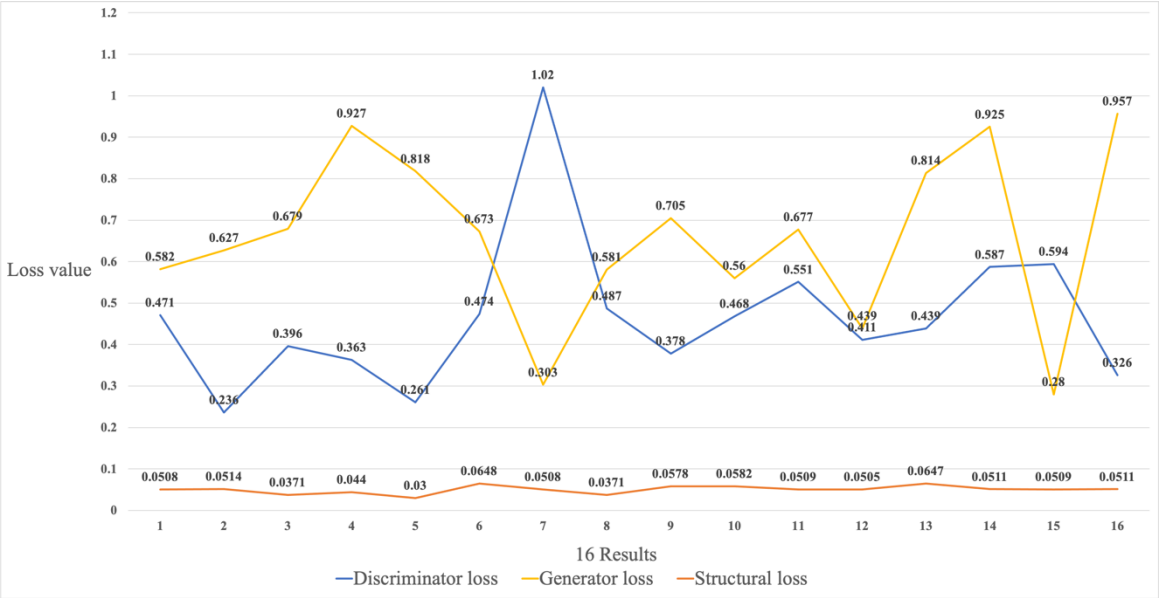

Figure 21. Loss distributions of GANs model

However, the distributions of generator and discriminator loss exhibit significant fluctuations, with a maximum loss difference of 1.323 observed in the 7th result. This variability points to instability during the training process, which, in turn, affects the stability of the causal results. This outcome can be attributed to the delicate balance inherent in GANs models between the generator and discriminator components. When one dominates the other, causing slow convergence, it can result in instability (Goodfellow et al., 2014).

5.3.2.2 Determine the stability of the causal relationship exploration results

The stability of these results can be influenced by several factors including model collapse, training instability, vanishing gradients, and data quality, which in turn, impact the accuracy of the generated outcomes (Radford et al., 2015). Typically, GANs evaluation metrics are employed to gauge the stability of model results (Spirtes & Zhang, 2016). Among these metrics, the maximum mean discrepancy (MMD) shown as following equation stands out as a kernel-based measure used to quantify the distinction between the distributions of two given samples (Gretton et al., 2012). When compared with other evaluation metrics for GANs, such as the Inception score, Wasserstein distance, Frechet inception distance, and 1-Nearest Neighbor, MMD's ability to assess the distance between samples can effectively accommodate most models (Xu et al., 2018).

$$\mathrm{MMD}\left[\mathrm{p_x}, \mathrm{p_y}\right] = \left\| \frac{1}{\mathrm{n}^2} \sum_{\mathrm{i}}^{\mathrm{n}} \sum_{\mathrm{i}'}^{\mathrm{n}} \mathrm{k}\left(\mathrm{x_i}, \mathrm{x_{i'}}\right) + \frac{1}{\mathrm{m}^2} \sum_{\mathrm{j}}^{\mathrm{m}} \sum_{\mathrm{j}'}^{\mathrm{m}} \mathrm{k}\left(\mathrm{y_j}, \mathrm{y_{j'}}\right) - \frac{2}{\mathrm{nm}} \sum_{\mathrm{i}}^{\mathrm{n}} \sum_{\mathrm{i}'}^{\mathrm{m}} \mathrm{k}\left(\mathrm{x_i}, \mathrm{y_{j,}}\right) \right\|_{\mathrm{H}}$$

p_x, p_y represent the probability distributions of samples x and y, respectively, while $k(x_i, x_{i'})$ and $k(y_j, y_{j'})$ are kernel functions. These kernel functions serve to compute the average similarity between all pairs of data points in the sets x and y. The paired kernel evaluations are then summed and averaged to quantify the self-similarity within these sets.

Additionally, $k(x_i, y_{j,})$calculates the cross-similarity between data points in sets x and y, while $\| \ \|_H$ signifies the norm in the Hilbert space. This norm measurement is instrumental in

gauging the distance between distributions within the high-dimensional feature space. Generally, $0 \leq \mathrm{MMD}\left[p_x, p_y\right] < \infty$. When MMD is close to zero, it indicates a high similarity between p_x and p_y. Conversely, larger values signify greater disparities between the distributions.

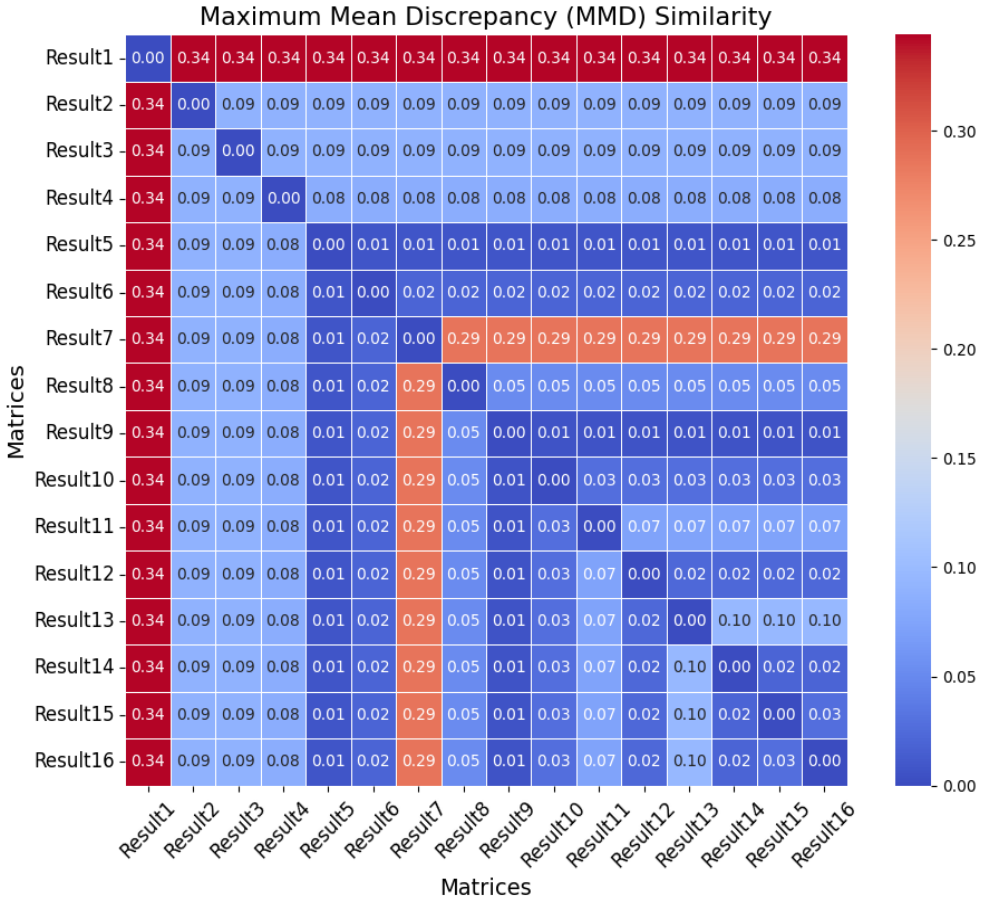

Figure 22. MMD similarity of 16 causal results

As depicted in Figure 22, the MMD results for the 16 causal outcomes are shown. Across the board, MMD values range from 0.01 to 0.34, and these values are relatively small overall. Among all the causal results, apart from Results 1 and 7, which have MMD values exceeding 0.1, the remaining results predominantly fall within the range of 0 to 0.1, indicating extremely low dissimilarity. This analysis of the graph reveals a high degree of similarity in MMD among the 16 causal relationship results. The small overall MMD value only has slight deviations in two results, which exhibit slightly higher values. The MMD distances between these two results and the others are relatively large, further confirming the instability of the model-generated outcomes.

5.3.2.3 Analysis of the causal relationship exploration results

Assuming the stability of both the model and analysis results is acceptable, a practical approach involves averaging multiple calculations. For instance, following the methodology employed by Kalainathan et al. (2022), who ran the model 8 times and used the average of the calculated results as the outcome, this study took a more rigorous approach and learned the model 15 times, subsequently computing the average of all results (Gawa, 2020).

Table 21. Probability of causal relationship between variables in shallow reading behavior

	RS	PJ	WNNS	INS	Score
RS	0	0.67375	0.37875	0.15625	0.07125
PJ	0.325	0	0.633125	0.613125	0.2675
WNNS	0.25875	0.52625	0	0.946875	0.163125
INS	0.335625	0.493125	0.27375	0	0.68875
Score	0.298125	0.35375	0.133125	0.18875	0

Table 21 shows a probability threshold of 0.6 is set to determine causal relationships. Any value surpassing this threshold indicates a higher likelihood of a causal relationship. Building upon this criterion and utilizing the causality probability values computed by the model across 16 runs, the causality matrix relationship values are then aggregated and averaged. The conclusive results are presented in Table 2, encompassing shallow reading behavior characteristics such as RS, PJ, WNNS, INS, and Score. Utilizing this table as a foundation, a causal relationship diagram is generated, denoted as figure 8. Within this figure, the arrows denote the method of causation, while the numerical values represent the probability of the causal impact.

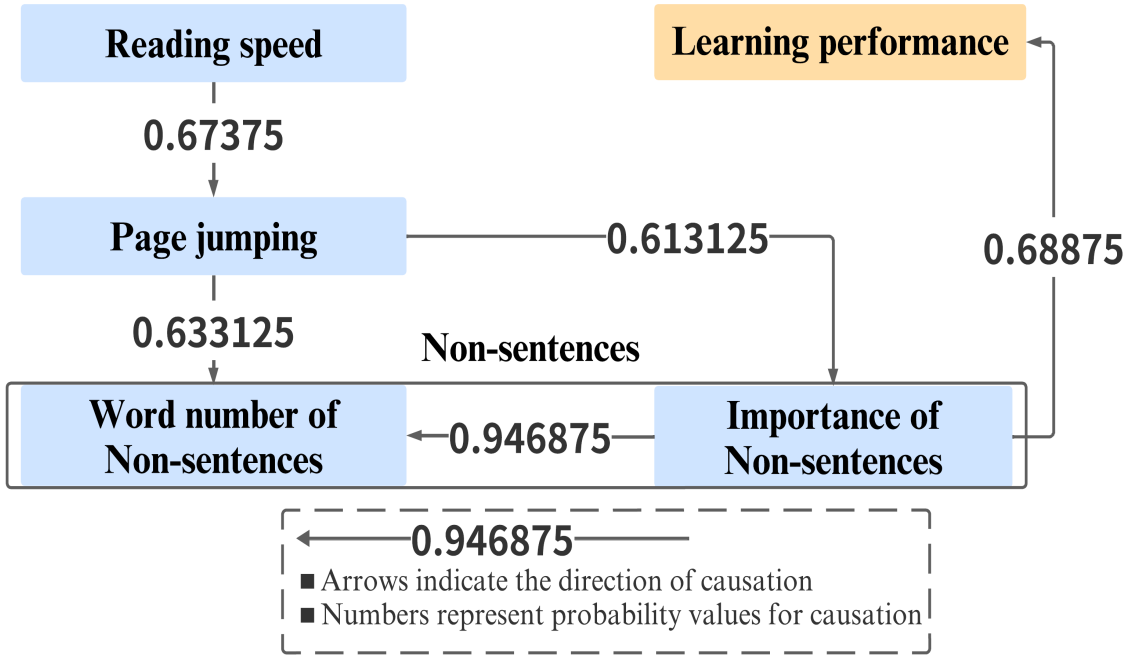

Figure 23. Causal relationship between variables in shallow reading behavior

In Figure 23, two significant causal relationships are evident: First, there is a causal link between shallow reading behavior and academic performance, with a probability of 0.68875, indicating the influence of shallow reading behavior on academic performance. Specifically, the characteristic of shallow reading, “Non sentence behavior”, directly impacts academic performance within this relationship. Second, causal relationships exist among various behavioral aspects of shallow reading. The figure shows that “Reading Speed” (RS) is the causal factor for “Page Jumping” (PJ) with a probability of 0.67375. Furthermore, PJ, in turn, affects the "no sentence behavioral characteristics." Within the latter, which includes two sub-behaviors, “Word Number of Non-sentence” (WNNS) and “Importance of Non-sentence” (INS), PJ forms causal links with them, with probabilities of 0.633125 and 0.613125, respectively. Notably, INS within the “Non sentences behavior” serves as the causal factor for WNNS, with a probability of 0.946875.

