Interleaved Design for E-learning: Theory, Design, and Empirical Findings

Li, Tao, Sean Xu, De Liu, and Yufang Wang. Accepted at MIS Quarterly, December 2023.

Abstract

The rapid development of e-learning has drawn increasing attention to the issue of how learners' learning activities can be better structured using technologies. This study focuses on how to improve e-learning performance by optimizing the structuring of learning sessions from the perspective of interleaving (i.e., mixing different topics in a learning session). Following the design science paradigm, this study chooses cognitive load theory as the kernel theory and proposes a new interleaving design — *related-interleaving* — that populates an interleaved session with related topics as a way of reducing cognitive load during an interleaved session. Drawing on the theoretical predictions, we design and instantiate a personalized learning system with the related-interleaving strategy by fusing educational strategies and machine learning techniques. The results from a two-month field experiment confirm that related-interleaving outperforms non-interleaving and unrelated-interleaving. Our findings also reveal that compared with unrelated-interleaving, related-interleaving benefits weak learners more and thus helps reduce learning performance disparities. This study demonstrates how personalized e-learning systems can be further improved from the perspective of interleaving.

Keywords: E-learning, interleaving, topic relatedness, machine learning, cognitive load theory, weak learner

INTRODUCTION

The e-learning industry has grown rapidly in recent years with over 60% of postsecondary degree seekers in the U.S. engaged in some form of e-learning.[1] Compared with traditional classroom-based learning, e-learning allows learners to access course materials at any time and from any location with an internet connection. Moreover, e-learning platforms can better personalize learning activities to suit each learner's progress and style (Chen et al. 2018; Park and Lee 2003). Despite these advantages, e-learning still faces criticism for its limited learning effectiveness (Bettinger et al. 2017; Figlio et al. 2013; Goudeau et al. 2021). Compared to learners in traditional classrooms, e-learners adopt a more passive mode of learning: they mostly consume and record information rather than actively reflecting upon it based on existing knowledge (Furenes et al. 2021; Shrivastav and Hiltz 2013). This can lead to a limited depth of understanding and a low ability to transfer the knowledge learned from one context to another (Delgado and Salmerón 2021).

One strategy to promote active thinking and learning effectiveness, as advocated by educational researchers, is to mix practices of different topics in the same learning session — called *interleaving* (Firth et al. 2021). In interleaved learning, learners are exposed to different topics in one session and learning of the same topic is spread across multiple sessions (Rohrer et al. 2020). For example, when learning Python data structures, such as matrix, tuple, and dictionary, an interleaved design would involve a series of sessions, each mixing exercises for different data structures instead of each focusing on one data structure. Advocates of interleaved learning suggest that it encourages learners to actively identify different topics and corresponding strategies based on their existing knowledge because they cannot simply rely on repetitive practices to solve the same type of problem over and over

[1] https://www.census.gov/library/stories/2020/08/schooling-during-the-covid-19-pandemic.html

again (Jaeger et al. 2016). This process can improve learners' ability to identify boundaries and connections among different topics, leading to a deeper understanding of the subject (Mielicki and Wiley 2022; Rohrer 2012; Rohrer et al. 2014). To our knowledge, however, e-learning platforms have not embraced interleaving. The prevailing design is still non-interleaving; that is, offering multiple practices for one topic in a session before moving on to the next topic (Hussain et al. 2019; Loghin et al. 2008). Given the potential benefits of interleaving, research is needed on how to leverage it in e-learning settings to improve e-learning effectiveness.

Existing interleaving designs, which are designed for traditional face-to-face instructions, may not work well for e-learning settings for a few reasons. First, past findings show that the effects of interleaving are not always positive (Firth et al. 2021; Rohrer et al. 2020), in part because students find interleaving more difficult than non-interleaving (Rohrer et al. 2015; Tauber et al. 2013; Yan et al. 2016). This could pose a special challenge for e-learning because learners on e-learning platforms may be more susceptible to distractions and cognitive overload (Delgado and Salmerón 2021). In particular, online learners are often learning in environments that are not specifically designed for focused learning (Conrad et al. 2022) and online platforms themselves can also be distracting, with various information competing for attention (Dontre 2021; Shrivastav and Hiltz 2013; Wang 2022). Hence, when designing interleaving for e-learning, it is important to consider how interleaving will affect the cognitive load on learners. To do so, one needs a better theoretical understanding of the relationship between interleaving and learners' cognitive resources and to design interleaving accordingly to alleviate the concern of further overloading e-learners.

Moreover, existing interleaving designs require teachers to pick topics for each interleaved session and the same design is offered to all learners. Such traditional designs do not take advantage of the rich data on learners' past activities and performance, while on e-

learning platforms, the easily accessible data can be leveraged to offer personalized and adaptive learning sessions for each individual learner. Therefore, for interleaving to be most effective, the traditional interleaving design must be modernized to suit the highly dynamic e-learning settings.

To address the aforementioned gaps, this research offers a theory-driven interleaving design for e-learning settings that is personalized, adaptive, and cognizant of each learner's cognitive load. To achieve this goal, we first draw on the cognitive load theory (CLT) to develop an understanding of the relationship between interleaved learning and learners' cognitive load. CLT is a fundamental theory of learning that focuses on the cognitive demands of learning (Sweller 2011). Based on CLT, we propose that while interleaved learning prompts learners to make connections between different topics and expand learning opportunities, it also increases learners' cognitive load compared to non-interleaving, which may reduce learning effectiveness. Accordingly, we propose a new interleaving design — *related-interleaving* — that requires an interleaved learning session to consist of related topics so that it reduces the cognitive resources required for basic processing while still offering opportunities for making connections between different topics. Based on this theoretical perspective, we also anticipate that weaker learners, who have less working memory for encoding new knowledge, are more likely to benefit from this related-interleaving design.

We then follow the design science guidelines to implement an interleaving design for e-learning that is data-driven, personalized, and adaptive. Our design framework includes the following components: (a) dynamic detection of learners' weak topics based on their past performance using a hidden Markov model (Reddy et al. 2016; Wilson et al. 2016), (b) a knowledge map for capturing relatedness between different topics, which is dynamically updated using fuzzy association rules (Tseng et al. 2007), and (c) a scheduling engine that

assembles practice materials in an adaptive, personalized manner. The scheduling engine ensures that the topics covered in each session are suitable for the learner's progress (based on the detected weak topics) and are related (based on the knowledge map).

To evaluate our design, we compare our *related-interleaving* design with *non-interleaving* (where each learning session focuses on a single topic) and *unrelated-interleaving* (where interleaved topics are chosen without considering topic relatedness) in a randomized field experiment involving 510 middle school students using an e-learning platform designed by this research team. Our results show that, in the context of e-learning, related-interleaving leads to better learning performance than non-interleaving and unrelated-interleaving. Furthermore, the benefit of related-interleaving over unrelated-interleaving is more prominent for weak learners than for strong learners. This suggests that our related-interleaving design for e-learning not only improves e-learning performance overall, but also reduces the performance disparities between weak and strong learners.

LITERATURE REVIEW

Research on E-learning Design

The low effectiveness of e-learning has been a major concern for educators and researchers (Bettinger et al. 2017; Figlio et al. 2013; Goudeau et al. 2021). To ensure the overall effectiveness of e-learning initiatives, there is an urgent need to develop targeted strategies and interventions that cater to the unique needs of online learners (Hansen and Reich 2015; Kizilcec et al. 2017; Reich and Ruipérez-Valiente 2019). In response to this challenge, Information Systems scholars have approached technology-based e-learning designs from multiple perspectives (Alavi and Leidner 2001; Gupta and Bostrom 2013; Gupta and Bostrom 2009; Piccoli et al. 2001). One stream of research focuses on technology-enabled behavioral nudges to engage online learners (Damgaard and Nielsen 2018) and facilitate self-regulation of the learning pace (Santhanam et al. 2008). Examples of these

nudges include on-the-hour cues (Huang et al. 2023), call-to-actions (Huang et al. 2021), and gamified interventions (Leung et al. 2023). A second stream studies the effects of communication or collaboration support tools that facilitate online learners' interaction with each other (Kulkarni et al. 2015) and with instructors (Dennen et al. 2007).

Our study belongs to the third stream, which focuses on the structuring of e-learning activities. For example, researchers have examined how to choose appropriate learning session lengths (Manasrah et al. 2021) and incorporate active learning activities during online lectures (Khan et al. 2017; Sandrone et al. 2021). A more recent focus is to personalize online learners' experience, including detecting weak topics (i.e., identifying gaps in learners' knowledge) for the purpose of recommending appropriate topics to learn next (e.g., Bauman and Tuzhilin 2018; Wilson and Nichols 2015), and adapting the challenge level of learning materials (Kim et al. 2020). Several researchers further explore how to optimize the sequence of learning sessions, with each session dealing with a different weak topic. For example, some studies suggest arranging topics from low to high difficulty levels across sessions (Hussain et al. 2019; Jiang et al. 2022). Others optimize the sequence of learning sessions based on topic similarity and learner preferences (Al-Muhaideb and Menai 2011; Chen 2008; Jeng and Huang 2019; Kurilovas et al. 2015). This research differs from prior studies in that, instead of studying the sequence of topics or learning sessions, we are concerned with how to interleave different topics in one session.

Interleaved Session Design

Interleaved learning, as proposed in the education field, exposes learners to a few different topics in a single learning session (e.g., ABC, BCD) (Taylor and Rohrer 2010). This contrasts with a conventional non-interleaved (or "blocking") design, which exposes learners to a single topic repeatedly in a learning session (e.g., AAA, BBB). Educational researchers propose that interleaving different topics in one session can encourage learners to probe the connections and differences among topics and associate problems with the corresponding

strategies, thus leading to better learning outcomes (Birnbaum et al. 2013; Rohrer et al. 2014; Taylor and Rohrer 2010).

Thus far, most research on interleaved learning has focused on examining *whether* interleaving can outperform non-interleaving. Some studies demonstrate that interleaving is more beneficial in various subject domains, such as math, category induction, sports, and medical training (Foster et al. 2019; Kornell and Bjork 2008; Rohrer et al. 2015). Other studies, however, suggest that interleaved learning may not always outperform non-interleaved learning (Carvalho and Goldstone 2014; Hausman and Kornell 2014). For example, an interleaved design is not as effective as a non-interleaved design when interleaved concepts are highly distinguishable (Carvalho and Goldstone 2014; Zulkiply and Burt 2013). A few recent studies further suggest that learners may struggle to process interleaved information when their memory capacities are limited, which may significantly dilute the benefits of interleaving (Firth et al. 2021; Sana et al. 2018).

The issue of *how* to design interleaving has received scant attention in this literature. One exception is Yan and Sana (2021), which explores interleaving at different *levels*; that is, whether to interleave *within* a domain (e.g., mixing different topics of statistics) or *across* domains (e.g., mixing topics of statistics and physics). Their research, however, does not address the question of how to choose topics from the same domain.