5.3.3 Results

Quantitative identification of shallow reading behaviors using online reading platforms is essential for analyzing behavioral causal relationship exploration. Current methods often use larger reading activities or presentation units, like course-level reading behavior or page-level textbook content. This is partly due to the limited accuracy of identification methods. Interviews and self-reports tend to rely on memory and can introduce errors. While high-precision technologies like eye tracking exist, they often focus on reading activities employing shallow reading strategies, not the behavior itself. This study employs an ebook reading system to collect reader-generated log data, combining three core behavioral characteristics of shallow reading (speed reading, page jumping, and non-sentence behavior) for identification. The structured log data and automated identification save time and minimize interference with the learning process, improving accuracy.

In GANs-based causal analysis, the primary challenge arises from model output instability. This instability may stem from data quality, model collapse, or vanishing gradients. To address these issues, this study employs ChannelBatchNorm 1d and the Tanh activation function for data feature extraction. Additionally, it explores global penalized min-max optimization in the SAM algorithm during model training to enhance stability. Evaluating the results of causal relationship exploration results presents a challenge due to the absence of true samples with causal relationships, rendering traditional metrics like Inception score or Wasserstein distance unsuitable. This study addresses this by averaging 16 results and using MMD to assess the difference in causal structure. The average is employed to estimate causal relationships with minimal result distribution differences.

Analyzing the causal relationship between shallow reading, learning performance, and various behavioral characteristics reveals that reading speed is fundamental among shallow reading behaviors. However, its impact on learning performance is mediated through other factors, particularly the importance of reading content. Reading quickly, on its own, does not directly affect learning performance. The significance lies in what students read, emphasizing the importance of content in the learning process.

6 Application of causal relationship exploration results: precision behavioral interventions

Evidence-based education aims to use empirical evidence, research, and data to provide evidence for educational decisions and behavioral interventions before they are made. First, Slavin (2002) argues that the development of instructional strategies, the implementation of interventions, and the introduction of educational policies should not be based on untested experiences, but need to rely on empirical research and data analysis. Secondly, as mentioned earlier, in the analysis of Education Big Data, causal relationship exploration mainly assumes the role of analyzing the relationship between variables. Therefore, when the causal mechanisms and relationships in different variables of the data are fully understood, behavioral interventions can be factually accurate (Yang, Chen, & Ogata, 2021). Also, when behavioral interventions are made, causal relationship exploration can be used to evaluate the effects of behavioral interventions and strategy applications. Third, since causal relationship exploration serves as one of the methods for exploring relationships in Education Big Data or learning analytics, the application of its results also belongs to the components in both models. Based on the performance of the results of the existing studies, there are no specific guiding steps for the application of results given in the models of either Education Big Data or Learning Analytics.

6.1 Requirements for the application of Education Big Data mining results

6.1.1 Evidence-based education applications and challenges in Education Big Data mining

Hargreaves (1997) first used the term evidence-based education (EBE), which was inspired by evidence-based practice in medicine. Like the medicine field, teaching is also a vocational field, in which decisions by teachers involving complex judgment need to be supported by their professional wisdom and relevant evidence (Haberfellnerm & Fenzl, 2017). The EBE was considered as a way of bridging the research-to-practice gap in teaching (Bauer & Prenzel, 2012), in an effort to make instructional programs and practices driven more by evidence than ideology, faddism, politics, and marketing (McKnight & Morgan, 2020). Inspired by evidence-based practice in medicine, EBE is an innovative approach that focuses on best available evidence, judgment and implementation for education and teaching, and value to students, administrators, and teachers.

Previous research has identified that EBE has contributed to a link between research and practice in teaching, with an attempt to move from individualistic and personal ways for determining instruction intervention practices to one in which all instruction practices are evaluated based on knowledge and ideas on scientific evidence (Cutspec, 2004). This approach helps to practice with sound evidence that is generated and evaluated from research (Bruniges, 2005). Additionally, the principle of sound evidence is constantly emphasized (Detrich & Lewis, 2013) when making, applying, and evaluating (Slavin, 2008) teaching skills, methods and strategies. As a result, this not only offers a guarantee for the correctness and usefulness of instruction activities but also brings more fresh practice attempts as the types of evidence to become more abundant (Moran & Malott, 2004).

Concerned with these benefits of EBE, many organizations and programs are committed to the development of EBE with numerous valuable findings. The Campbell Collaboration was launched in 1999 focusing on systematic reviews, in terms of educational effectiveness and

social policy practices. Three approaches to completing evidence-based education by Evidence-based Education UK, include evidence-based policies development, practices promotion, and the advocacy of a culture of evidence. To promote student academic performance, the Center for Evidence-based Education developed a systematic procedure for investigating literature evidence to determine what characteristics make students successful. As one of the important implementation methods of the No Child Left Behind Act of 2001, the EBE was selected as a strategy to improving student learning and intervening learning behavior (Moran & Malott, 2004).

Despite the fact that a growing body of research in EBE has been received. So far, however, some emerging difficulties and challenges surrounding EBE should be paid enough attention to, as well as addressing ways. The primary difficulty is to define the relevant evidence (Berliner, 2002; Kvernbekk, 2019), while it is practically impossible to support all instruction activity with one-size-fits-all evidence (Biesta, 2010; Farley-Ripple et al., 2018). With the increasing popularity of education big data (EBD) and learning analytics (LA), EBE will provide new opportunities as well as unknown challenges, in terms of obvious benefits and potentially negative factors (Yin & Hwang, 2018; Zhao, Hwang, & Yin, 2021) that they cause to stakeholders in the education field.

6.1.2 Low application effect in analysis result application

The application of analysis results in LA is a key step, aiming to apply the analysis results to educational practice for monitoring, prediction, intervention, assessment, adaptation, personalization, recommendation, and reflection (Chatti et al., 2012). To achieve the intended application targets, it is critical to fit the analysis results into implementations of education activities when developing an application.

However, having good analysis results is not always successful in facilitating educational activities, and it is especially crucial to apply them correctly (Viberg et al., 2018). Hanna (2004) attributes this to the ambiguity of the analysis results: the results may yield useful insights and clues without providing definitive answers. Furthermore, Saks, Pedaste and Rannastu (2018) found that the ambiguity of the analysis results has caused most users to hold an ambiguous view of application effectiveness. On the other hand, most studies have concentrated on how to meet application requirements along with the advent of these learning scenarios, tools, and data. In contrast, little attention has been given to concrete steps to use the analysis results to create an optimal application strategy.

6.1.3 Challenges in the application of Education Big Data mining results

Although studies of LA models have been undertaken, they have not examined the details of the implementation of the analysis result application in those models. Campbell, DeBlois and Oblinger (2007) initially proposed the “Five Steps of Analytics” model, comprising the steps “capture,” “report,” “predict,” “act,” and “refine.” The term “act” refers to the application of analysis results. As a type of intervention, they suggest supporting the above applications with a personal phone call or e-mail.

As the introduction of LA to various learning environments increased, different perspective-based explorations on the model also saw urgent development. Dron and Anderson (2009) defined their “Collective Application Model” model taking into consideration the characteristics of e-learning. As it highlighted information gathering, processing, and

presentation, the analysis result application step was excluded from the model, which comprised only "select," "capture," "aggregate," "process," and "display." After a comparison of existing models, Elias (2011) proposed a comprehensive model comprising "select," "capture," "aggregate and report," "predict," "use," "refine," and "share." Regarding the application step, it only provides a brief description of attempts to improve the learning system.

The Chatti model (Chatti et al., 2012) incorporates analysis and its corresponding application into one step to compose the processing step, with the remaining two steps pre-processing and post-processing. In contrast, Siemens (2013) not only added an application step to his model but also highlighted the purposes of the step, such as "intervention," "optimization," "alters," "warning," "guiding" / "nudging," "systemic," and "improvement." Two issues exist in these models. First, the reason for the analysis results cannot be sufficiently understood due to the lack of confirmation of the analysis results by AI algorithms, reducing the explanation and trustworthiness of analysis results. Second, existing models cannot provide specific ways to apply the analysis results of AI algorithms, as shown in Table 22. In this context, analysis results cannot be successfully translated into the proven application strategy and thus cannot effectively support human welfare. The former was addressed in the ReCoLBA process (Zhao et al., 2021), and the latter is the focus of this study.

Table 22. Features and drawbacks of the application step in the existing models

Title	Features	Drawbacks
"Five Steps of Analytics" framework	Specific intervention approaches (phone calls or e-mail)	Limited application range
"Collective Application framework"	Skips the application step	Lack of function
Elias' framework	Only defines the application	No specific guidance steps
Chatti's framework	Considers the application a sub-step of data processing; no other details are available	The independent properties of the application are not established
Siemens' model	Describes the purposes of the application	Lack of application methods
ReCoLBA process	Describes the stakeholders of the application	Lack of specific guidance

6.2 Strategies and steps for applying causal relationship exploration results

6.2.1 Big data-driven education application strategy development

The term evidence-driven education (EDE) was first used by Hargreaves (1997) and was inspired by evidence- driven practice in the field of medicine. EDE aims to bridge the research-to-practice gap in teaching as well as to shift the driving force of instructional programs and practices to evidence rather than ideology, faddism, politics, and marketing (Davies, 1999). EDE is supported by the findings of multiple, high-quality, experimental studies (Cook et al., 2008) and quantitative analysis (Moran & Malott, 2004), which provides sound evidence that an educational practice truly works. Kuromiya et al. (2020) used evidence obtained from

reading behavior analysis conducted when students were learning via a learning management system to evaluate whether a new learning intervention has a positive learning effect.

6.2.2 Steps for applying the results of evidence-based education analyses

To find the optimal application strategy, an evidence-driven education policy was employed in the application-step sequence, which is responsible for building a corresponding link between the identified learning strategy and the analysis result. It is the evidence-driven education policy that establishes the hypothesis of matching between the analysis results and the application strategy. Hence, the application-step sequence helps instructors to find the optimal application strategy. In addition, the evidence for constructing hypotheses with the optimal applied strategy is based on the analysis of data from any learning scenario and is not for any specific one. Therefore, the application strategy obtained can be applied to any learning scenario. Four steps were designed in the application-step sequence, including identification, presentation & demonstration, participation, and assessment. The hypothesis between analysis results and application strategies is fulfilled in the identification step, which reflects the design concept of evidence-based education.

Identification. Since the existing models lack specific application guidance, especially are not available to the function of strategy identification. This step aims to find an optimal learning strategy suitable for the confirmed analysis results. Thus, the degree to which the identified learning strategy matches the confirmed analysis results determines the application effect.

Presentation and demonstration. This step allows the instructor to disseminate information to learners using verbal information in writing, and visual symbols. It aims to gain learners' attention, inform learners of objectives, and combine new strategies with prior knowledge (Baldwin-Evans, 2006). When illustrating, concise and concrete steps for implementing a strategy are crucial.

Participation. This application step allows the learner to use the identified strategy to affect learning achievement. In this step, learners are asked to participate in a learning scenario and retry the strategy until they can use it successfully in real-world practice (Baldwin-Evans, 2006; Dick et al., 2015).

Assessment. This step has two dimensions. The first is to provide accurate feedback about learner performance when using this strategy, focusing on the academic achievement in the learning content (Dick et al., 2015). The second is to measure the cognitive load on the learner, which can reveal the impact of the learning strategy. The combination of academic performance and cognitive load measures is considered to provide a reliable estimate of the efficiency of instruction methods (Paas et al., 2003).

6.3 Application of causal relationship exploration results: an example analysis of back tracking behavior

To examine the effectiveness and contribution of the proposed application strategy, a case study was conducted using two independent experiments. The first experiment is designed to analyze reading behavior data. The learning behavior that contributes most to learning achievement in the experiment is expected to be found and will be used as evidence for an application strategy to intervene in the learning behavior. To verify the effectiveness of SQ3R, which is

hypothesized based on the evidence that SQ3R could promote students to turn back to read pages, another experiment will be conducted.

6.3.1 Experiment for confirming the proposed method effectiveness

A total of 234 freshmen were recruited, and gender distribution was 65 males and 169 females, with an average of 19 years old. Five experimental steps were employed according to the HAILA model, namely, (1) collecting students' reading behaviors using an e-book system, (2) using a machine learning library (scikit-learn) to process raw data, (3) utilizing classification prediction algorithms and feature importance calculation methods in the analysis step, (4) adopting different algorithms to confirm the analysis results, and (5) applying an identified learning strategy.

6.3.1.1 Data collection by e-book system

An e-book system was developed (Yin et al., 2018) to capture learners' reading learning behavior. Learners can conveniently read materials using several operating tools, such as (1) turning the page forward or backward, (2) resizing the view by zoom, (3) making a memo, (4) adding or removing underlines in the learning material, and (5) making or deleting highlights with a variety of color options, as shown in Figure 24.

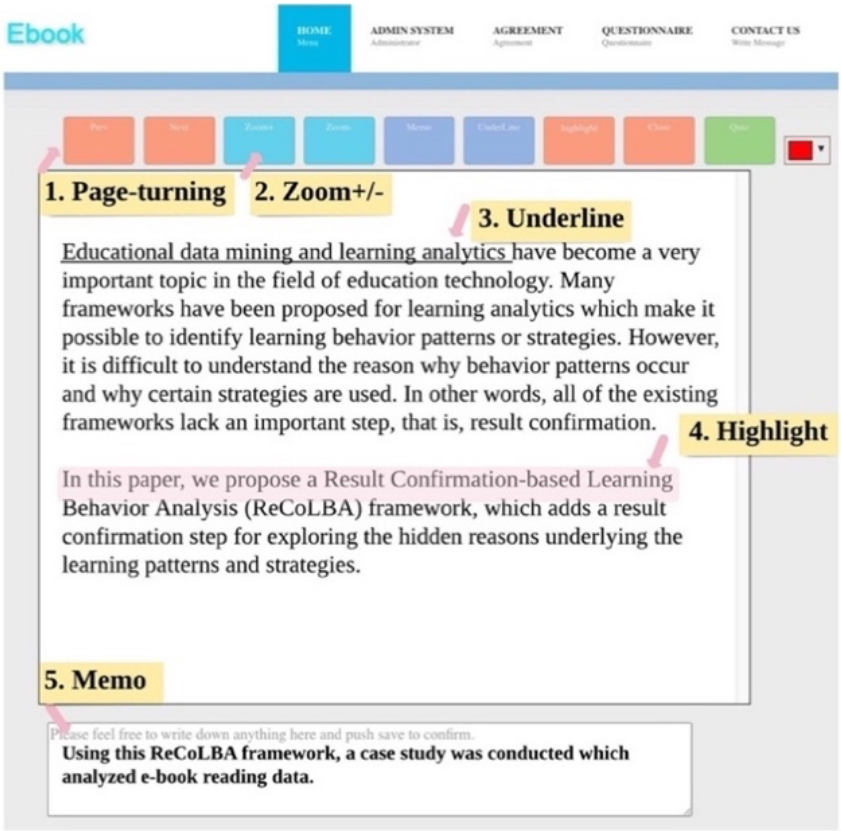

Figure 24. E-book system interface

All the learning behavior observed in the learning activities were recorded in a data log with 12 data features, as shown in Table 23. Among the e-book features, the most used were the tools associated with basic reading behaviors, such as Next (NE), Prev (PR), Highlight (HL), Underline (UL), Memo (ME), Bookmaker (BM), Read time (RT), and Read page (RP). In addition, the BacktrackRate (BR) feature is a statistical ratio dividing Next and Prev, which

provides a new view on repeated learning behavior. The equipment used for learning, such as PC, mobile, and tablet, was adopted as a parameter.

Table 23. Samples of reading behavior data by the e-book system

Id	PC	Mobile	Tablet	BR	ME	HL	UL	PR	NE	RT	RP	BR
1	0	1	0	7	1	29	3	9	12	60	51051	0.75
2	0	0	1	8	0	20	4	79	96	107	42043	0.823

6.3.1.2 Data processing using scikit-learn

The data processing steps were divided into two categories. The first consisted of basic processing, mainly involving missing values removal to maximize the validity of data, and standardization to unify the values of each data feature. As a result, valid data from 229 participants remained, and all data values were compressed from 0 to 1. The second category focused on the analysis method–specific data preparation. Considering the demands of data balance for binary classification prediction (Krawczyk, 2016), we adjusted the passing score 60, which had a huge gap in proportion, to 70. In addition, splitting data into train- and test-datasets is necessary as part of regular data processing. The *train_test_split* function built into scikit-learn was adopted to achieve the specific ratio of splitting 3:7, in which the training dataset accounted for 70%.

6.3.1.3 Data analysis by decision tree model

In this study, the decision tree model was used to explore which data feature contributes most to predicting academic performance (Hamoud et al., 2018; Mesarić & Šebalj, 2016). After the prediction model was created and tuned by scikit-learn, we obtained the final model with five metrics: accuracy (0.739), F1-score (0.763), recall (0.763), precision (0.763), and AUC (0.81). Following the rule of thumb, the predictive performance of this model can be viewed as fair. Note that an AUC score over 0.8 represents a good comprehensive performance in a model's effectiveness.

Scikit-learn offers an impurity-based feature importance calculation function oriented to the decision tree model. Subsequently, the 12 data features were analyzed using the feature importance calculation function. As fundamental splitting parameters to calculate the feature importance for classification prediction, Gini Index and Information Gain prevail in the splitting criteria area. Raileanu and Stoffel (2004) found that there are no obvious differences by comparing the efficiency of splitting features for tree models. Moreover, Gini Index has proved to be better than the other splitting parameters specifically for unbalanced datasets (Park & Kwon, 2011). Since our sample size of students who fail the exam is unbalanced in contrast to those who passed, accounting for a minority, the Gini Index was adopted to calculate the probability weighting of each node in the tree model, varies between 0 and 1, where 0 expresses the purity of classification. The PC, Prev, Nex, Read time, and BacktrackRate are 0.176, 0.496, 0.134, 0.103, and 0.115, respectively, and the other features are 0. It was found that the Prev feature had the greatest impact on prediction performance. In other words, students who have Prev learning behavior are more likely to pass the exam and obtain better academic performance.

6.3.1.4 Result confirmation by a comparative method

The most effective data feature, Prev, was successfully identified for predicting students' academic performance. However, it was unclear whether this analysis result was sufficiently

accurate to obtain the same results in other algorithms. Guided by the analysis result confirmation step, a comparative confirmation method was conducted to confirm the correctness and reproducibility of the analysis results. As an innovative algorithm based on the tree model (Liaw & Wiener, 2002), the random forest algorithm outperforms the other algorithms (Breiman, 2001), and is commonly used for binary prediction models.

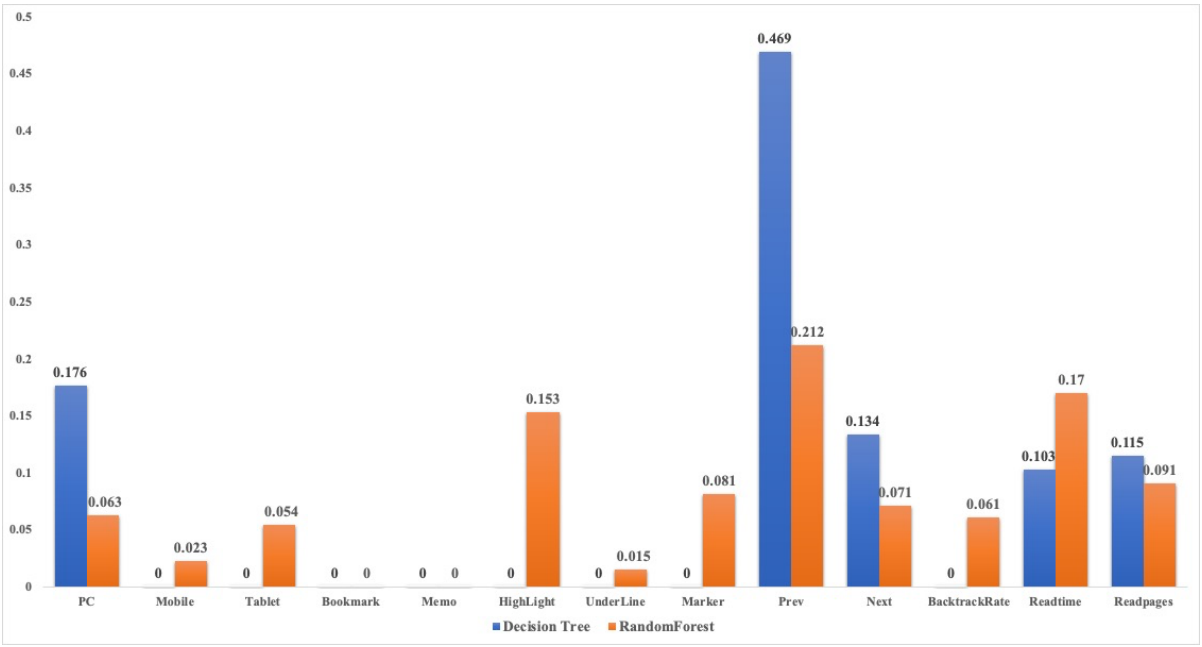

Figure 25. Importance of 12 features in predicting academic performance

In the confirmation step, the same experimental conditions were set, and only the selected algorithm was shifted from the decision tree to the random forest model. Following the rule of thumb, the prediction based on the random forest model was also acceptable in terms of accuracy (0.753), precision (0.729), recall (0.794), F1-score (0.76), and AUC (0.75), according to the results of data feature importance in the decision tree and random forest models, as shown in Figure 3. This shows that Prev parameter consistently contributes the most in the prediction of academic performance, indicating that students who show the Prev learning behavior have a higher probability of passing the exam.

6.3.1.5 Result application by SQ3R

After the researchers analyzed learning behavior by using AI algorithms, the learning behavior which could mostly contribute to learning efficiency is to be found as the evidence of strategy making. Meanwhile, a hypothesis on application strategy was presented to facilitate the improvement of learning efficiency. Firstly, the identification step is entered. The Prev behavior raised the need to increase learning frequency. Consistent with the demand in the learning frequency, we found that the SQ3R learning strategy was primarily designed for university students reading academic textbooks (Li et al., 2014). It is a reading comprehensive method comprising five steps: survey, question, read, recite, and review (Flippo & Caverly, 2009). Importantly, multiple learning steps of SQ3R can improve the occurrence of repeated learning behaviors (Huber, 2004). As shown in Figure 26, a short-term memory achieved by repeatedly turning pages is necessary, when students quickly browsing on outstanding features and writing down reading questions in survey and question steps. In subsequent steps, reciting and confirming what has been read also need to be realized by turning pages. To this end, it is hypothesized that SQ3R could facilitate the occurrence of the Prev behavior.