DESIGNING RELATED-INTERLEAVING FOR E-LEARNING

This study focuses on the design of interleaving. We follow the design science research approach (Abbasi and Chen 2008; Walls et al. 1992) to develop a system for intelligent, data-driven interleaved learning for e-learning platforms (see Table 1). In doing so, we select CLT (Sweller 2011) as the kernel theory to motivate our related-interleaving design. CLT is a fundamental theory about the relationship between the cognitive demands of learning and

learning performance. Given the recent suggestions that a learner's memory capacity can be a barrier to realizing the benefits of interleaving (Firth et al. 2021; Sana et al. 2018), we posit that CLT is the appropriate framework for analyzing both the benefits and costs of interleaving. This cost perspective is especially relevant when guiding the design of interleaving sessions for e-learning as e-learners typically have limited cognitive resources available (e.g., Delgado and Salmerón 2021). Guided by CLT, we propose a new related-interleaving design to lessen the cognitive load on learners while still offering opportunities to make connections between different topics. We then identify meta-requirements for the related-interleaving design for e-learning systems and propose meta-designs that satisfy these requirements. We also instantiate an e-learning system that integrates the meta-designs. We present our design framework in Table 1 and discuss the details in the following sections.

A Brief Overview of Cognitive Load Theory

CLT is one of the most widely used theories in learning and instruction (Kalyuga 2007; Sweller 2010). CLT is built on an understanding of the role of memory in learning activities. It recognizes that human beings use working memory for receiving and processing new information and long-term memory for storing and organizing processed information in the form of *schemas* (Sweller 2011). Such cognitive schemas can be quickly retrieved and used in a flexible way to resolve problems; hence, building such schemas is an essential goal of learning. When a learner approaches new information, basic processing occurs first. This includes receiving information and retrieving existing schemas to understand the information. Then, schema building may occur (Sweller et al. 2019). The latter includes, for example, categorizing information, abstracting away unnecessary details, identifying relationships with other information, and integrating with existing schemas.

Table 1: Theory-driven Design Framework for an Interleaved E-learning System

Research goal	Improve the e-learning session design from the interleaving perspective
Kernel theory	Cognitive load theory (CLT) theorizes that the ultimate design goal to achieve effective learning is to manage learners' basic processing load and maximize their schema building. Based on CLT, increasing the relatedness of interleaved topics (i.e., as in related-interleaving) can reduce basic processing load and provide more opportunities for schema building, thus leading to better learning performance.
Meta-requirements	1. Dynamically detect the learner's weak topics. 2. Increase the relatedness of weak topics within the same learning session. 3. Schedule available learning materials according to the criteria of weak topics and topic relatedness.
Meta-design	1. Use the hidden Markov model to dynamically identify the mastery level of each topic for a focal learner. 2. Include expert knowledge and fuzzy association rules to build a knowledge map to detect topic relatedness. 3. Design a scheduling engine that implements the scheduling goals.
Testable hypotheses	**H1**: Compared with unrelated-interleaving, related-interleaving leads to better learning performance. **H2**: Compared with non-interleaved learning, related-interleaving leads to better learning performance.
System instantiation	Instantiate the meta-design artifacts and implement the designed system.
Experimental evaluation	Empirically evaluate the testable hypotheses via a randomized field experiment and post-hoc analyses on the heterogenous effect of related-interleaving.

Crucially, both basic processing and schema building require working memory, which is very limited in capacity and duration (Zhu and Watts 2010). When basic processing consumes too much working memory, schema building is reduced, leading to suboptimal learning. In particular, the amount of working memory used (or the *cognitive load*) for basic processing is a function of the complexity of the materials, the presentation format, and whether the learner can retrieve and apply relevant schemas from long-term memory. Applying existing schemas can drastically lessen the cognitive load needed for basic information processing (Kleider et al. 2008). Overall, a fundamental principle of CLT for

effective learning session design is to manage the cognitive load for basic processing, allowing sufficient resources for schema building.

Applying CLT to instructional design, scholars have focused on methods to reduce the basic processing load, such as presenting information in an easy-to-understand format (e.g., by incorporating multimedia and diagrams) (Brunken et al. 2003; Mayer and Moreno 2003), providing worked examples or partial solutions for practices (Renkl 2014; Sweller et al. 2019) and incorporating multi-modal information (e.g., visual and auditory information) (Ginns 2005). A few other studies have explored approaches to deliberately increase the basic processing load within learners' working memory capacity to provide more opportunities for schema building. For instance, it is suggested that providing practices with higher variability for the same topic leads to better learning outcomes when the total cognitive load remains within limits (Likourezos et al. 2019).

The above studies primarily focus on designing practices or learning materials for a specific learning topic, whereas this study focuses on how to interleave different topics in a learning session. Therefore, this research addresses a gap in the literature of CLT-based instructional design and adds new perspectives on how interleaving designs may affect cognitive load and learning performance.

CLT and Interleaving

In non-interleaved learning, learners encounter materials of the same topic repeatedly in a learning session. Observing the commonality between materials of the same topic can facilitate building schemas of this topic, which, in turn, can be used to quickly process other materials of the same topic. Non-interleaved learning thus drastically reduces learners' cognitive load for basic processing. However, because learners can apply the same schemas in the entire session, opportunities for refining or building new schemas are also limited.

In interleaved learning, learners have opportunities to process materials for different topics in the same session, which may lead to the construction of higher-level schemas that can help learners "connect the dots" among different topics (Rohrer et al. 2014). Furthermore, learners may also contrast between different topics and are hence exposed to the limitation of schemas built for specific topics, which may lead to more refined and robust schemas (Birnbaum et al. 2013; Rohrer 2012; Rohrer et al. 2015). Thus, interleaved learning can expand "**schema-building opportunities**." However, interleaved learning may also increase learners' cognitive load and elevate the "**overload risk.**" Given that learners cannot easily leverage schemas developed for one topic to the next topic, the cognitive load for basic processing in interleaved learning can be substantially higher than that in non-interleaved learning. As learners devote more cognitive resources to basic processing, they may not have enough cognitive capacity for schema building (Sana et al. 2018). This is highly relevant in e-learning where learners tend to operate with reduced cognitive resources, but this downside has not been adequately recognized in the literature.

To mitigate the overload risk of interleaved learning and promote its schema-building benefits, we propose an interleaving design called *related-interleaving*, which involves purposefully choosing highly related topics for inclusion in an interleaved learning session. [2] When the topic relatedness is relatively high, schemas built for one topic can be partially reused for a related topic, lessening the cognitive load for basic processing and thus

[2] Topic relatedness in this study differs from concepts raised in prior interleaving studies such as topic similarity (Foster et al. 2019). Specifically, topic similarity refers to whether the topics belong to the same category, while topic relatedness in this study emphasizes the dependency relationship: whether understanding topic A can enhance the learning of a subsequent topic B. To illustrate this distinction, let us consider the examples from Foster et al. (2019). In their study, computing volumes of spheroids and wedges are similar topics. However, in our study, these topics are not related because knowing how to compute volumes of spheroids does not depend on the knowledge of computing volumes of wedges, and vice versa; instead, computing volumes of spheroids (i.e., the area of the circle times 2/3 of the height) should be related to computing the area of circles, because computing volumes of spheroids depends on correctly calculating the area of circles.

mitigating the "**overload risk**" (O'donnell et al. 2002; Sweller 2010). For example, the schemas built for Python array can facilitate the understanding of matrix. Hence, these two topics can be included in the same interleaved session. Related-interleaving stands in contrast to traditional interleaving designs that do not consider topic relatedness and thus have low or no topic relatedness ("*unrelated-interleaving*" hereafter).

Furthermore, high topic-relatedness can also provide more "**schema-building opportunities**." When topics in a learning session are highly related, more "dots" can be connected, which can facilitate the construction of high-level schemas. Furthermore, when topics are highly related, there is also a greater need to compare and contrast them, which can help learners fix misconceptions about a particular topic and develop a more robust and nuanced understanding of different topics (Birnbaum et al. 2013; Carvalho and Goldstone 2014). For instance, when Python learners resolve a question on array next to one on matrix, they are likely prompted to deliberate on the connections and distinctions between the two topics and may thus form a deeper understanding of both topics.

Meta-Requirements for Related-interleaving

Guided by the kernel theory, we now discuss the meta-requirements for our proposed design artifact. As stated in the introduction, our design goals include improving e-learning session design using related-interleaving and making learning session design more personalized and adaptive by leveraging the rich data generated in e-learning. To fulfill these goals, we propose a related-interleaving design with the following key components: weak topic detection, topic-relatedness modeling, and a scheduling engine. The weak-topic-detection component draws upon a learner's past learning record to produce a personalized list of weak topics at a specific moment. The topic-relatedness-modeling component builds and updates a knowledge map that models relatedness among topics. The scheduling engine chooses from available learning materials that address a learner's weak topics while meeting

the criteria of related-interleaving. Next, we discuss the three components separately in further detail and how they work together as a system.

Meta-Design I: Weak Topic Detection Using Hidden Markov Model

According to CLT, learning occurs when novel information prompts learners to build new schemas or to enhance existing ones (Sweller 2011). By detecting each learner's weak topics and letting her focus on unmastered topics, the session design uses the learner's time efficiently. For this purpose, our e-learning system maintains a list of weak topics for each learner at any time. Weak topics are topics not yet mastered by a learner, as indicated by the learner's poor performance on the topic (Bauman and Tuzhilin 2018). Previous work has explored several approaches to detect learners' weak topics. One approach involves comparing a learner's overall performance in practicing a specific topic with a predefined threshold to identify weak topics (Bauman and Tuzhilin 2018). Another approach uses user-based collaborative filtering to suggest unmastered topics to learners with similar learning conditions (Klašnja-Milićević et al. 2011). More recently, Bayesian knowledge tracing models have been introduced to consider learners' performance at the individual practice level and to model the hidden transition of a learner's states in terms of topic mastery (Pelánek 2017). This approach acknowledges that a learner can gain mastery of a topic over time through practice but may also lose mastery as time passes. It has become the prevailing approach on personalized learning platforms (Abdelrahman et al. 2023).

Our model builds on previous work on Bayesian knowledge tracing that captures temporal changes in students' topic mastery (Pardos et al. 2013; Reddy et al. 2016). The evolution of topic mastery can be represented by the transition process in a hidden Markov model (HMM) and the observed performance on a topic can be captured by the emission process of the HMM (Chen et al. 2018). To model the evolution of a learner's topic mastery, we build an HMM, as shown in Figure 1, to consist of (1) a hidden transition process for capturing the evolution of topic mastery and (2) an observed emission process for capturing

the observed learning performance. Formally, we model a learner *s*'s mastery of the topic *c* at time *t* (defined as the time of the *t*-th practice of the topic) as a latent state (y_{st}^{c}) with N ordered levels.[3] The emission (x_{st}^{c}) is defined as answer correctness, which takes a value of "1" if the learner *s* answers the question on the topic *c* correctly at time *t* and "0" otherwise.