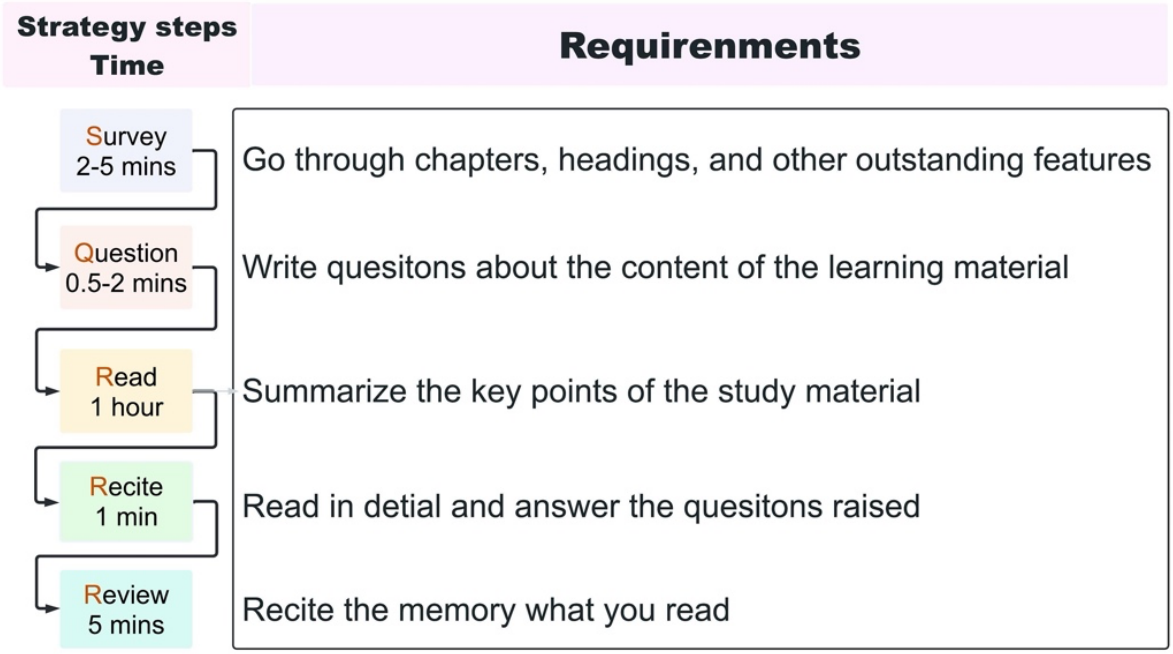

Figure 26. SQ3R schedule

In the presentation and demonstration step, an introduction of SQ3R explaining how to use it in the e-book system was provided by an instructor. In Figure 4, a schedule was also designed based on previous instruction experience. For example, the survey step is recommended to last 2 to 5 min go through outstanding features in the learning textbook by turning pages, highlighting or underlining keywords, followed by 30 seconds to 2 mins for writing down questions by memos, 1 hour for reading in detail, 1 min for reciting the memory, and 5 mins for summarizing and reviewing the learning emphasis by quickly turning pages. Subsequently, learners could undertake the participation step with the guidance of the presentation and demonstration steps. Finally, a post-test about the learning content and a test of cognitive load provide feedback to learners in the assessment step.

6.3.2 Experiment for verifying the model contribution

This study examines the contribution of this model by exploring the effectiveness of the SQ3R. Particularly, this experiment aims to evaluate the effectiveness of the SQ3R in helping students improve their academic performance. To this end, an e-book system was used to complete the above experiment to evaluate whether significant changes in academic performance and reading behaviors were related to the SQ3R application.

Participants**.** 37 male and 32 female freshmen, aged from 19 to 21 years, participated in this study. All participants were randomly assigned to two groups: the experimental group and the control group. The 35 participants in the experimental group were asked to read a learning material using the SQ3R, while the control group, which comprised 34 participants, was assigned to read the learning material without the guidance of the SQ3R.

6.3.2.1 Measuring method

The measuring method comprised of a pre-test and a post-test. Before the experiment, a pre-test was conducted to check whether the two groups had the same equivalent prior knowledge

about the learning content and the SQ3R. It consisted of 10 multiple-choice items. The post-test aimed to measure whether the SQ3R was beneficial to participants' academic performance. This test was similar to the pre-test, comprising 10 multiple-choice items related to the learning emphasis in the upcoming learning materials. The pre-test and post-test were both scored on 10.

A post-questionnaire was employed to identify participants' cognitive load when using the SQ3R, and their attitude about introducing new learning technologies into the reading environment. Based on the measurement created by Sweller (1988), a questionnaire to investigate cognitive load was modified. The developed questionnaire consisted of eight questions with two dimensions: mental load and mental effects. Each question was scored on a 5-point scale, where 5 represented "strongly agree" and 1 indicated "strongly disagree." The Cronbach's alpha values of the two dimensions were 0.91 and 0.95.

6.3.2.2 Experimental procedure

According to Figure 27, this study consisted of a pre-test, a reading learning activity using the e-book system, a post-test, and a post-questionnaire on cognitive load and learning strategy acceptance, which together took 2 weeks.

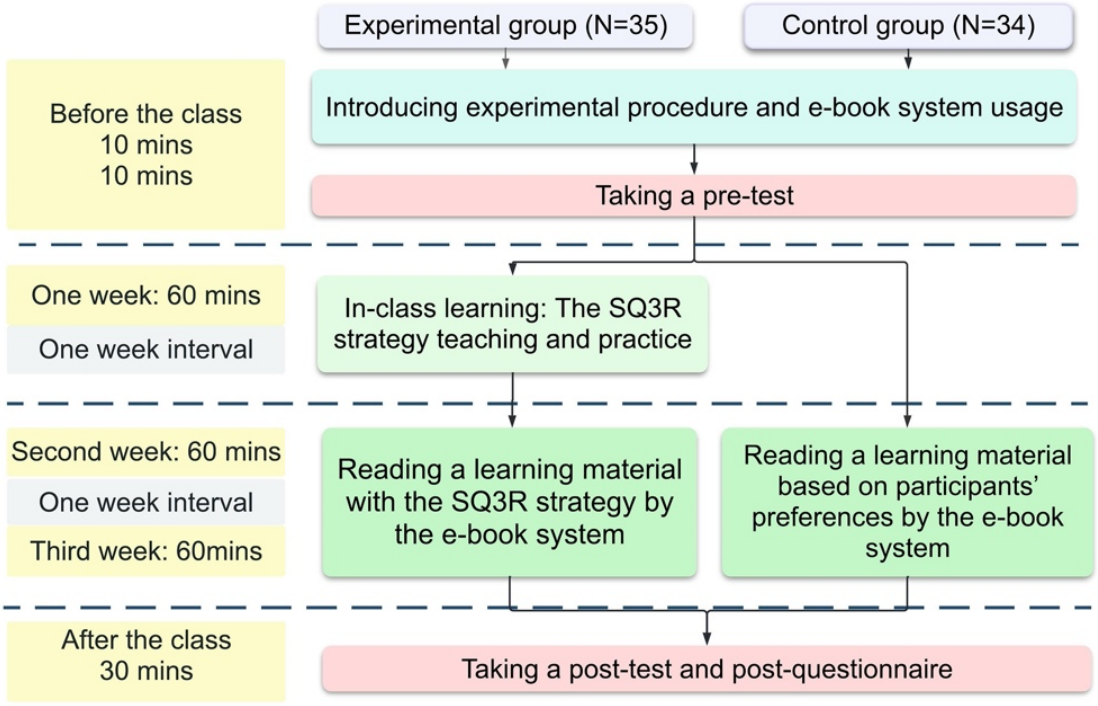

Figure 27. Experimental procedure

Initially, an orientation was provided for participants to introduce the experimental procedure and the e-book system operation and precautions. Following that, the experimental and control groups were asked to complete a pre-test to evaluate their knowledge of the SQ3R. Afterward, in-class learning of the SQ3R was only given to the experimental group within 60 mins in the first class, followed by an independent practice targeting the SQ3R proficiency during 1 week. Then, the experiment entered two learning activities of 60 mins each, with an interval of 1 week. The participants in the experimental group read the learning material using the SQ3R, whereas those in the control group read the same learning material, but the reading strategy was based on their preferences. Subsequently, a post-test and post-questionnaire were conducted for the two groups within 30 mins.

6.3.2.3 Experimental results

We firstly explored the learning behavior that contributes most to learning achievement by analyzing learning behavior in an e-book system, and use the analysis results as evidence for developing application strategies, hypothesizing that the developed strategies could promote learning achievement. Finally, the hypothesis has been verified by the following experiment. Hypothesis testing to 61 valid samples, including an independent sample t-test, and one-way analysis of variance (ANOVA) in SPSS Statistics were used to examine the effectiveness of the SQ3R for improving academic performance and affecting learning behavior, respectively.

The experimental results show that the experimental group using the SQ3R significantly outperformed the control group in terms of academic performance, at the equivalent prior knowledge. Also, regarding Prev, Read time, and Read page behaviors, the experimental group has a significant advantage in the number of occurrences. In addition, the above differences are also verified by data visualization. Finally, the cognitive load test revealed that the SQ3R did not impose additional cognitive load on students' learning.

Analysis of academic performance.

The pre-test results show the standard deviation and mean values were 2.18 and 6.67 for the experimental group, and 2.15 and 7.35 for the control group. According to the t-test results (t = -1.23, $p > .05$) (Table 24), no statistically significant differences were found between the two independent groups. Thus, all participants in both groups were known to have equivalent prior knowledge about the SQ3R.

Table 24. Descriptive data and t-test result of the pre-test results

Variable	Group	N	Mean	Std. Deviation	t
Pre-test	Control group	30	6.67	2.18	-1.23*
	Experiment group	31	7.35	2.15	

Note. *$p > .05$

After the learning activity, ANOVA was conducted to determine whether there was any statistically significant difference in the post-test results between the two groups. Using the groups as a fixed factor and post-test scores as the dependent variable, a common assessment for homogeneity of variance was performed using Levene's test. A Levene's test score above 0.05 (F = .09, P = .75) indicated that this test is robust to violations of the assumption. It was concluded that the experimental group was statistically different from the other groups. Based on the statistical results (F = 32.86, $P < .01$) in Table 25, it is seen that the participants who learned with the SQ3R showed significantly better academic performance than those who did not. In other words, the SQ3R was helpful for participants in improving their academic performance.

Table 25. Descriptive data and ANOVA result of the post-test results

Variable	Group	N	Mean	Std. Deviation	F
Post-test	Control group	30	4.50	1.77	32.86**
	Experiment group	31	7.06	1.71	

Note. **$p < .01$

Analysis of learning behavior.

First, a t-test was used to analyze the repeated learning behavior based on Prev, Read time, and Read page in two groups, as shown in Table 26. Based on the hypothesis that SQ3R contributes to the emergence of repetitive learning behavior like Prev, the SQ3R was applied in the experiment particularly to facilitate the occurrence of Prev behavior and achieve this purpose. For the Prev, the mean and standard deviation were 24.43 and 26.07 for the control group, while 60.39 and 35.71 for the experimental group. Moreover, the t-test result (t = -4.50, p < .01) showed that there was a significant difference between the two groups, implying that the SQ3R was able to promote the occurrence of Prev.