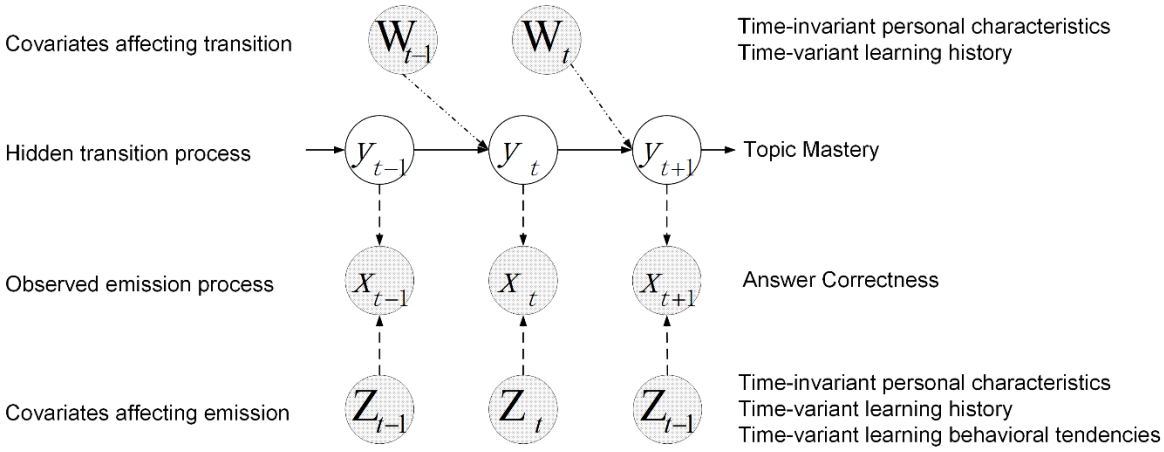


Figure 1. Hidden Markov Model for Learning Processes

Hidden transition process: HMM assumes that the evolution of hidden states over time follows a Markov chain, in which the next state (y_{st+1}^{c}) depends only on the current state (y_{st}^{c}) and the transition covariates ($\boldsymbol{W_{st}^{c}}$). We include a vector of covariates ($\boldsymbol{W_{st}^{c}}$) affecting state transitions, including both time-invariant learner characteristics (e.g., *gender* and *age*) and time-variant learning history (e.g., the number of correct answers on topic *c*) (Kim and Krishnan 2019). We provide more details about the covariates in Appendix A.

Let $p_{st}^{c}(i,j)$ represent the probability of learner *s*'s mastery of topic *c* transiting from state *i* at time t to state *j* at time t+1. $\boldsymbol{P_{st}^{c}} = [p_{st}^{c}(i,j)]$ is an $N \times N$ transition matrix for learner *s* on topic *c*. Following Singh et al. (2011), we assume that the topic mastery state can only transit from one state to its adjacent states and that the transition follows a random walk. Consequently, the transition matrix $\boldsymbol{P_{st}^{c}}$ is illustrated below:

[3] We have explored N = 2, 3, and 4, and the optimal number of hidden states is N = 2 according to the Bayesian information criterion.

$$\boldsymbol{P}_{st}^{c} = \begin{bmatrix} p_{st}^{c}(1,1) & p_{st}^{c}(1,2) & \cdots & 0 & 0 \\ p_{st}^{c}(2,1) & \ddots & & & 0 \\ \vdots & & p_{st}^{c}(\mathrm{i},\mathrm{j}) & & \vdots \\ 0 & & & \ddots & p_{st}^{c}(N-1,N) \\ 0 & 0 & \cdots & p_{st}^{c}(\mathrm{N},\mathrm{N}-1) & p_{st}^{c}(N,N) \end{bmatrix}$$

Following Singh et al. (2011), we assume that the hidden probability follows an ordered logit model. Specifically, we let the probability of a learner's mastery of topic *c* transiting to a lower-ordered state, a higher-ordered state, or the same state, respectively, as follows:

$$p_{st}^{c}(\mathrm{i},\mathrm{i}-1) = \frac{\exp\left(u_i^{lc} - \boldsymbol{\beta_i^{c\prime}}\boldsymbol{W_{st}^{c}} - \delta_{st}^{c}\right)}{1+\exp\left(u_i^{lc} - \boldsymbol{\beta_i^{c\prime}}\boldsymbol{W_{st}^{c}} - \delta_{st}^{c}\right)}$$

$$p_{st}^{c}(\mathrm{i},\mathrm{i}+1) = 1 - \frac{\exp\left(u_i^{hc} - \boldsymbol{\beta_i^{c\prime}}\boldsymbol{W_{st}^{c}} - \delta_{st}^{c}\right)}{1+\exp\left(u_i^{hc} - \boldsymbol{\beta_i^{c\prime}}\boldsymbol{W_{st}^{c}} - \delta_{st}^{c}\right)}$$

$$p_{st}^{c}(\mathrm{i},\mathrm{i}) = 1 - \mathrm{p}(\mathrm{i},\mathrm{i}-1)_{st} - \mathrm{p}(\mathrm{i},\mathrm{i}+1)_{st}$$

where u_i^{lc} and u_i^{hc} ($u_i^{lc} < u_i^{hc}$) are two threshold values that are used to divide the transition probabilities. $\boldsymbol{\beta_i^{c}}$ is the vector of topic-specific and state-dependent parameters for $\boldsymbol{W_{st}^{c}}$. δ_{st}^{c} is the random noise.

State-Dependent Emission Process: The probability of a learner *s* answering a question on topic *c* correctly (x_{st}^{c}) at time t is a function of the topic mastery state y_{st}^{c} and some covariates ($\boldsymbol{Z_{st}^{c}}$). The covariates include time-invariant learner characteristics, time-variant learning history (e.g., the number of correct answers on the focal topic *c*), and time-variant learner behavioral tendencies (e.g., the average time spent on each answer of the focal topic *c* and its standard deviation) (Ayabakan et al. 2016).

Learners with different topic mastery levels (y_{st}^{c}) will have different distributions of correctly answering the topic-related questions (x_{st}^{c}). We thus model x_{st}^{c} as a Gaussian mixture distribution, which is generated by the different discrete hidden states (y_{st}^{c}). Answer correctness x_{st}^{c} depends on the probabilities of the learner being in different mastery states

of topic *c*. Following Ayabakan et al. (2016), we choose the logit model to represent the emission probability as follows:

$$p(x_{st}^{c} = 1_{|y_{st}^{c}=i}) = \frac{1}{1 + e^{-(\gamma^{c}+\alpha_i+\boldsymbol{\theta_i'}\boldsymbol{Z_{st}^{c}}+\varepsilon_{st})}}$$

where γ^c represents the topic-level heterogeneity (e.g., different difficulty levels) and α_i captures the heterogeneity associated with the mastery state *i*. $\boldsymbol{\theta_i}$ is the vector of state-dependent parameters for $\boldsymbol{Z_{st}^{c}}$ and ε_{st} is the random noise.

We then estimate the HMM by maximizing the likelihood of the observed emission sequences. More details about the estimation of the HMM are presented in Appendix A. The outcomes of this process yield a personalized list of weak topics that are specifically tailored to the focal learner's dynamic learning progress.

Meta-Design II: Topic Relatedness Learning Using a Knowledge Map

With weak topics detected, the next question is what topics can be mixed in a learning session. As discussed earlier, CLT suggests that related-interleaving can reduce learners' cognitive workload for basic information processing while still providing knowledge integration opportunities. Hence, our second meta-requirement is to model topic relatedness. Topic relations are typically represented in the form of the knowledge map, which is a graph model where nodes (points/vertices) represent topics and edges (arcs/links) portray the dependency relationships between topics (Atapattu et al. 2017; Balaid et al. 2016; Lee and Segev 2012). An intuitive and common understanding of topic dependency in learning contexts is that topic B depends on topic A if mastering topic A can facilitate learners to master topic B (Tseng et al. 2007). This understanding is also consistent with the principles of CLT, that is, learners can leverage the schema built from learning topic A to understand topic B more efficiently.

Existing e-learning platforms employ various methods to construct knowledge maps. Some platforms depend on domain experts to manually create knowledge maps, which can

often ensure a high level of reliability but requires substantial labor (Wilson and Nichols 2015). Other platforms employ text mining techniques, such as TF-IDF and NLP, to extract knowledge maps from learning materials such as syllabi and reading materials (Bauman and Tuzhilin 2018). In addition, a data-driven approach involves learning knowledge maps based on learners' learning records, enabling the capture of subtle changes in topic relationships over time (Balaid et al. 2016; Tseng et al. 2007).

We choose an approach of combining expert knowledge and data insights by initializing the knowledge map using expert knowledge and then refining it dynamically based on learners' learning records. This hybrid approach is based on the following considerations. On one hand, many teachers have some expert knowledge about topic dependency, which is a valuable source of information, especially when topic dependency data are sparse. On the other hand, building an exhaustive knowledge map is a time-consuming process and is prone to incompleteness. Therefore, the creation of a knowledge map needs to be automated using topic dependencies observed in the data. Specifically, we develop a set of expert rules (M_0) by encoding the knowledge of several senior teachers at the school where we conducted our experiment. Then, we use a data-driven method to dynamically fine-tune the knowledge map. Specifically, every day *d* at midnight, we retrieve all the historical learning records and discover the data-driven rule set (M_d). We then combine the two rule sets to form an updated knowledge map.

We represent topic dependencies learned from the data as fuzzy association rules. The conditional dependence nature of fuzzy association rules corresponds well to the topic dependencies that we intend to capture. Specifically, we infer topic dependencies from learning data if we observe conditional dependence between learners' performances on two

topics (Tseng et al. 2007). For example, if we observe many concurrences of incorrectly answered records of topics A and B, and learners tend to perform poorly on topic B when they perform poorly on topic A, we can infer that topic B depends on topic A (i.e., an association rule A → B).

An important decision in fuzzy association rule mining is to determine the thresholds for support (i.e., the concurrence of poor performance on both topics), α, and confidence (i.e., the proportion of poor performance on topic B given poor performance on topic A), β (Chen and Wei 2002; Tseng et al. 2007). Choosing higher support and confidence thresholds (α and β) can lead to a higher quality of the discovered rules but may result in omissions of qualified rules. In addition, in our context, these choices also have implications for the degree of conflict between the set of discovered rules M_d and the set of expert rules M_0, which are held to be true but incomplete. Combining these considerations, we choose α and β to maximize $(\alpha + \beta) \sum_{(l \in M_0 \cap l \in M_d)} (Confidence(l))$. This heuristic objective function trades off the agreement between the discovered rules and expert rules, as measured by the sum of the confidence of rules at the intersection of the two sets of rules, and the quality of discovered rules (combining support and confidence). Once we determine the two thresholds, we use them in subsequent rule mining to reveal the hidden dependencies among topics and then dynamically update the knowledge map. Appendix B shows the knowledge map updating details.

For example, in the current context, "making inferences" and "retrieving relevant information" are two different learning topics in English reading comprehension. Learning how to retrieve relevant information from an article can facilitate learners in making correct inferences. Hence, "making inferences" depends on "retrieving relevant information." Figure 2 shows the terminal state knowledge map in our context. We validate our entire approach by

demonstrating the terminal-state knowledge map to the senior teachers, who agreed with the new rules generated by fuzzy association rule mining.