Table 26. Descriptive data and T-test results of the learning behavior

Variable	Group	N	Mean	Std. Deviation	*t*
Prev	Control group	30	24.43	26.07	-4.50**
	Experiment group	31	60.39	35.71	
Read time	Control group	30	0:49:18	0:20:31	-2.35*
	Experiment group	31	1:00:30	0:16:30	
Read page	Control group	30	78.57	61.30	-4.83**
	Experiment group	31	182.32	101.96	
Underline	Control group	30	5.40	10.19	-3.62**
	Experiment group	31	37.52	48.16	
Highlight	Control group	30	5.30	18.22	-.060
	Experiment group	31	5.52	8.57	
Bookmark	Control group	30	.43	2.37	-4.97**
	Experiment group	31	6.90	6.82	
Memo	Control group	30	.00	.000	-3.16**
	Experiment group	31	2.19	3.85	
Next	Control group	30	43.00	28.42	-3.42**
	Experiment group	31	72.00	37.04	

Note. *p < .05; **p < .01

Analyzing the Read time variable alone, the difference between the two groups is obvious, and the T-test result (t = -2.35, p < .05) also further verifies this conclusion, as shown in Table 6. However, the Read time variable was combined with the Read page variable for analysis. It is worth noting that the participants using the SQ3R spent more time than those who did not, but after taking the Read page variable into account based on the t-test result (t = -4.83, p < .01), it was found that the former read nearly twice as efficiently as the latter. On average, the experimental group read approximately 3.01 pages per minute, compared with the control group's 1.59 pages per minute.

Second, we used the t-test to explore whether the SQ3R affects other reading behaviors. The results showed that there was a significant difference between the two groups in terms of Next (t = -3.42, p < .01), Memo (t = -3.16, p < .01), Underline (t = -3.62, p < .01), and Bookmark (t = -4.97, p < .01), though not for Highlight (t = .06, p > .05). These findings reveal that the SQ3R also promoted the occurrence of reading behaviors such as Next, Memo, Underline, and Bookmark. As for Highlight, no significant difference was found between the two groups.

Third, the Figures 28 and 29 show learning patterns in terms of the time distribution of reading behaviors. The X-axis represents the time participants spent reading the material in the e-book system, and the Y-axis indicates the reading behaviors. The upper part of the two figures is the time distribution for each step of the SQ3R, where the time division between steps is based on

the application guidance designed in the model application step. Although there exists an obvious time division between various steps, this does not mean that all learning behaviors have a similar distribution to the pre-set learning steps.

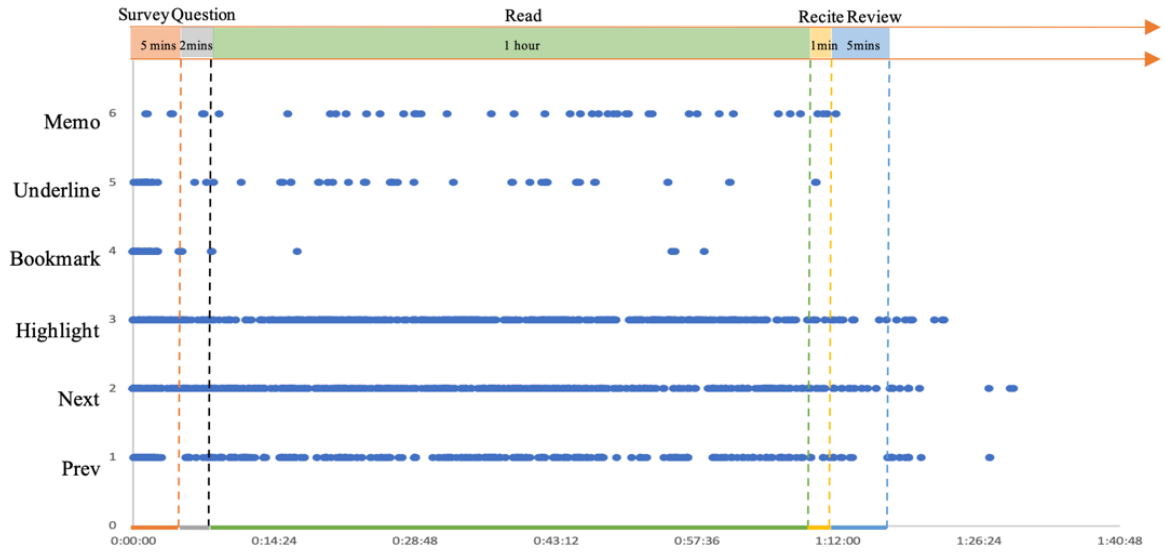

Figure 28. Distribution of reading behaviors of the experimental group in the e-book system

Figure 29 shows a rough distribution trend of reading behaviors representing three stages in terms of period: 0 min to 3 min; 3 min to 1 h 12 min; and 1 h 12 min to the end. Combining the five divisions of the SQ3R, it is seen that in Stage 1, Prev, Underline, and Bookmark appear intensively in a short period, which indicates that the participants repeatedly read the material by a survey step, marking some knowledge points in a short time. Stage 2 accounts for most of the reading behavior. This stage might include the question and reading. The reading behaviors in Stage 3 are mainly made up of Prev and Next, and other behaviors account for a smaller proportion. The time distribution of reading behaviors in the experimental group is consistent with the SQ3R, which suggests spending 2 to 3 min for the first survey step, 2 min for the question step, and 5 min each for the reciting and review step.

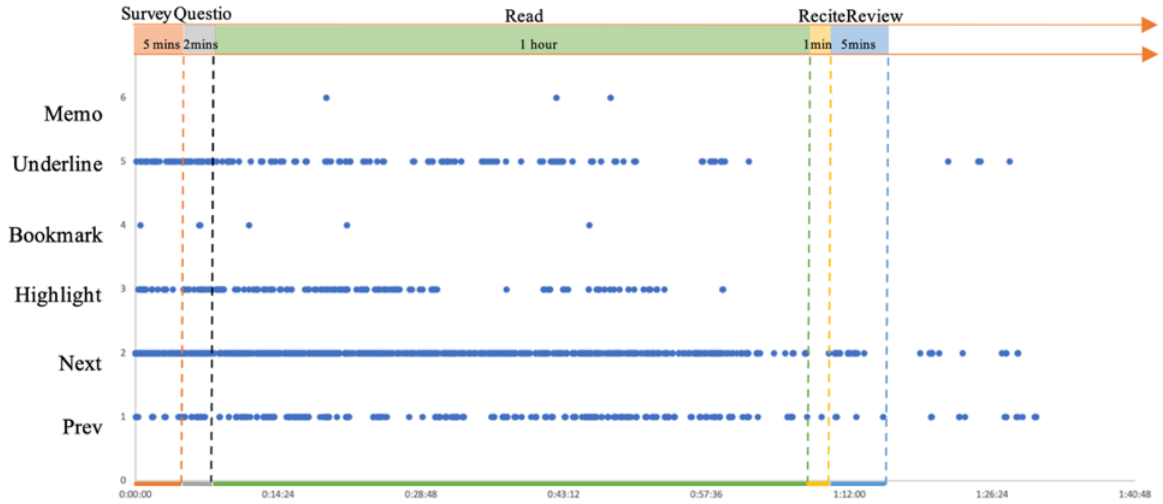

Figure 29. Distribution of reading behaviors of the control group in the e-book system

Figure 29 shows that the control group is divided into two stages, in which Prev and Next occur continuously and are the most frequent. Memo and Bookmark are less frequent, scattered, and irregular. In the first stage, Highlight and Underline appear frequently. Highlight has two distinct distributions, and it appears intensively and continuously for 30 min. Underline remains stable across several occurrences.

Analysis of cognitive load.

A cognitive load post-questionnaire, which includes two test dimensions, mental load, and mental effort, was employed to investigate the differences between the two groups, in terms of the learning pressure and load on the participants. Because the mental load effect depends on the information being processed, which imposes a heavy cognitive load (Sweller, 1988), the first dimension focuses on the pressure caused by the amount of information the participants process. In addition, the second dimension carries out the mental effort test, which reflects the controlled consumption of psychological information processing resources in the cognitive process (Sweller, 1988; Hwang & Chang, 2011).

The t-test results in Table 27 show that for the mental load dimension, the mean and standard deviation were 11.91 and 3.25 in the control group, while 11.17 and 2.35 for the experimental group. No significant difference was found between the two groups ($t = 1.01$, $p > .05$), implying that the SQ3R did not increase the pressure of amount information to participants. For the mental effort also, no significant difference was found between the two groups ($t = 1.66$, $p > .05$). Therefore, the SQ3R did not cause participants additional pressure in terms of mental load or mental effort. All participants in both groups had a moderate level of learning pressure, because the two groups had similar standard deviation concentrating at 3.

Table 27. Descriptive data and T-test result of the cognitive load results

Variable	Group	N	Mean	Std. Deviation	*t*
Mental Load	Control group	30	11.91	3.25	1.01*
	Experiment group	31	11.17	2.35	
Mental Effort	Control group	30	10.47	3.27	1.66*
	Experiment group	31	9.16	2.82	

Note. *$p > .05$

6.3.2.4 Discussion

This study explored whether the HAILA model would affect identifying optimal learning strategies. RQ1 investigated the effectiveness of the evidence-driven model to offer guidance for transforming analysis results into the most suitable application strategy. For the application step performance, the SQ3R was obtained and optimally matched with the analysis results of the first experiment. The need for the analysis result application has been proven by previous studies (Barry & Reschly, 2012), but no previous study has provided concrete guidance for implementing the application. Inspired by HAI, evidence-driven education was drawn into the design of the application step. Thus, this study explored the extent to which evidence-based education can facilitate the identification and transformation of analysis results to the optimal strategy. The experiment results show that evidence-driven education can sufficiently support application step design.

There remain negative sides of AI in LA, particularly related to algorithmic bias (Carter & Egliston, 2021). For example, decision-making based on AI analytics with unrepresentative

datasets and algorithm design bias, results in not only incomprehensible analysis results, but also inappropriate or inapplicable result application. Specifically, training AI algorithms on historical and complicated learning behavior data may reinforce the difficulty of understanding the reason which potentially undermines the learning behavior patterns. In that case, the HAILA model contributes to increasing the interpretability of learning behavior patterns, not just for exploring learning behavior patterns themselves. For example, the database on which AI algorithms is inevitably biased by gender, family background, ethnicity, resulting in the data itself containing bias (West et al., 2018). Also, due to concerns that AI analysis results can trigger artificial biases against learning behavior without one inappropriate or inapplicable application strategy, which reduces the trustworthiness of the application of AI analytics. So, the proposed model with HAI consideration focuses on identifying the best application strategy that matches the analysis results. This function works to avoid those artificial biases on learning behavior caused by inappropriate application strategies.

6.3.2.5 Conclusions

The HAILA model is significantly effective in terms of analysis result application, and it highlights an implementable way to identify and apply the optimal applying strategy, especially concerning learning strategies. To verify the effectiveness of the modified model and application strategy, two independent experiments were conducted. The results of the first experiment show that a learning strategy that best matched the analysis results were found through the application step in the model. In addition, the findings of comparative experiments showed that students who applied the learning strategy achieved better learning results. This result is consistent with the study by Li et al. (2014) that the SQ3R contributes to good learning achievement. However, it is unclear whether SQ3R brings an additional cognitive load. Moreover, a t-test result showed that the experimental group students who applied the learning strategy were not burdened with an additional cognitive load, in comparison with the control group students.

In contrast with the current analysis result application approaches, the HAILA model consists of four steps, in which there are two design focuses. In particular, evidence-driven education is used to determine the optimal learning strategy, and the cognitive load test provides feedback on the application of the learning strategy. According to the results of the experiment, it was concluded that the SQ3R can improve academic performance by motivating the occurrence of repeated learning. Moreover, statistical results show that the SQ3R helps increase the frequency of Prev learning behavior. In terms of the cognitive load test, there was no significant difference between the students who used the SQ3R and those who did not.

One of the purposes of HAI is to accurately identify students at risk by AI algorithms and provide a timely intervention with considering human education welfare. Some students are inevitably at risk of low academic performance; therefore, how to intervene promptly is a crucial problem. Based on previous studies, it was obvious that most research emphasizes identifying students who are at risk in academic performance by analyzing learning behavior, rather than the application of learning strategies to overcome these risks. The main contribution of this study is that it succeeded in not only enhancing the trustworthiness of AI algorithms analysis results and verifying which factors contribute most to learning performance, but also in determining the optimal learning strategy for intervention in learning behavior to guarantee the effectiveness of AI algorithms analysis application.