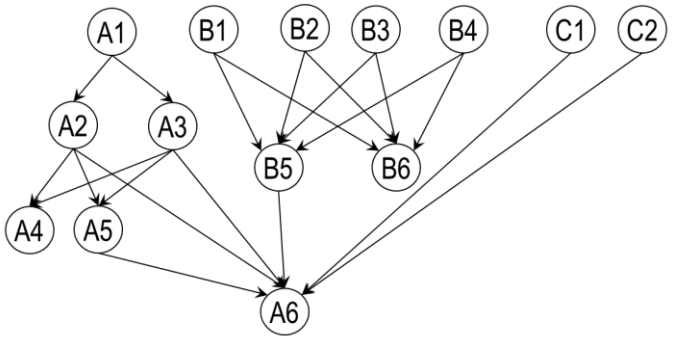


Figure 2. Terminal-State Knowledge Map

Notes: A circle represents a topic. An edge represents that a topic depends on a preceding topic. For example, A1 → A2 captures that topic A2 depends on topic A1.

Meta-Design III: Scheduling Engine for Choosing Learning Materials

With weak topics detected and topic relatedness discovered, the next question is how to deliver materials that cover mixed topics to each learner. From the perspective of CLT, learning is more effective when the learning materials are personalized and dynamically adjusted to reflect learners' progress so that their cognitive load can be optimized. This calls for a scheduling engine. In our learning context, learners learn through exercises. Each exercise consists of a set of multiple-choice assessment questions and each question covers a single topic (but questions in the same exercise may collectively cover a few topics). The goal of the scheduling engine is to choose the set of questions for each learning session to meet the session design objectives, including the number of topics covered, topic mastery level, and topic-relatedness.

We implement both *related-interleaving* and two benchmark session designs (i.e., *non-interleaving* and *unrelated-interleaving*). For the *non-interleaving* design, we choose exercises that have the highest concentration on the topic with the highest weakness ranking (i.e., probability of being unmastered), with concentration defined as the percentage of

questions covering a topic in an exercise. In most cases, the chosen exercise was 100% concentrated on one topic, meaning that all questions were about the same topic.

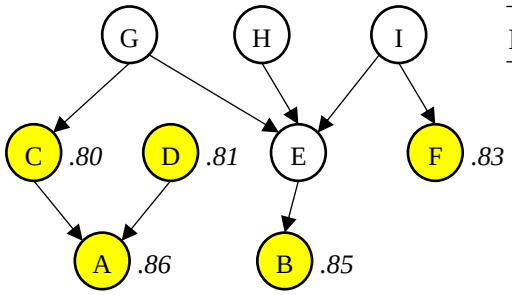
As an illustrative example, Figure 3 depicts a knowledge map that models the relatedness among nine topics (A to I). In this example, the learner has five unmastered topics A, B, C, D, and F (yellow-shaded). The number shown next to an unmastered topic denotes the *weakness* probability; i.e., the probability that the learner has not mastered the topic. Suppose we have a list of unattempted exercises covering different topics as shown in the figure. By our design, topic A has the highest weakness probability (p=0.86) and thus will be the topic for the current learning session. Exercise a, which has the highest concentration of topic A, is therefore used for the current learning session for the *non-interleaving* design.

For the *unrelated-interleaving* design, the scheduling engine chooses exercises that cover the largest number of unmastered topics and breaks ties by the aggregated weakness ranking of the topics covered in the exercise.[4] Continuing with the example in Figure 3, the schedule engine will choose Exercise c, which covers three unmastered topics with the highest weakness ranks.

For the *related-interleaving* design, we choose exercises that cover the largest number of unmastered topics and break ties based on a modified topic ranking that considers both weakness and relatedness among topics. The modified topic ranking is constructed as follows. Starting from topic *A*, which has the highest weakness ranking, we find all the topics that topic A depends on in the knowledge map, such as topic *C*. If *C* is an unmastered topic, we insert *C* just before *A* in the ranking. This is to ensure that when we select topic *A*, we also include unmastered topic *C*, which topic A depends on. We do this repeatedly for all

[4] For example, consider Exercise b, which covers three unmastered topics with weakness rankings of 1, 2, and 4; Exercise c, which covers three unmastered topics with weakness rankings of 1, 2, and 3; and Exercise d, which covers two unmastered topics with weakness rankings of 1 and 2. We prefer Exercises b and c over d because the former two cover more unmastered topics. Between Exercises b and c, we prefer Exercise c because it has an aggregated weakness ranking of 6 (= 1 + 2 + 3), which ranks higher than Exercise b with aggregated weakness ranking of 7 (= 1 + 2 + 4).

unmastered topics to arrive at a modified ranking of topics (see Appendix C for a pseudo algorithm). We then calculate an aggregated ranking for each exercise by summing up the modified rankings of all unmastered topics covered. We select the exercise with the highest aggregated ranking to represent a related-interleaving design (see Appendix C for a pseudo-scheduling algorithm). Continuing with the example in Figure 3, the schedule engine will generate a modified topic ranking (i.e., D, C, A, B, F), based on which the modified aggregated weakness will be calculated for each exercise. Then, the schedule engine will choose Exercise e for related-interleaving, since its modified aggregated weakness ranks higher than the other exercises.



Numbers are weakness probabilities

Topic Ranking: *A, B, F, D, C*

Modified Topic Ranking: *D, C, A, B, F*

Exercise	Covered Topics	
a	A, A, A	← Non-interleaving
b	D, A, B	
c	A, B, F	← Unrelated-interleaving
d	E, A, B	
e	C, D, A	← Related-interleaving

Figure 3. An Example of Related-Interleaving Session Design

Overall System Architecture

Figure 4 depicts how the three components combine to form a whole system. At each iteration, the Weak Topic Detection module learns a user's weak topics from existing records. The Topic Relatedness Modeling module updates the knowledge map reflecting the new performance data. The Scheduling Engine module selects and ranks the weak topics and uses the ranked topics to choose unattempted exercises for the learner.

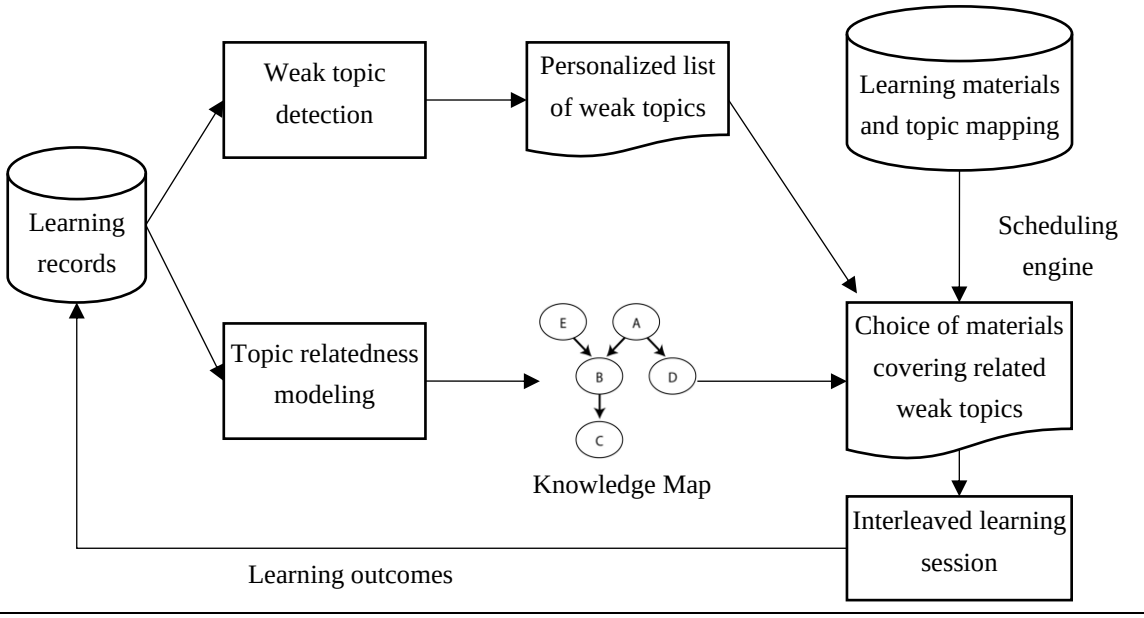


Figure 4: A System Architecture for Related-interleaving

Please note that after learners complete a learning session, we store new learning outcomes in the repository so that they can be considered when choosing materials for the next learning session. For example, because topic weakness rankings may change, newly mastered topics could be removed and the knowledge map may be adjusted as the estimations of dependencies change. Figure C1 in Appendix C provides an example of how topic weakness evolves across multiple learning sessions.

EXPERIMENTAL EVALUATION

After designing related-interleaving for e-learning platforms, the next step is to evaluate our design against the benchmarks. We start by formulating testable hypotheses for our experimental evaluation and then describe the study context and experiment design.

Testable Hypotheses

We use a field experiment to test two hypotheses derived from our design goals. Based on our theory-driven design discussed earlier, compared with unrelated-interleaving, related-interleaving mitigates the “overload risk” and expands “schema-building opportunities,” thereby leading to improved learning. In addition, our previous discussion indicates that the

relative advantage of interleaving versus non-interleaving primarily hinges upon whether learners are overloaded in interleaved learning. We thus propose that since related-interleaving possesses the strength of interleaving in schema building while reducing learners' cognitive load, it is also likely to outperform non-interleaving. Thus, we hypothesize that,

H1: *Compared with unrelated-interleaving, related-interleaving leads to better learning performance.*

H2: *Compared with non-interleaved learning, related-interleaving leads to better learning performance.*

While our main focus is on the advantage of the related-interleaving design, we are also interested in testing the heterogeneous effects of related-interleaving across learner types, which is useful for guiding future designs. CLT suggests that individuals differ in cognitive capacities and the amount of schemas they can leverage to lessen their cognitive load (Sweller et al. 2019). As a result, weak learners who have not built strong schemas from their past learning may face greater "overload risk" under interleaving than strong learners. For weak learners, unrelated-interleaving may leave few cognitive resources for schema building and prevent them from realizing the "schema-building enhancement" benefit, leading to suboptimal performance. Increasing topic relatedness in interleaving is thus beneficial for weak learners because it mitigates the "overload risk" and leaves more resources for schema building (Rau et al. 2010). Strong learners, however, with large cognitive capacities and more existing schemas to rely on, are less resource constrained and may thus not benefit as much from the reduced cognitive load of increasing topic relatedness.

Furthermore, weak learners also stand to benefit more from the increased "schema-building opportunities" in related-interleaving. This is because weak learners are generally less capable of information abstraction and connection building than strong learners. Juxtaposing related topics in the same session thus facilitates weak learners' information connection and induction to a greater extent (Lambiotte and Dansereau 1992; Nesbit and Adesope 2006). Hence, we expect that in the interleaving design, *increasing topic relatedness is more beneficial for weak learners than for strong learners*.

Study Context and System Instantiation

To evaluate the effectiveness of our proposed related-interleaving session design, we collaborate with a middle school in China to supplement a mandatory English course. During summer and winter breaks, English teachers at this school assign learners English reading exercises. Each exercise consists of one article and three to five multiple-choice questions for assessing learners' comprehension. Traditionally, teachers distribute booklets that contain the same exercises for all learners. Using a third-party developer, we develop an e-learning system to replace the booklets. The system works as follows: The system assigns two exercises to each learner at the beginning of every learning session, each consisting of two consecutive days. Exercises that are not completed within the learning session will expire. After a learner submits his/her answers to each exercise online, the system provides immediate feedback, including displaying the correct answers, the learning topics covered, and explanations of why an answer is correct. We explain the system details in Appendix D.

There are about 200 exercises in our system database. Each question in an exercise is designed to cover one of the 14 topics designated by the education bureau for English reading. These topics are related to specific skills in reading comprehension, such as "making inferences," "event sorting," "summarizing the main idea," and "retrieving relevant information." The exercises, along with the assessment questions, are designed to reinforce

and evaluate a learner's mastery of these core topics. Experienced English teachers from leading middle schools in the region coded the topic covered by each assessment question. Some exercises have all questions covering the same topic, whereas other exercises have questions covering different topics. We leverage such natural variations and implement different learning session designs by choosing exercises with a desirable number of topics and a level of topic relatedness, as shown in the section on the scheduling engine design.

Experiment Design and Procedure

We conducted a field experiment during the summer break from July 13 to August 31, 2017. We adopted a between-subject design with three conditions: non-interleaving, unrelated-interleaving, and related-interleaving. Our participants included 510 eighth-grade learners from 17 different classes in the middle school, taught by nine different English teachers. We did not inform learners of their assigned conditions upfront nor did we inform the teachers. Because the system interface was identical for the three groups, it was unlikely that the learners could infer their assignment group. Learners from the same class were randomly assigned to the three groups with equal probability so that any teacher effect would be canceled in cross-group comparisons. After the experiment, we debriefed the learners and teachers who participated in the experiment.

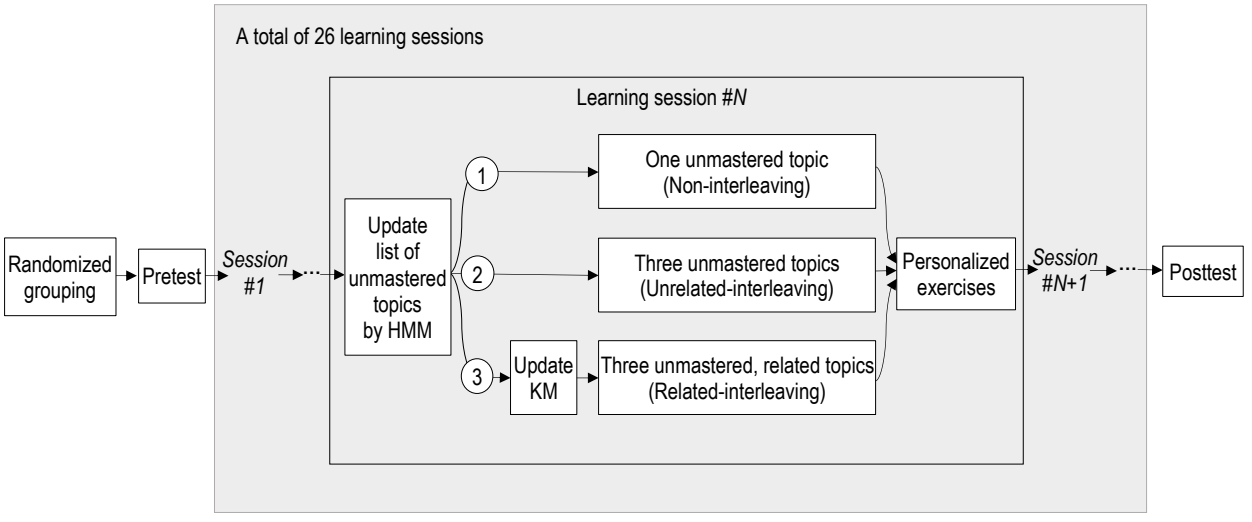


Figure 5. Experimental Procedure

Notes: HMM = Hidden Markov Model; KM = Knowledge Map. ①, ②, and ③ indicate the three treatment groups.

The experimental procedure is depicted in Figure 5. Before the experiment, we collected the learners' demographic information and their final exam scores in English, taken about one week before the experiment. These scores were later used to classify learners into weak and strong learners. Next, we randomly assigned the learners to the three groups with equal probability. The learners stayed in the same group throughout the experiment. We then conducted an online pretest that included several exercises to control for the learners' reading comprehension skills before the experiment.

The experiment included 26 learning sessions. Before each session, the HMM automatically recalculated each learner's topic mastery based on up-to-date learning records. The system also updated the knowledge map based on pooled performance data. The system then selected personalized exercises for the learning session based on the learner's treatment group. After the experiment, we conducted a posttest of reading comprehension to evaluate the learners' performance.

RESULTS AND ANALYSIS

Among the 510 learners invited to use the system, 435 voluntarily participated and finished at least one exercise during the experiment. The three groups (non-interleaving, unrelated-interleaving, and related-interleaving) were roughly equal in size, with 140, 149, and 146 participants, respectively. Among all the participants, 381 took the posttest, and 337 took the pretest. Our analyses focus on learners who took both the posttest and the pretest. We classify learners into weak and strong based on their English final-exam scores before the experiment. Specifically, we define those whose scores were above the class median as strong learners and the remaining as weak learners. We use the *Posttest* score (based on a 100-point scale) as our main measure of learning performance.

In Appendix E, we report a series of randomization and manipulation checks. The randomization check suggests no significant differences in terms of pre-experiment exam

scores, pretest scores, and gender distributions across groups. Our manipulation check confirms that our intervention is valid. Specifically, we find that the non-interleaving design covered 1.16 topics per exercise on average, whereas the two interleaving designs covered 3.40 topics per exercise on average. In addition, the related-interleaving design had 2.29 topic dependencies per exercise on average, which is a significant increase from 0.47 under the unrelated-interleaving design.

We report the summary statistics in Table 2. Among the participants, 46% were female. The learners spent 159.47 minutes in the system, on average, during the entire experiment period and each exercise took approximately 5.04 minutes to finish. The learners completed 31.65 exercises out of the 54 assigned, on average, for a 61% completion rate. The average accuracy among the completed exercises was 65%.

Table 2: Summary Statistics

Variables	Description	N	Mean	SD	Min	Max
Interleaving	=1 if the learner was randomly assigned to the unrelated-interleaving or related-interleaving group	435	0.68	0.47	0	1
Relatedness	=1 if the learner was randomly assigned to the related-interleaving group	435	0.34	0.47	0	1
Female	=1 if learner was female	435	0.46	0.50	0	1
FinalScore	English final exam score before the experiment	427	81.39	13.12	23.33	98.75
StrongLearner	=1 if learner's pre-experiment final exam score was above the class median	427	0.51	0.50	0	1
Pretest	Online pretest score before the experiment	337	64.77	24.91	0	100
Posttest	Online posttest score after the experiment	381	69.22	28.29	0	100

Model-free Evidence

We first conduct a model-free analysis by comparing the posttest scores across the three groups (Figure 6). Overall, related-interleaving leads to an 8.28-point increase in posttest

scores compared with non-interleaving (p=0.021) and a 10.90-point increase compared with unrelated-interleaving (p<0.001). Unrelated-interleaving results in a 2.82-point decrease compared with non-interleaving, though the effect is insignificant (p=0.459).

We find a similar pattern among weak learners. Related-interleaving enables them to gain 7.40 points relative to non-interleaving (p=0.200) and 15.88 points relative to unrelated-interleaving (p=0.002). Unrelated-interleaving leads to an 8.47-point drop compared with non-interleaving, though the effect is insignificant (p=0.14). The effects among strong learners are different: related-interleaving leads to a significant 10.58-point increase compared with non-interleaving (p=0.017) and a less prominent 6.03-point increase compared with unrelated-interleaving (p=0.179). For stronger learners, unrelated-interleaving does not differ from non-interleaving (β=4.55, p=0.291).

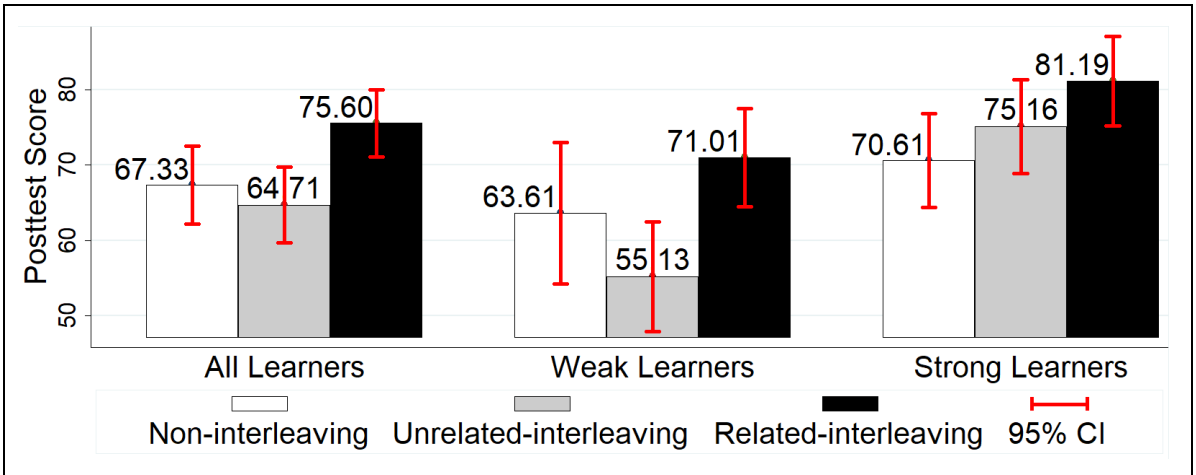


Figure 6: Model-free Evidence

Effect of Related-interleaving on Posttest Scores

To test the effect of related-interleaving on learning performance, we model the learning performance of a learner i as follows:

$$Posttest_i = \alpha + \beta_1 Interleaving_i + \beta_2 Interleaving_i \times Relatedness_i + \boldsymbol{x_i} + \boldsymbol{\omega_i} + \epsilon_i$$

where $Posttest_i$ represents the posttest performance of learner i and vector $\boldsymbol{x_i}$ denotes learner characteristics including the pre-experiment final exam score (*FinalScore*), the pretest score (*Pretest*), and gender (*Female*). Vector $\boldsymbol{\omega_i}$ is a class-fixed effect. ϵ_i denotes the idiosyncratic variation in learning performance.

Table 3: Effect of Interleaving and Topic Relatedness

	Posttest	Posttest
	(1)	(2)
Interleaving	-2.97	-19.86***
	(3.33)	(5.38)
Interleaving×*Relatedness*	10.05**	18.81***
	(3.21)	(4.48)
Interleaving×*StrongLearner*		25.18***
		(6.98)
Interleaving×*Relatedness*×*StrongLearner*		-15.86*
		(6.56)
StrongLearner		-7.13
		(5.35)
FinalScore	0.83***	
	(0.16)	
Pretest	0.30***	0.36***
	(0.06)	(0.06)
Female	6.10*	6.84*
	(2.78)	(2.82)
Constant	-18.81	58.46***
	(16.22)	(7.48)
Class fixed effect	YES	YES
N	306	306
R^2	0.359	0.343

Notes: *Interleaving* represents the effect of unrelated-interleaving relative to non-interleaving. *Relatedness* is meaningful only in the interleaving condition (i.e., coded as 0 for both unrelated-interleaving and non-interleaving conditions). Therefore, *Interleaving*×*Relatedness* is equivalent to *Relatedness* and captures the effect of related-interleaving relative to unrelated-interleaving. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors are in parentheses.