7 Conclusions

This paper focuses on causal relationship analysis in Education Big Data mining and addresses the following three issues. 1) It is difficult to identify the cause of learning behavior, 2) It is difficult to identify reading patterns from a single behavior, and 3) It is a technique for extracting causal relationships in large-scale data. In order to solve these problems, we developed a method for analyzing causal relationships, a method for identifying learning behaviors using multidimensional scales, and a method for analyzing causal relationships using deep learning analysis technology.

First, regarding the analysis of causal relationships, correlation analysis is the mainstream in conventional Education Big Data research, and the focus is not on confirming causal relationships in behavior, so the results of correlation analysis cannot be applied to actual classes. The problem is that it is difficult to do so. Therefore, in this study, we proposed incorporating a causal relationship confirmation step into the existing analysis flow. It adds a causal analysis step to the normal flow of data collection, data processing, and data analysis, and also proposes a clear implementation method for causal relationships. Using the proposed method, we actually collected data using e-books and conducted a causal analysis, which revealed that students "get lost in important parts" when reading academic papers. Based on this result, the strategy of "marking important points in advance" was introduced into actual classes, and the results of the experiment confirmed the effectiveness of the strategy.

Next, in research on reading pattern identification, we addressed the problem of low reading pattern identification rates based on a single action. In this study, we considered reading behavior and related behaviors, and proposed a method for identifying reading patterns using a multidimensional scale. There are two types of reading: "shallow reading" and "deep reading." This research focuses on the identification of shallow reading behavior. Existing methods identify shallow reading based only on browsing speed. This study developed a method to identify shallow reading activities by considering factors such as browsing speed, Jump Reading, and Keyword Reading. The results of an experiment in which the proposed method was applied to the analysis of "shallow reading" showed that the proposed method has a higher identification rate than existing methods.

Finally, we are working on the problem of extracting causal relationships in big data. There is a problem that existing causal relationship analysis methods are not suitable for big data. For example, taking Bayesian networks as an example, only discrete dimensions cannot be applied. Many Education Big Data cannot meet these conditions.

In this study, we extract behavioral features from Education Big Data using deep learning technology, such as multilayer neural networks, and propose a causal relationship analysis method based on deep learning. In this proposal, we express the formation process of causal relationships using a Markov chain mathematical model, and design a causal relationship analysis method using the mechanism of generators and discriminators in generative adversarial networks. The proposed method was applied to the analysis of shallow reading behavior. We analyzed the causal relationship between reading behaviors such as reading speed, skipping, and keyword reading, which are related to shallow reading behavior. Reading speed was found to be the root cause of shallow reading behavior.

Acknowledgement

I would like to express my sincere gratitude to everyone who has contributed to my academic journey over the past four years. First and foremost, I extend my heartfelt thanks to Kobe University for providing me with an excellent learning opportunity.

I am deeply grateful to my supervisors, Professor Chengjiu Yin and Kumamoto Etsuko, for their guidance, support, and invaluable insights throughout my research. Special thanks to my lab members, Luo Danqing, Wang Siqi, 白木亮也, Wang Wenhao, and Wu Ziyi, for their collaborative efforts and assistance.

I would also like to acknowledge the significant contributions of Juan Zhou and Ling Xu, whose support has been instrumental in my academic endeavors. Studying in Japan has been an interesting and enriching experience, and I am thankful for the friendships I have formed, particularly with Professor Zhe Li, who has provided me with valuable assistance.

Last but not least, I express my deepest gratitude to my wife, Yu Bai, whose combination of wisdom and beauty has been a constant inspiration. I am also grateful for the unwavering support of my adorable children, Haku and Haru.

Thank you all for being an integral part of my academic and personal growth.

References

Adler, M. J., & Van Doren, C. (2014). How to read a book: The classic guide to intelligent reading. Simon and Schuster.

Alrizq, M., Mehmood, S., Mahoto, N. A., Alqahtani, A., Hamdi, M., Alghamdi, A., & Shaikh, A. (2021). Analysis of Skim Reading on Desktop versus Mobile Screen. Applied Sciences, 11(16), 7398.

Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings?. The Quarterly Journal of Economics, 106(4), 979-1014.

Angrist, J. D., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. Journal of Economic perspectives, 15(4), 69-85.

Bakeman, R., & Gottman, J. M. (1997). Observing interaction: An Introduction to sequential analysis. Cambridge, UK: Cambridge university press.

Baker, B. M. (2007). A Conceptual model for making knowledge actionable through capital formation (Unpublished doctoral dissertation). University of Maryland University College.

Baker, L., & Beall, L. (2014). Metacognitive Processes and Reading Comprehension. In E. Israel & G. Duffy (Eds.), Handbook of Research on Reading Comprehension (pp. 373–388). Taylor & Francis.

Baker, R. S. (2019). Challenges for the future of educational data mining: The Baker learning analytics prizes. JEDM Journal of Educational Data Mining, 11(1), 1-17.

Baker, R. S., & Inventado, P. S. (2014). Chapter 4: Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), Learning Analytics: From Research to Practice (pp. 61-75). New York, NY: The Springer Science+Business Media Company.

Baker, R. S., Martin, T., & Rossi, L. M. (2016). Educational data mining and learning analytics. The Wiley handbook of cognition and assessment: Models, methodologies, and applications, 379-396.

Baldwin-Evans, K. (2006). Key steps to implementing a successful blended learning strategy. Industrial and Commercial Training, 38(3), 156–163.

Bareinboim, E., Tian, J., & Pearl, J. (2022). Recovering from selection bias in causal and statistical inference. In Probabilistic and Causal Inference: The Works of Judea Pearl (pp. 433-450).

Barry, M., & Reschly, A. L. (2012). Longitudinal predictors of high school completion. School Psychology Quarterly, 27(2), 74–84.

Bauer, J., & Prenzel, M. (2012). European teacher training reforms. Science, 336(6089), 1642-1643.

Berliner, D. C. (2002). Comment: Educational research: The hardest science of all. Educational researcher, 31(8), 18-20.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates?. The Quarterly journal of economics, 119(1), 249-275.

Biesta, G. J. (2010). Why 'what works' still won't work: From evidence-based education to value-based education. Studies in philosophy and education, 29(5), 491-503.

Birkerts, S. (1994). The Gutenberg elegies: The fate of reading in an electronic age, Faber and Faber, Boston, MA.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5–32.

Brown, G., Pocock, A., Zhao, M. J., & Luján, M. (2012). Conditional likelihood maximisation: a unifying model for information theoretic feature selection. The journal of machine learning research, 13, 27-66.

Bruniges, M. (2005). An evidence-based approach to teaching and learning.

Burks, A. W. (1951). The logic of causal propositions. Mind, 60(239), 363-382.
Cabrera, A. F., Crissman, J. L., Bernal, E. M., Nora, A., Terenzini, P. T., & Pascarella, E. T. (2002). Collaborative learning: Its impact on college students' development and diversity. Journal of College Student Development, 43(1), 20-34.
Campbell, D. T., & Riecken, H. W. (1968). Quasi-experimental design. International encyclopedia of the social sciences, 5(3), 259-263.
Campbell, J. P., DeBlois, P. B., & Oblinger, D. G. (2007). Academic analytics: A New tool for a new era. EDUCAUSE Review, 42(4), 40-57.
Carr, N. (2020). The shallows: What the Internet is doing to our brains. WW Norton & Company.
Carroll, J. B. (1971). The nature of the reading process. In Reading forum (NINDS Monograph No. 11) Washington, DC: US Government Printing Office (Vol. 197, No. 1, pp. 1-10).
Carter, M., & Egliston, B. (2021). What are the risks of Virtual Reality data? Learning Analytics, Algorithmic Bias and a Fantasy of Perfect Data. New Media & Society, 146144482110127.
Chakrabarty, A., Mannan, S., & Cagin, T. (2015). Multiscale modeling for process safety applications. Waltham, MA: The Butterworth-Heinemann Company.
Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2012). A Reference model for learning analytics. International Journal of Technology Enhanced Learning, 4(5), 318-331.
Chen, C. M., & Chen, F. Y. (2014). Enhancing digital reading performance with a collaborative reading annotation system. Computers & Education, 77, 67-81.
Chen, D. W., & Catrambone, R. (2015, September). Paper vs. screen: Effects on reading comprehension, metacognition, and reader behavior. In Proceedings of the human factors and ergonomics society annual meeting (Vol. 59, No. 1, pp. 332-336). Sage CA: Los Angeles, CA: SAGE Publications.
Chen, G., Cheng, W., Chang, T. W., Zheng, X., & Huang, R. (2014). A comparison of reading comprehension across paper, computer screens, and tablets: Does tablet familiarity matter?. Journal of computers in education, 1, 213-225.
Clow, D. (2012). The learning analytics cycle: Closing the loop effectively. In Proceedings of the 2nd international conference on learning analytics and knowledge (pp. 134–138). New York, NY: ACM.
Coiro, J. (2014). Online reading comprehension: Challenges and opportunities. Texto Livre, 7(2), 30-43.
Coiro, J., & Dobler, E. (2007). Reading Comprehension on the Internet: Exploring the Online Comprehension Strategies Used by Sixth-Grade Skilled Readers to Search for and Locate Information on the Internet. Reading Research Quarterly, 42, 214-257.http://dx.doi.org/10.1598/RRQ.42.2.2
Cook, B. G., Tankersley, M., Cook, L., & Landrum, T. J. (2008). Evidence-Based Practices in Special Education: Some Practical Considerations. Intervention in School and Clinic, 44(2), 69–75.
Cook, T. D., Campbell, D. T., & Shadish, W. (2002). Experimental and quasi-experimental designs for generalized causal inference (Vol. 1195). Boston, MA: Houghton Mifflin.
Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). Applied behavior analysis (2nd ed.). Prentice Hall: Upper Saddle River, NJ.
Cutspec, P. A. (2004). Bridging the research-to-practice gap: Evidence-based education. Centerscope: evidence-based approaches to early childhood development, 2(2), 1-8.
Dalgarno, B., & Lee, M. J. (2010). What are the learning affordances of 3-D virtual environments?. British journal of educational technology, 41(1), 10-32.
Davies, P. (1999). What is evidence-based education? British Journal of Educational Studies, 47(2), 108–121.