Based on the regression results (see Table 3, column 1), *Relatedness* significantly moderates the effect of *Interleaving*. To aid understanding, we compare related-interleaving with two other conditions in Figure 7. Compared with unrelated-interleaving, learners in the related-interleaving group score 10.05 points higher ($p = 0.002$). Compared with non-

interleaving, related-interleaving leads to a 7.08-point increase (p = 0.038)[5]. We thus find support for hypotheses **H1** and **H2.** Overall, we find that related-interleaving significantly increased learning performance.

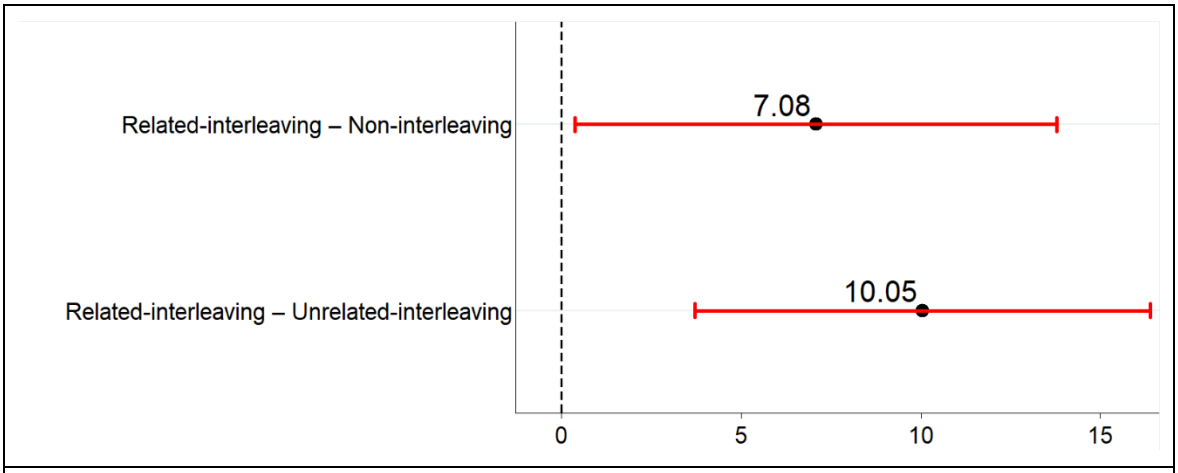


Figure 7. Posttest Comparison among Three Experimental Conditions
Notes: The dots represent mean values. The error bars represent the 95% confidence intervals.

We also compare the posttest performance between non-interleaving and unrelated-interleaving and find no significant difference (β = -2.97, p = 0.374). This indicates that the traditional unrelated-interleaving design does not yield better learning performance than the non-interleaving design.

To further examine whether related-interleaving increases the amount of schema building as predicted by CLT, we also compare the three groups based on two indirect measures of schema-building loads as suggested by prior literature; namely, a learner's *topic mastery* and *practice accuracy* after each learning session (Brunken et al. 2003; Orru and Longo 2019). As shown in Appendix F, related-interleaving leads to increased topic mastery and practice accuracy in each learning session than unrelated-interleaving and non-interleaving, supporting our theoretical predictions.

[5] This is calculated as a combined effect of *Interleaving* and *Interleaving*×*Relatedness* in Table 3.

Effect of Related-interleaving by Learner Type

We further explore the heterogeneous effects of related-interleaving across different learners. To understand how the benefit of related-interleaving differs by learner type, we add a three-way interaction term between interleaved learning (*Interleaving*), topic relatedness (*Relatedness*), and learner type (*StrongLearner*).[6]

$$Posttest_i = \alpha + \beta_1 Interleaving_i + \beta_2 Interleaving_i \times Relatedness_i + \beta_3 Interleaving_i \times StrongLearner_i + \beta_4 Interleaving_i \times Relatedness_i \times StrongLearner_i + \boldsymbol{x_i} + \boldsymbol{\omega_i} + \epsilon_i.$$

The results from Table 3 Column 2 show a significantly negative three-way interaction, which confirms that *increasing topic relatedness benefited weak learners more than strong learners*. To aid understanding, we first plot the findings in Figure 8, which shows that weak learners (solid line) achieve significantly better performance (an 18.81-point increase) with related-interleaving than unrelated-interleaving. In contrast, the improvement brought by related-interleaving for strong learners (dashed line) is not significant (β = 2.94, p = 0.538). Increasing topic relatedness also significantly reduces the performance gap between strong and weak learners from 18.04 points (with unrelated-interleaving) to 2.18 points (with related-interleaving). Appendix G shows that our heterogeneous analyses are not sensitive to the division between strong and weak learners.

Another notable finding is that a significant positive interaction exists between interleaved learning (*Interleaving*) and learner type (*StrongLearner*) (Table 3, column 2). This indicates that traditional unrelated-interleaving (relative to non-interleaving) benefits strong learners (a 5.31-point increase in posttest scores), but hurts weak learners (a 19.86-point drop). This is consistent with the CLT framework that suggests that interleaving may significantly increase cognitive overload risks among weak learners (who have not built

[6] *Interleaving*×*Relatedness*×*StrongLearner* is equivalent to *Relatedness*×*StrongLearner* because *Relatedness* is coded as 0 for both unrelated-interleaving and non-interleaving conditions.

strong schemas previously), dampening their learning performance. For strong learners, interleaving is less likely to lead to cognitive overload because they have more existing schemas to rely on. Overall, these findings support the CLT perspective that when learners' cognitive resources are strained (as in the case of weak learners), unrelated-interleaving can hurt learning performance, highlighting the importance of topic relatedness.

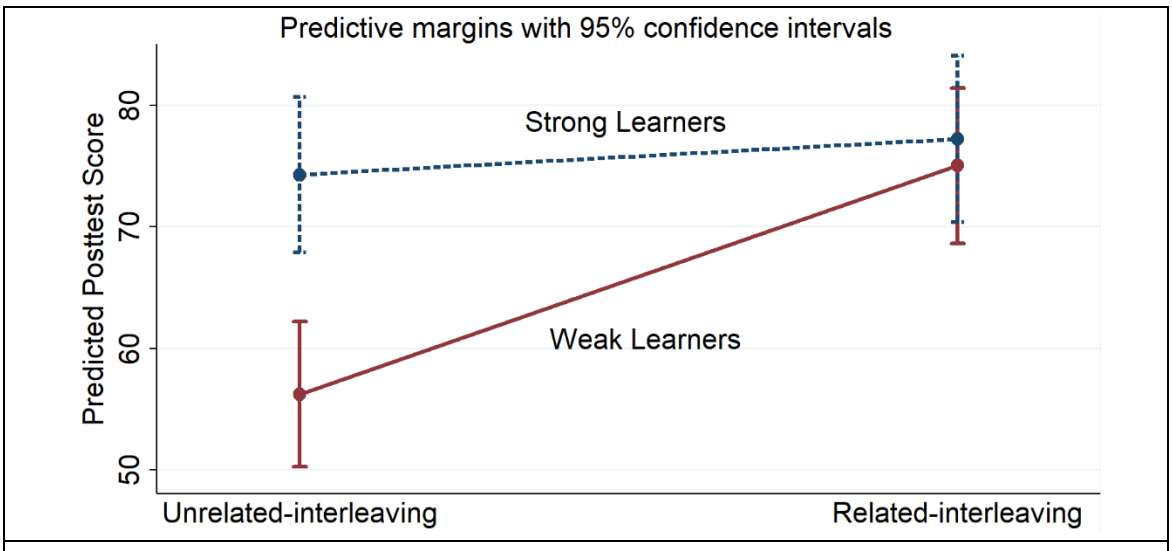


Figure 8. Heterogeneous Effect of Topic Relatedness on Posttest Score
Notes: Dots represent the mean values. Error bars represent the 95% confidence intervals.

DISCUSSION

Contributions to the Literature

This study contributes to the literature in four ways. First, following the design science paradigm, we contribute to the e-learning literature by developing and testing a theory-grounded interleaving design — related-interleaving. Our related-interleaving design incorporates (a) the dynamic detection of learners' weak topics at each moment using a hidden Markov chain, (b) a knowledge-map-based representation for capturing topic dependencies and a fuzzy association rule algorithm for data-driven augmentation of the knowledge map, and (c) a scheduling engine that assembles a set of exercises that meet the requirements of related-interleaving, personalization, and adaptation. Based on a field

experiment, we showcase how one can harness the power of machine learning in data-rich e-learning environments to make learning session design more adaptive and effective.

Second, more broadly, we extend the stream of IS research on structuring e-learning activities (Alavi and Leidner 2001; Gupta and Bostrom 2013; Gupta and Bostrom 2009; Piccoli et al. 2001) by examining topic design in a learning session. Prior studies have mainly focused on weak topic detection and how to optimize the design across multiple learning sessions. We have gone further to show the importance of choosing multiple weak topics to practice in one learning session. We believe that this new design, related-interleaving, can generate meaningful implications in other e-learning design elements such as multimedia design and instructional strategies.

Third, our work contributes to the literature on interleaving. Existing literature has shown mixed evidence in terms of the interleaving effect, but there is very limited theorization about the potential downsides of interleaving. Additionally, there is also limited attention on how to design interleaving effectively. This study uses CLT as a theoretical framework to explain the benefits and risks of interleaving and to motivate a new related-interleaving design that mitigates the risk of cognitive overload while maintaining schema-building opportunities. Our findings confirm two novel predictions of the theory: the benefit of related-interleaving and the differential effects of interleaving across learner types. The contingency factors identified in this study — topic relatedness and learner type (i.e., strong and weak learners) — may help explain the mixed findings about interleaving in the literature. Moreover, our CLT-based framework may serve as a theoretical foundation for new interleaving research and designs. For instance, researchers can use it to investigate the optimal spacing of interleaved topics and the role of complexity level of exercises in interleaved designs.

Lastly, this study contributes to the existing CLT literature by adding that topic design in a learning session also holds important implications for cognitive load and learning

performance. Prior research has used CLT to guide various aspects of instructional design, such as information presentation formats, the use of worked examples, and the modality of instructions. This research shows that CLT can also guide interleaving designs, both in terms of the number of topics in a learning session and the relationship between these topics. Building on tenets of CLT, our research suggests a nuanced relationship between interleaving and learning performance: increasing the number of topics in a session (i.e., interleaving) can elevate cognitive load and potentially hinder learning, particularly among weak learners; however, when the topics included in the same session are more related, the risk of overload is reduced, allowing learners, especially weak learners, to benefit more from schema-building opportunities brought by interleaving and thus obtain better learning performance. Overall, our work extends the domain of CLT by revealing a complex relationship between interleaving designs, individual differences, cognitive load, and learning performance.

Implications for Practice

Our findings have several actionable implications for practice. First, we highlight an overlooked issue in e-learning session design and offer a few alternatives — non-interleaving, unrelated-interleaving, and related-interleaving session designs. As more e-learning platforms begin to offer personalized learning materials for learners, the issue will become increasingly relevant. Second, our findings suggest that e-learning platforms should not blindly mix topics in an interleaved learning session; ensuring topic relatedness is a great way of enhancing the benefits of interleaved learning while mitigating overload risks. Third, our findings suggest that e-learning platforms should consider learners' capacities when designing learning sessions. Interleaving generally benefits strong learners and our proposed related-interleaving is likely the best design for them. For weak learners, both related-interleaving and traditional non-interleaving designs are suitable, but unrelated-interleaving design (i.e., interleaving without enforcing topic relatedness) should be avoided. Finally, we provide several tools, such as a hidden Markov chain and a knowledge map, that can be

directly appropriated by practitioners to support personalized, adaptive, and data-driven designs, such as related-interleaving. We hope that our work can inspire more e-learning practitioners to incorporate data-driven decisions and machine intelligence into their learning designs.

Limitations and Future Research

Our study has several limitations that can be addressed in future research. *First*, our findings are based on a field experiment on a specific subject (English) and population. The research can benefit from replications to other subject domains and learner populations. *Second*, our field experiment lasted only two months and thus may not capture long-term effects. *Third*, we used certain heuristics in our implementation of related-interleaving that could be further optimized. Future research could experiment with a different number of interleaved topics and different ways of choosing among weak topics. For example, for strong learners, given that handling three related topics with the highest weakness rankings works well in our study, future studies could consider increasing the number of related topics to fully tap the potential of strong learners. However, for weak learners, our findings show that, although mixing three related topics may mitigate the downside of unrelated-interleaving, it is no better than non-interleaving. Hence, mixing fewer related topics or choosing topics with a lower weakness ranking may further lessen cognitive load and potentially benefit weak learners more. Existing CLT-based design principles could also be utilized to reduce cognitive load for weaker learners. Strategies like offering worked examples, partial solutions, and integrating multi-modal information could be leveraged to further assist weak learners in getting the full benefits of related-interleaving. In addition, future research could further explore different implementations of topic relatedness (e.g., ones focusing on topic similarity) and how optimal implementations may depend on specific contexts (e.g., different subject domains and learning conditions). *Fourth*, given our findings on how the effects of different interleaving designs differ across weak and strong learners,

future research could explore how the design of interleaving should be adapted dynamically as learners gain proficiency. *Finally*, although our findings support CLT predictions, we could not directly test the theory in our field experiment. Further tests of theoretical mechanisms may be a good subject for future research.

REFERENCE

Abbasi, A., and Chen, H. 2008. "Cybergate: A Design Framework and System for Text Analysis of Computer-Mediated Communication," *MIS Quarterly* (32:4), pp. 811-837.

Abdelrahman, G., Wang, Q., and Nunes, B. 2023. "Knowledge Tracing: A Survey," *ACM Computing Surveysy* (55:11), p. Article 224.

Al-Muhaideb, S., and Menai, M. E. B. 2011. "Evolutionary Computation Approaches to the Curriculum Sequencing Problem," *Natural Computing* (10:2), pp. 891-920.

Alavi, M., and Leidner, D. E. 2001. "Research Commentary: Technology-Mediated Learning - a Call for Greater Depth and Breadth of Research," *Information Systems Research* (12:1), pp. 1-10.

Atapattu, T., Falkner, K., and Falkner, N. 2017. "A Comprehensive Text Analysis of Lecture Slides to Generate Concept Maps," *Computers & Education* (115:1), pp. 96-113.

Ayabakan, S., Bardhan, I., and Zheng, E. 2016. "What Drives Patient Readmissions? A New Perspective from the Hidden Markov Model Analysis," *in Proceedings of the 37th International Conference on Information Systems*, Dublin, Ireland.

Balaid, A., Abd Rozan, M. Z., Hikmi, S. N., and Memon, J. 2016. "Knowledge Maps: A Systematic Literature Review and Directions for Future Research," *International Journal of Information Management* (36:3), pp. 451-475.

Bauman, K., and Tuzhilin, A. 2018. "Recommending Remedial Learning Materials to Students by Filling Their Knowledge Gaps," *MIS Quarterly* (42:1), pp. 313-332.

Bettinger, E. P., Fox, L., Loeb, S., and Taylor, E. S. 2017. "Virtual Classrooms: How Online College Courses Affect Student Success," *American Economic Review* (107:9), pp. 2855-2875.

Birnbaum, M. S., Kornell, N., Bjork, E. L., and Bjork, R. A. 2013. "Why Interleaving Enhances Inductive Learning: The Roles of Discrimination and Retrieval," *Memory & Cognition* (41:3), pp. 392-402.

Brunken, R., Plass, J. L., and Leutner, D. 2003. "Direct Measurement of Cognitive Load in Multimedia Learning," *Educational Psychologist* (38:1), pp. 53-61.

Carvalho, P. F., and Goldstone, R. L. 2014. "Putting Category Learning in Order: Category Structure and Temporal Arrangement Affect the Benefit of Interleaved over Blocked Study," *Memory & Cognition* (42:3), pp. 481-495.

Chen, C. M. 2008. "Intelligent Web-Based Learning System with Personalized Learning Path Guidance," *Computers & Education* (51:2), pp. 787-814.

Chen, G., and Wei, Q. 2002. "Fuzzy Association Rules and the Extended Mining Algorithms," *Information Sciences* (147:1), pp. 201-228.

Chen, Y., Li, X., Liu, J., and Ying, Z. 2018. "Recommendation System for Adaptive Learning," *Applied Psychological Measurement* (42:1), pp. 24-41.

Conrad, C., Deng, Q., Caron, I., Shkurska, O., Skerrett, P., and Sundararajan, B. 2022. "How Student Perceptions About Online Learning Difficulty Influenced Their Satisfaction During Canada's Covid-19 Response," *British Journal of Educational Technology* (53:3), pp. 534-557.

Damgaard, M. T., and Nielsen, H. S. 2018. "Nudging in Education," *Economics of Education Review* (64), pp. 313-342.

Delgado, P., and Salmerón, L. 2021. "The Inattentive on-Screen Reading: Reading Medium Affects Attention and Reading Comprehension under Time Pressure," *Learning and Instruction* (71), p. 101396.

Dennen, V. P., Aubteen Darabi, A., and Smith, L. J. 2007. "Instructor–Learner Interaction in Online Courses: The Relative Perceived Importance of Particular Instructor Actions on Performance and Satisfaction," *Distance Education* (28:1), pp. 65-79.

Dontre, A. J. 2021. "The Influence of Technology on Academic Distraction: A Review," *Human Behavior and Emerging Technologies* (3:3), pp. 379-390.

Figlio, D., Rush, M., and Yin, L. 2013. "Is It Live or Is It Internet? Experimental Estimates of the Effects of Online Instruction on Student Learning," *Journal of Labor Economics* (31:4), pp. 763-784.

Firth, J., Rivers, I., and Boyle, J. 2021. "A Systematic Review of Interleaving as a Concept Learning Strategy," *Review of Education* (9:2), pp. 642-684.

Foster, N. L., Mueller, M. L., Was, C., Rawson, K. A., and Dunlosky, J. 2019. "Why Does Interleaving Improve Math Learning? The Contributions of Discriminative Contrast and Distributed Practice," *Memory & Cognition* (47:6), pp. 1088-1101.

Furenes, M. I., Kucirkova, N., and Bus, A. G. 2021. "A Comparison of Children's Reading on Paper Versus Screen: A Meta-Analysis," *Review of Educational Research* (91:4), pp. 483-517.

Ginns, P. 2005. "Meta-Analysis of the Modality Effect," *Learning and Instruction* (15:4), pp. 313-331.

Goudeau, S., Sanrey, C., Stanczak, A., Manstead, A., and Darnon, C. 2021. "Why Lockdown and Distance Learning During the Covid-19 Pandemic Are Likely to Increase the Social Class Achievement Gap," *Nature Human Behaviour* (5:10), pp. 1273-1281.

Gupta, S., and Bostrom, R. 2013. "Research Note—an Investigation of the Appropriation of Technology-Mediated Training Methods Incorporating Enactive and Collaborative Learning," *Information Systems Research* (24:2), pp. 454-469.

Gupta, S., and Bostrom, R. P. 2009. "Technology-Mediated Learning: A Comprehensive Theoretical Model," *Journal of the Association for Information Systems* (10:9), pp. 686-714.

Hansen, J. D., and Reich, J. 2015. "Democratizing Education? Examining Access and Usage Patterns in Massive Open Online Courses," *Science* (350:6265), pp. 1245-1248.

Hausman, H., and Kornell, N. 2014. "Mixing Topics While Studying Does Not Enhance Learning," *Journal of Applied Research in Memory and Cognition* (3:3), pp. 153-160.

Huang, N., Wang, L., Hong, Y., Lin, L., Guo, X., and Chen, G. 2023. "When the Clock Strikes: A Multimethod Investigation of on-the-Hour Effects in Online Learning," *Information Systems Research* (Forthcoming).

Huang, N., Zhang, J., Burtch, G., Li, X., and Chen, P. 2021. "Combating Procrastination on Moocs Via Optimal Calls-to-Action: Evidence from a Field Experiment," *Information Systems Research* (32:2), pp. 301-317.

Hussain, M., Zhu, W., Zhang, W., Abidi, S. M. R., and Ali, S. 2019. "Using Machine Learning to Predict Student Difficulties from Learning Session Data," *Artificial Intelligence Review* (52:1), pp. 381-407.

Jaeger, A. J., Taylor, A. R., and Wiley, J. 2016. "When, and for Whom, Analogies Help: The Role of Spatial Skills and Interleaved Presentation," *Journal of Educational Psychology* (108:8), pp. 1121-1139.

Jeng, Y.-L., and Huang, Y.-M. 2019. "Dynamic Learning Paths Framework Based on Collective Intelligence from Learners," *Computers in Human Behavior* (100), pp. 242-251.

Jiang, B., Li, X., Yang, S., Kong, Y., Cheng, W., Hao, C., and Lin, Q. 2022. "Data-Driven Personalized Learning Path Planning Based on Cognitive Diagnostic Assessments in Moocs," *Applied Sciences* (12:8), p. 3982.

Kalyuga, S. 2007. "Enhancing Instructional Efficiency of Interactive E-Learning Environments: A Cognitive Load Perspective," *Educational Psychology Review* (19:3), pp. 387-399.

Khan, A., Egbue, O., Palkie, B., and Madden, J. 2017. "Active Learning: Engaging Students to Maximize Learning in an Online Course," *Electronic Journal of e-learning* (15:2), p. 107-115.

Kim, N. J., Belland, B. R., Lefler, M., Andreasen, L., Walker, A., and Axelrod, D. 2020. "Computer-Based Scaffolding Targeting Individual Versus Groups in Problem-Centered Instruction for Stem Education: Meta-Analysis," *Educational Psychology Review* (32:2), pp. 415-461.

Kim, Y., and Krishnan, R. 2019. "The Dynamics of Online Consumers' Response to Price Promotion," *Information Systems Research* (30:1), pp. 175-190.

Kizilcec, R. F., Saltarelli, A. J., Reich, J., and Cohen, G. L. 2017. "Closing Global Achievement Gaps in Moocs," *Science* (355:6322), pp. 251-252.