Dawson, S., & Siemens, G. (2014). Analytics to literacies: The development of a learning analytics model for multiliteracies assessment. International Review of Research in Open and Distributed Learning, 15(4), 284-305.
Delgado, P., Vargas, C., Ackerman, R., and Salmerón, L. (2018). Don't throw away your printed books: a meta-analysis on the effects of reading media on reading comprehension. Educ. Res. Rev. 25, 23–38.
Detrich, R., & Lewis, T. (2013). A decade of evidence-based education: Where are we and where do we need to go?. Journal of Positive Behavior Interventions, 15(4), 214-220.
Dick, W., Carey, L., & Carey, J. O. (2015). The systematic design of instruction (6th ed.). Upper Saddle River, New Jersey: Pearson.
Dietze, S., Siemens, G., Taibi, D., & Drachsler, H. (2016). Datasets for learning analytics. Journal of Learning Analytics, 3(2), 307-311.
Dillenbourg, P., & Traum, D. (2006). Sharing solutions: Persistence and grounding in multimodal collaborative problem solving. The Journal of the Learning Sciences, 15(1), 121-151.
Dron, J., & Anderson, T. (2009). On the design of collective applications. In Proceedings of the 2009 International Conference on Computational Science and Engineering (pp. 368-374). Vancouver, BC: IEEE.
Eichler, M. (2012). Causal inference in time series analysis. Causality: Statistical perspectives and applications, 327-354.
Elias, T. (2011). Learning analytics: Definitions, processes and potential. Retrieved August 25, 2021, from http://learninganalytics.net/LearningAnalyticsDefinitionsProcessesPotential.pdf
Ellis, B., & Wong, W. H. (2008). Learning causal Bayesian network structures from experimental data. Journal of the American Statistical Association, 103(482), 778-789.
Faisal A., Borodin Y., Puzis Y., & Ramakrishnan, I. V. (2012). Why read if you can skim: towards enabling faster screen reading. Proceedings of the International Cross-Disciplinary Conference on Web Accessibility (pp. 1-10). New York, USA: ACM.
Farley-Ripple, E., May, H., Karpyn, A., Tilley, K., & McDonough, K. (2018). Rethinking connections between research and practice in education: A conceptual model. Educational Researcher, 47(4), 235-245.
Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. International Journal of Technology Enhanced Learning, 4(5-6), 304-317.
Fisher, R. A. (1935). The logic of inductive inference. Journal of the royal statistical society, 98(1), 39-82.
Fitzsimmons, G., Jayes, L. T., Weal, M. J., & Drieghe, D. (2020). The impact of skim reading and navigation when reading hyperlinks on the web. PloS One, 15(9), e0239134. DOI: 10.1371/journal.pone.0239134
Flippo, R. F., & Caverly, D. C. (2009). Handbook of college reading and study strategy research (2nd ed). Madison Avenue, NY: Taylor & Francis Routledge.
Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. Review of educational research, 74(1), 59-109.
Gao, Y., Shen, L., & Xia, S. T. (2021). DAG-GAN: Causal Structure Learning with Generative Adversarial Nets. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3320-3324). IEEE.
García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: Methods and prospects. Big Data Analytics, 1(1), 9-31.
Gawa, Y. O. (2020). Learn while creating! Advanced deep learning with PyTorch. Mynavi.
Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. Advances in neural information processing systems, 27.1-9.
Grari, V., Lamprier, S., & Detyniecki, M. (2023). Adversarial learning for counterfactual fairness. Machine Learning, 112(3), 741-763.
Greco, S., Słowiński, R., & Szczęch, I. (2016). Measures of rule interestingness in various perspectives of confirmation. Information Sciences, 346(1), 216-235.
Gresham, F. M. (2004). Current Status and Future Directions of School-Based Behavioral Interventions. School Psychology Review, 33(3), 326–343.
Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. The Journal of Machine Learning Research, 13(1), 723-773.
Haberfellner, C., & Fenzl, T. (2017). The utility value of research evidence for educational practice from the perspective of preservice student teachers in Austria-A qualitative exploratory study. Journal for educational research online, 9(2), 69-87.
Hammerton, G., & Munafò, M. R. (2021). Causal inference with observational data: the need for triangulation of evidence. Psychological medicine, 51(4), 563-578.
Hamoud, A. K., Hashim, A. S., & Awadh, W. A. (2018). Predicting student performance in higher education institutions using decision tree analysis. International Journal of Interactive Multimedia and Artificial Intelligence, 5(2), 26.
Han, J., Kamber, M., & Mining, D. (2006). Concepts and techniques. Morgan kaufmann, 340, 94104-3205.
Hanna, M. (2004). Data mining in the e-learning domain. Campus-Wide Information Systems, 21(1), 29–34.
Hargreaves, D. H. (1997). In defence of research for evidence-based teaching: A rejoinder to Martyn Hammersley. British Educational Research Journal, 23(4), 405–419.
Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The Rise of "big data" on cloud computing: Review and open research issues. Information Systems, 47(7), 98-115.
Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.
Hauger, D., Paramythis, A., & Weibelzahl, S. (2011). Using browser interaction data to determine page reading behavior. In User Modeling, Adaption and Personalization: 19th International Conference, UMAP 2011, Girona, Spain, July 11-15, 2011. Proceedings 19 (pp. 147-158). Springer Berlin Heidelberg.
Hayes, A. F. (2017). Introduction to mediation, moderation, and conditional process analysis: A regression-based approach. Guilford publications.
Hendricks, M., Plantz, M. C., & Pritchard, K. J. (2008). Measuring outcomes of United Way–funded programs: Expectations and reality. New Directions for Evaluation, 119(9), 13-35.
Hernán, M., & Robins, J. (2020). Causal inference: What if. Boca Raton: Chapman & Hill/CRC.
Holland, P. W. (1986) Statistics and Causal Inference, Journal of the American Statistical Association, 81:396, 945-960
Hou, H. T. (2012). Exploring the behavioral patterns of learners in an educational massively multiple online role-playing game (MMORPG). Computers & Education, 58(4), 1225-1233.
https://net.educause.edu/ir/library/pdf/PUB6101.pdf
Huber, J. A. (2004). A closer look at SQ3R. Reading Improvement, 41(2), 108–112.
Hwang, G. J., & Chang, H. F. (2011). A formative assessment-based mobile learning approach to improving the learning attitudes and achievements of students. Computers & Education, 56(4), 1023–1031.
Imbens, G. W., & Rubin, D. B. (2015). Causal inference in statistics, social, and biomedical sciences. Cambridge University Press.

Janzing, D., & Schölkopf, B. (2010). Causal inference using the algorithmic Markov condition. IEEE Transactions on Information Theory, 56(10), 5168-5194.
Junco, R., Heiberger, G., & Loken, E. (2011). The effect of Twitter on college student engagement and grades. Journal of computer assisted learning, 27(2), 119-132.
Kalainathan, D., Goudet, O., Guyon, I., Lopez-Paz, D., & Sebag, M. (2022). Structural agnostic modeling: Adversarial learning of causal graphs. The Journal of Machine Learning Research, 23(1), 9831-9892.
Kantardzic, M. (2011). Data mining concepts models methods and algorithms. Hoboken, NJ: Wiley-IEEE Press.
Kennedy, G. E., & Judd, T. S. (2004). Making sense of audit trail data. Australasian Journal of Educational Technology, 20(1), 18-32.
Keohane, R. O. (2009). Counterfactuals and causal inference: Methods and principles for social research. Social Forces, 88(1), 466-467.
Khalil, M., Prinsloo, P., & Slade, S. (2022, March). A comparison of learning analytics models: A systematic review. In LAK22: 12th international learning analytics and knowledge conference (pp. 152-163).
Kim H. Y. (2013). Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. Restorative Dentistry & Endodontics, 38(1), 52–54.
Kocaoglu, M., Snyder, C., Dimakis, A. G., & Vishwanath, S. (2017). Causalgan: Learning causal implicit generative models with adversarial training. arXiv preprint arXiv:1709.02023.1-37.
Kormos, C., & Gifford, R. (2014). The validity of self-report measures of proenvironmental behavior: A meta-analytic review. Journal of Environmental Psychology, 40, 359-371.
Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. Progress in Artificial Intelligence, 5(4), 221–232.
Kucirkova, N. (2019). Reader, come home: the reading brain in a digital world. Journal of Children and Media.
Kuromiya, H., Majumdar, R., & Ogata, H. (2020). Fostering evidence-based education with learning analytics. Educational Technology & Society, 23(4), 14–29.
Kvernbekk, T. (2019). Practitioner tales: possible roles for research evidence in practice. Educational Research and Evaluation, 25(1-2), 25-42.
Lai, C. L., & Hwang, G. J. (2016). A self-regulated flipped classroom approach to improving students' learning performance in a mathematics course. Computers & Education, 100, 126-140.
Lauritzen, S. L., & Jensen, F. (2001). Stable local computation with conditional Gaussian distributions. Statistics and Computing, 11, 191-203.
Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. Journal of economic literature, 48(2), 281-355.
Leshner, A., & Pfaff, D. W. (2011). Quantification of behavior. Proceedings of the National Academy of Sciences, 108(3), 15537-15541.
Li, L. Y., Fan, C. Y., Huang, D. W., & Chen, G. D. (2014). The effects of the e-book system with the reading guidance and the annotation map on the reading performance of college students. Journal of Educational Technology & Society, 17(1), 320–331.
Li, L., Uosaki, N., Ogata, H., Mouri, K., Yin, C. (2018) Analysis of Behavior Sequences of Students by Using Learning Logs of Digital Books. In Proceedings of the 26th International Conference on Computers in Education (pp. 367–376). Manila, Philippines: APSCE.
Liang, T. H., & Huang, Y. M. (2014). An investigation of reading rate patterns and retrieval outcomes of elementary school students with e-books. Journal of Educational Technology & Society, 17(1), 218-230.

Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. R news, 2(3), 18–22.
Liñán, L. C., & Pérez, Á. A. J. (2015). Educational Data Mining and Learning Analytics: differences, similarities, and time evolution. RUSC. Universities and Knowledge Society Journal, 12(3), 98-112.
Liu, Z. (2005). Reading behavior in the digital environment: Changes in reading behavior over the past ten years. Journal of documentation, 61(6), 700-712.
Locher, F. M., & Philipp, M. (2023) Measuring reading behavior in large-scale assessments and surveys. Frontiers in Psychology, 13, 1044290.
Loh, K. K., & Kanai, R. (2016). How Has the Internet Reshaped Human Cognition?. The Neuroscientist : a review journal bringing neurobiology, neurology and psychiatry, 22(5), 506–520.
Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., & Welling, M. (2017). Causal effect inference with deep latent-variable models. Advances in neural information processing systems, 30.
Lovmar, L., Ahlford, A., Jonsson, M., & Syvänen, A. C. (2005). Silhouette scores for assessment of SNP genotype clusters. BMC genomics, 6(1), 1-6.
Lu, O. H., Huang, A. Y., Huang, J. C., Lin, A. J., Ogata, H., & Yang, S. J. (2018). Applying learning analytics for the early prediction of Students' academic performance in blended learning. Educational Technology & Society, 21(2), 220-232.
Lu, Y., Zheng, Q., & Quinn, D. (2023). Introducing Causal Inference Using Bayesian Networks and do-Calculus. Journal of Statistics and Data Science Education, 31(1), 3-17.
Maldonado-Mahauad, J., Pérez-Sanagustín, M., Kizilcec, R. F., Morales, N., & Munoz-Gama, J. (2018). Mining theory-based patterns from Big data: Identifying self-regulated learning strategies in Massive Open Online Courses. Computers in Human Behavior, 80(3), 179-196.
Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). Big data: The next frontier for innovation, competition, and productivity.
Martin, G., & Pear, J. J. (2019). Behavior modification: What it is and how to do it (11th Edition). Routledge. https://www.routledge.com/9780815366546.
McCallum, R. S., Sharp, S., Bell, S. M., & George, T. (2004). Silent versus oral reading comprehension and efficiency. Psychology in the Schools, 41(2), 241–246.
McGeown, S. P., Duncan, L. G., Griffiths, Y. M., & Stothard, S. E. (2015). Exploring the relationship between adolescent's reading skills, reading motivation and reading habits. Reading and Writing, 28, 545-569.
McKnight, L., & Morgan, A. (2020). A broken paradigm? What education needs to learn from evidence-based medicine. Journal of Education Policy, 35(5), 648-664.
McLean, C. A. (2019). The shallows? The nature and properties of digital/screen reading. The Reading Teacher, 73(4), 535–542.
McSpadden, K. (2015). You now have a shorter attention span than a goldfish. Time. https://time.com/3858309/attention-spans-goldfish/
Mesarić, J., & Šebalj, D. (2016). Decision trees for predicting the academic success of students. Croatian Operational Research Review, 7(2), 367–388.
Microsoft. (n.d.). How do I change mouse sensitivity (DPI)? - Microsoft Support. https://support.microsoft.com/en-us/topic/how-do-i-change-mouse-sensitivity-dpi-11c0e36c-e348-526b-fdde-80c5d41f606f
Miller, P., Kargin, T., & Guldenoglu, B. (2014). Differences in the reading of shallow and deep orthography: Developmental evidence from Hebrew and Turkish readers. Journal of Research in Reading, 37, 409-432.
Misanchuk, E. R., & Schwier, R. A. (1992). Representing interactive multimedia and hypermedia audit trails. Journal of Educational Multimedia and Hypermedia, 1(3), 35-72.

Moran, D. J., & Malott, R. W. (2004). Evidence-based educational methods. Elsevier Academic Press.
Morgan, S. L., & Winship, C. (2015). Counterfactuals and causal inference. Cambridge University Press.
Morineau, T., Blanche, C., Tobin, L., & Guéguen, N. (2005). The emergence of the contextual role of the e-book in cognitive processes through an ecological and functional analysis. International Journal of Human-Computer Studies, 62, 329–348.
Morrison, K. (2012). Causation in educational research. Routledge.
Muller, C., Winship, C., & Morgan, S. L. (2014). Instrumental variables regression. The SAGE Handbook of Regression Analysis and Causal Inference. London: SAGE, 277-300.
Murnane, R. J., & Willett, J. B. (2010). Methods matter: Improving causal inference in educational and social science research. Oxford University Press.
Myers, M., & Paris, S. (1978). Children's metacognitive knowledge about reading. Journal of Educational Psychology, 70, 680–690.
National Center for Education Statistics. (1991). SEDCAR (Standards for Education Data Collection and Reporting). Rockville, MD: Westat, Inc.
Nauta, M., Bucur, D., & Seifert, C. (2019). Causal discovery with attention-based convolutional neural networks. Machine Learning and Knowledge Extraction, 1(1), 312-340.
Neyshabur, B., Bhojanapalli, S., McAllester, D., & Srebro, N. (2017). Exploring generalization in deep learning. Neural Information Processing Systems, 30, 5947–5956.
Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. Educational Psychologist, 38(1), 63–71.
Park, H., & Kwon, H. C. (2011). Improved Gini-Index Algorithm to Correct Feature-Selection Bias in Text Classification. IEICE Transactions on Information and Systems, E94-D(4), 855–865.
Pearl, J. (1995). From Bayesian networks to causal networks. In Mathematical models for handling partial knowledge in artificial intelligence (pp. 157-182). Boston, MA: Springer US.
Pearl, J. (2000). Models, reasoning and inference. Cambridge, UK: CambridgeUniversityPress, 19(2), 3.
Pearl, J. (2009). Causal inference in statistics: An overview. Statistics Surveys. 3, 96-146
Pearson, K. (1920). Notes on the history of correlation. Biometrika, 13(1), 25-45.
Pellet, J. P., & Elisseeff, A. (2008). Using markov blankets for causal structure learning. Journal of Machine Learning Research, 9(7). 1295-1342
Peters, J., Janzing, D., & Schölkopf, B. (2017). Elements of causal inference: foundations and learning algorithms (p. 288). The MIT Press.
Phillips, R. A. (2006). Tools used in Learning Management Systems: Analysis of WebCT usage logs. In L. Markauskaite, P. Goodyear & P. Reimann (Eds.), Proceedings of the 23rd Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education: Who's Learning? Whose Technology? (pp. 663-673). Sydney, Australia: Sydney University
Phillips, R. A., Baudains, C., & Van Keulen, M. (2002, December). An Evaluation of student learning in a web-supported
Ping, L., & Cunningham, D. (2013). Application of structural equation modeling in educational research and practice (Vol. 7). M. S. Khine (Ed.). Rotterdam: SensePublishers.
Pintrich, P. R., & Zusho, A. (2002). The development of academic self-regulation: The role of cognitive and motivational factors. In A. Wigfield & J. S. Eccles (Eds.), Development of achievement motivation (pp. 249–284). Academic Press.
Porter-O'Donnell, C. (2004). Beyond the yellow highlighter: Teaching annotation skills to improve reading comprehension. English Journal, 93(5), 82-89.

Qiao, S., Wang, H., Liu, C., Shen, W., & Yuille, A. (2019). Micro-batch training with batch-channel normalization and weight standardization. Journal of Latex Class Files, 14(8), 1-15
Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434. 1-16
Raileanu, L. E., & Stoffel, K. (2004). Theoretical Comparison between the Gini Index and Information Gain Criteria. Annals of Mathematics and Artificial Intelligence, 41(1), 77–93.
Richmond, R. C., Al-Amin, A., Smith, G. D., & Relton, C. L. (2014). Approaches for drawing causal inferences from epidemiological birth cohorts: a review. Early human development, 90(11), 769-780.
Riley-Tillman, T. C., Chafouleas, S. M., Christ, T., Briesch, A. M., & LeBel, T. J. (2009). The impact of item wording and behavioral specificity on the accuracy of direct behavior ratings (DBRs). School Psychology Quarterly, 24(1), 1.
Romero, C., & Ventura, S. (2010). Educational data mining: A Review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), 601-618.
Romero, C., & Ventura, S. (2013). Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12-27.
Rose, E. (2010). Continuous partial attention: Reconsidering the role of online learning in the age of interruption. Educational Technology, 50(4), 41-46.
Rosenbaum, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. Statistical Science, 17(3), 286-327.
Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55.
Rubin, D. B. (2008). For objective causal inference, design trumps analysis. The Annals of Applied Statistics, 2(3), 808-840.
Saks, K., Pedaste, M., & Rannastu, M. (2018, July 9–13). University teachers' and students' expectations on learning analytics. Proceedings of 18th International Conference on Advanced Learning Technologies (ICALT) (pp. 183–187). IEEE.
Sarris, M. (2022). Learning to read in a shallow orthography: the effect of letter knowledge acquisition. International Journal of Early Years Education, 30(4), 661-678.
Savage N. (2023) Why artificial intelligence needs to understand consequences. Nature.
Scheines, R. (1997). An introduction to causal inference. V.R. McKim, S.P. Turner (Eds.), Causality in crisis?, University of Notre Dame Press, IN, pp. 185-199.
Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, 85-117.
Schommer-Aikins, M., & Easter, M. (2018). Cognitive flexibility, procrastination, and need for closure linked to online self-directed learning among students taking online courses. Journal of Business and Educational Leadership, 8(1), 112-123.
Shalit, U., Johansson, F. D., & Sontag, D. (2017, July). Estimating individual treatment effect: generalization bounds and algorithms. In International conference on machine learning (pp. 3076-3085). PMLR.
Shapiro, A. M., & Waters, D. L. (2005). An investigation of the cognitive processes underlying the keyword method of foreign vocabulary learning. Language Teaching Research, 9(2), 129-146.
Shimizu, S. (2014). LiNGAM: Non-Gaussian methods for estimating causal structures. Behaviormetrika, 41, 65-98.
Shuell, T. J. (1986). Cognitive conceptions of learning. Review of Educational Research, 56(4), 411-436.
Siemens, G. (2013). Learning analytics: The emergence of a discipline. American Behavioral Scientist, 57(10), 1380-1400.

Siemens, G., & Baker, R. S. D. (2012, April). Learning analytics and educational data mining: towards communication and collaboration. In Proceedings of the 2nd international conference on learning analytics and knowledge (pp. 252-254).
Simamora, R. M. (2020). The Challenges of online learning during the COVID-19 pandemic: An essay analysis of performing arts education students. Studies in Learning and Teaching, 1(2), 86-103.
Simon, H. A. (1952). On the definition of the causal relation. The Journal of Philosophy, 49(16), 517-528.
Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. Educational researcher, 31(7), 15-21.
Slavin, R. E. (2008). Perspectives on evidence-based research in education—What works? Issues in synthesizing educational program evaluations. Educational Researcher, 37(1), 5-14.
Slavin, R. E., Lake, C., Davis, S., & Madden, N. A. (2011). Effective programs for struggling readers: A best-evidence synthesis. Educational Research Review, 6(1), 1-26.
Spirtes, P. & Zhang, K. (2016) Causal discovery and inference: concepts and recent methodological advances. Applied Informatics, 3(3), 1-28.
Spirtes, P. (2001, January). An anytime algorithm for causal inference. In International Workshop on Artificial Intelligence and Statistics (pp. 278-285). PMLR.
Spirtes, P., Glymour, C. N., & Scheines, R. (2000). Causation, prediction, and search. MIT press.
Stroud, B. (1978). Hume and the idea of causal necessity. Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 33(1), 39-59.
Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. Statistical science: a review journal of the Institute of Mathematical Statistics, 25(1), 1.
Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257–285.
Tabri, N., & Elliott, C. M. (2012). Principles and practice of structural equation modeling. Canadian Graduate Journal of Sociology and Criminology, 1(1), 59-60.
Ten Have, T. R., & Joffe, M. M. (2012). A review of causal estimation of effects in mediation analyses. Statistical Methods in Medical Research, 21(1), 77-107.
US Department of Education (Ed.). (2009). Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies. https://www2.ed.gov/rschstat/eval/tech/evidence-based-practices/finalreport.pdf
Van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Van Aken, M. A. (2014). A gentle introduction to Bayesian analysis: Applications to developmental research. Child development, 85(3), 842-860.
Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. Computers in Human Behavior, 89, 98–110.
Walsh, G. (2016). Screen and paper reading research–a literature review. Australian Academic & Research Libraries, 47(3), 160-173.
Wang, Q., Woo, H. L., Quek, C. L., Yang, Y., & Liu, M. (2012). Using the Facebook group as a learning management system: An exploratory study. British journal of educational technology, 43(3), 428-438.
West, D., Tasir, Z., Luzeckyj, A., Si Na, K., Toohey, D., Abdullah, Z., Searle, B., Farhana Jumaat, N., & Price, R. (2018). Learning analytics experience among academics in Australia and Malaysia: A comparison. Australasian Journal of Educational Technology, 34(3).
Winne, P. H. (2010). Improving measurement of self-regulated learning. Educational Psychologist, 45(4), 267-276.

Wong, B. T. M. (2017). Learning analytics in higher education: An Analysis of case studies. Asian Association of Open Universities Journal, 12(1), 21-40.
Wooldridge, J. M. (2019). Correlated random effects models with unbalanced panels. Journal of Econometrics, 211(1), 137-150.
Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2013). Data mining with big data. IEEE transactions on knowledge and data engineering, 26(1), 97-107.
Xu, Q., Huang, G., Yuan, Y., Guo, C., Sun, Y., F. Wu, & Weinberger, K., (2018). An empirical study on evaluation metrics of generative adversarial networks, arXiv:1806.07755, 1-14
Yang, C. C., Chen, I. Y., & Ogata, H. (2021). Toward precision education. Educational Technology & Society, 24(1), 152-163.
Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., & Zhang, A. (2021). A survey on causal inference. ACM Transactions on Knowledge Discovery from Data (TKDD), 15(5), 1-46.
Yin, C., & Hwang, G. J. (2018). Roles and strategies of learning analytics in the e-publication era. Knowledge Management & E-Learning: An International Journal, 10(4), 455-468.
Yin, C., Yamada, M., Oi, M., Shimada, A., Okubo, F., Kojima, K., & Ogata, H. (2019). Exploring the relationships between reading behavior patterns and learning outcomes based on log data from e-books: A Human factor approach. International Journal of Human–Computer Interaction, 35(4), 313-322.
Yu, K., Liu, L., & Li, J. (2021). A unified view of causal and non-causal feature selection. ACM Transactions on Knowledge Discovery from Data (TKDD), 15(4), 1-46.
Zhao, F., & Yin, C. (2021). Data collection in the learning behavior analysis: A systematic review. In 2021 10th International Congress on Advanced Applied Informatics (IIAI-AAI) (pp. 178-181). IEEE.
Zhao, F., & Yin, C. (2022). A Technique for Tracking the Reading Rate to Provide Students' Learning Feedback. In S. Iyer, J.-L. Shih, W. Chen, & M. N. M. Khambari (Eds.), The 30th International Conference on Computers in Education (ICCE 2022) 1, 381–386. APSCE.
Zhao, F., Hwang, G. J., & Yin, C. (2021). A result confirmation-based learning behavior analysis model for exploring the hidden reasons behind patterns and strategies. Educational Technology & Society, 24(1), 138-151.
Zhao, F., Liu, G. Z., Zhou, J., & Yin, C. (2023). A learning analytics model based on human-centered artificial intelligence for identifying the optimal learning strategy to intervene in learning behavior. Educational Technology & Society, 26(1), 132-146.
Zouaghi, R. (2016). A research study about: Computer-Assisted Reading instruction. International Journal of Instructional Technology and Distance Learning, 13(8), 61–65.

Doctor Thesis, Kobe University
"Study on Exploring Causal Relationships in Educational Big Data Mining", 89 pages
Submitted on January, 15, 2024
When published on the Kobe University institutional repository /Kernel/, the publication date shall appear on the cover of the repository version.