Klašnja-Milićević, A., Vesin, B., Ivanović, M., and Budimac, Z. 2011. "E-Learning Personalization Based on Hybrid Recommendation Strategy and Learning Style Identification," *Computers & Education* (56:3), pp. 885-899.

Kleider, H. M., Pezdek, K., Goldinger, S. D., and Kirk, A. 2008. "Schema-Driven Source Misattribution Errors: Remembering the Expected from a Witnessed Event," *Applied Cognitive Psychology* (22:1), pp. 1-20.

Kornell, N., and Bjork, R. A. 2008. "Learning Concepts and Categories: Is Spacing the 'Enemy of Induction'?," *Psychological Science* (19:6), pp. 585-592.

Kulkarni, C., Cambre, J., Kotturi, Y., Bernstein, M. S., and Klemmer, S. R. 2015. "Talkabout: Making Distance Matter with Small Groups in Massive Classes," in: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. Vancouver, BC, Canada: pp. 1116–1128.

Kurilovas, E., Zilinskiene, I., and Dagiene, V. 2015. "Recommending Suitable Learning Paths According to Learners' Preferences: Experimental Research Results," *Computers in Human Behavior* (51), pp. 945-951.

Lambiotte, J. G., and Dansereau, D. F. 1992. "Effects of Knowledge Maps and Prior Knowledge on Recall of Science Lecture Content," *The Journal of Experimental Education* (60:3), pp. 189-201.

Lee, J. H., and Segev, A. 2012. "Knowledge Maps for E-Learning," *Computers & Education* (59:2), pp. 353-364.

Leung, A. C. M., Santhanam, R., Kwok, R. C.-W., and Yue, W. T. 2023. "Could Gamification Designs Enhance Online Learning through Personalization? Lessons from a Field Experiment," *Information Systems Research* (34:1), pp. 27-49.

Likourezos, V., Kalyuga, S., and Sweller, J. 2019. "The Variability Effect: When Instructional Variability Is Advantageous," *Educational Psychology Review* (31:2), pp. 479-497.

Loghin, G. C., Carron, T., Marty, J. C., and Vaida, M. 2008. "Observation and Adaptation of a Learning Session Based on a Multi-Agent System: An Experiment," *4th International Conference on Intelligent Computer Communication and Processing*, pp. 17-24.

Manasrah, A., Masoud, M., and Jaradat, Y. 2021. "Short Videos, or Long Videos? A Study on the Ideal Video Length in Online Learning," *International Conference on Information Technology*, Amman, Jordan, pp. 366-370.

Mayer, R. E., and Moreno, R. 2003. "Nine Ways to Reduce Cognitive Load in Multimedia Learning," *Educational Psychologist* (38:1), pp. 43-52.

Mielicki, M. K., and Wiley, J. 2022. "Exploring the Necessary Conditions for Observing Interleaved Practice Benefits in Math Learning," *Learning and Instruction* (80), p. 101583.

Nesbit, J. C., and Adesope, O. O. 2006. "Learning with Concept and Knowledge Maps: A Meta-Analysis," *Review of Educational Research* (76:3), pp. 413-448.

O'donnell, A. M., Dansereau, D. F., and Hall, R. H. 2002. "Knowledge Maps as Scaffolds for Cognitive Processing," *Educational Psychology Review* (14:1), pp. 71-86.

Orru, G., and Longo, L. 2019. "The Evolution of Cognitive Load Theory and the Measurement of Its Intrinsic, Extraneous and Germane Loads: A Review," *Human Mental Workload: Models and Applications*, L. Longo and M.C. Leva (eds.), Cham: Springer International Publishing, pp. 23-48.

Pardos, Z., Bergner, Y., Seaton, D., and Pritchard, D. 2013. "Adapting Bayesian Knowledge Tracing to a Massive Open Online Course in Edx," *International Conference on Educational Data Mining*, Memphis, TN.

Park, O.-c., and Lee, J. 2003. "Adaptive Instructional Systems," *Educational Technology Research and Development* (25), pp. 651-684.

Pelánek, R. 2017. "Bayesian Knowledge Tracing, Logistic Models, and Beyond: An Overview of Learner Modeling Techniques," *User Modeling and User-Adapted Interaction* (27:3), pp. 313-350.

Piccoli, G., Ahmad, R., and Ives, B. 2001. "Web-Based Virtual Learning Environments: A Research Framework and a Preliminary Assessment of Effectiveness in Basic It Skills Training," *MIS Quarterly* (25:4), pp. 401-426.

Rau, M. A., Aleven, V., and Rummel, N. 2010. "Blocked Versus Interleaved Practice with Multiple Representations in an Intelligent Tutoring System for Fractions," *in Proceedings of the 10th International Conference on Intelligent Tutoring Systems*, Pittsburgh, PA, pp. 413-422.

Reddy, S., Labutov, I., and Joachims, T. 2016. "Latent Skill Embedding for Personalized Lesson Sequence Recommendation." working paper.

Reich, J., and Ruipérez-Valiente, J. A. 2019. "The Mooc Pivot," *Science* (363:6423), pp. 130-131.

Renkl, A. 2014. "Toward an Instructionally Oriented Theory of Example-Based Learning," *Cognitive Science* (38:1), pp. 1-37.

Rohrer, D. 2012. "Interleaving Helps Students Distinguish among Similar Concepts," *Educational Psychology Review* (24:3), pp. 355-367.

Rohrer, D., Dedrick, R. F., and Burgess, K. 2014. "The Benefit of Interleaved Mathematics Practice Is Not Limited to Superficially Similar Kinds of Problems," *Psychonomic Bulletin & Review* (21:5), pp. 1323-1330.

Rohrer, D., Dedrick, R. F., Hartwig, M. K., and Cheung, C.-N. 2020. "A Randomized Controlled Trial of Interleaved Mathematics Practice," *Journal of Educational Psychology* (112:1), pp. 40-52.

Rohrer, D., Dedrick, R. F., and Stershic, S. 2015. "Interleaved Practice Improves Mathematics Learning," *Journal of Educational Psychology* (107:3), pp. 900-908.

Sana, F., Yan, V. X., Kim, J. A., Bjork, E. L., and Bjork, R. A. 2018. "Does Working Memory Capacity Moderate the Interleaving Benefit?," *Journal of Applied Research in Memory and Cognition* (7:3), pp. 361-369.

Sandrone, S., Scott, G., Anderson, W. J., and Musunuru, K. 2021. "Active Learning-Based Stem Education for in-Person and Online Learning," *Cell* (184:6), pp. 1409-1414.

Santhanam, R., Sasidharan, S., and Webster, J. 2008. "Using Self-Regulatory Learning to Enhance E-Learning-Based Information Technology Training," *Information Systems Research* (19:1), pp. 26-47.

Shrivastav, H., and Hiltz, S. R. 2013. "Information Overload in Technology-Based Education: A Meta-Analysis," *Americas Conference on Information Systems*, Chicago, Illinois.

Singh, P. V., Tan, Y., and Youn, N. 2011. "A Hidden Markov Model of Developer Learning Dynamics in Open Source Software Projects," *Information Systems Research* (22:4), pp. 790-807.

Sweller, J. 2010. "Element Interactivity and Intrinsic, Extraneous, and Germane Cognitive Load," *Educational Psychology Review* (22:2), pp. 123-138.

Sweller, J. 2011. "Cognitive Load Theory," in *Psychology of Learning and Motivation*, J.P. Mestre and B.H. Ross (eds.). Cambridge, UK: Elsevier Academic Press, pp. 37-76.

Sweller, J., van Merriënboer, J. J., and Paas, F. 2019. "Cognitive Architecture and Instructional Design: 20 Years Later," *Educational Psychology Review* (31:2), pp. 261-292.

Tauber, S. K., Dunlosky, J., Rawson, K. A., Wahlheim, C. N., and Jacoby, L. L. 2013. "Self-Regulated Learning of a Natural Category: Do People Interleave or Block Exemplars During Study?," *Psychonomic Bulletin & Review* (20:2), pp. 356-363.

Taylor, K., and Rohrer, D. 2010. "The Effects of Interleaved Practice," *Applied Cognitive Psychology* (24:6), pp. 837-848.

Tseng, S., Sue, P., Su, J., Weng, J., and Tsai, W. 2007. "A New Approach for Constructing the Concept Map," *Computers & Education* (49:3), pp. 691-707.
Walls, J. G., Widmeyer, G. R., and El Sawy, O. A. 1992. "Building an Information System Design Theory for Vigilant Eis," *Information Systems Research* (3:1), pp. 36-59.
Wang, C. 2022. "Comprehensively Summarizing What Distracts Students from Online Learning: A Literature Review," *Human Behavior and Emerging Technologies* (2022), pp. 1-15.
Wilson, K., and Nichols, Z. 2015. "The Knewton Platform. A General-Purpose Adaptive Learning Infrastructure," Knewton White Paper.
Wilson, K. H., Karklin, Y., Han, B., and Ekanadham, C. 2016. "Back to the Basics: Bayesian Extensions of Irt Outperform Neural Networks for Proficiency Estimation," *International Conference on Educational Data Mining*, Raleigh, NC.
Yan, V. X., Bjork, E. L., and Bjork, R. A. 2016. "On the Difficulty of Mending Metacognitive Illusions: A Priori Theories, Fluency Effects, and Misattributions of the Interleaving Benefit," *Journal of Experimental Psychology: General* (145), pp. 918-933.
Yan, V. X., and Sana, F. 2021. "Does the Interleaving Effect Extend to Unrelated Concepts? Learners' Beliefs Versus Empirical Evidence," *Journal of Educational Psychology* (113:1), pp. 125-137.
Zhu, B., and Watts, S. A. 2010. "Visualization of Network Concepts: The Impact of Working Memory Capacity Differences," *Information Systems Research* (21:2), pp. 327-344.
Zulkiply, N., and Burt, J. S. 2013. "The Exemplar Interleaving Effect in Inductive Learning: Moderation by the Difficulty of Category Discriminations," *Memory & Cognition* (41:1), pp. 16-27.

APPENDIX A: ESTIMATION OF THE HIDDEN MARKOV MODEL

Covariate Specification

According to Figure 1 in the main text, the transition covariates ($\boldsymbol{W_{st}^{c}}$) include time-invariant *learner characteristics* and time-variant *learning history*. The former includes gender, age, and pre-experiment exam scores and the latter includes the number of correctly answered questions on the focal topic, the number of incorrectly answered questions on the focal topic, and the total number of questions answered for all topics.

The emission covariates ($\boldsymbol{Z_{st}^{c}}$) also include the same time-invariant *learner characteristics* and time-variant *learning-history* variables. The emission process can also be

affected by time-variant *learner's behavioral tendencies*. The latter includes the average duration spent on each question of a given topic and the standard deviation of durations spent on questions of this topic.

Maximizing the Likelihood of the Observed Learning Records Over Time

We estimate a learner's mastery level of different topics by maximizing the likelihood of the observed learning outcomes over time. The model specification and estimation are as follows.

For each specific topic *c*, we use X_{st}^c to represent the outcome of the *t*-th assessment (i.e., whether the *t*-th question on topic *c* is correctly answered) for learner *s*. We define $\boldsymbol{X_s^{c(T)}} = \left(X_{s1,}^c X_{s2}^c, \dots, X_{sT}^c\right)$, representing a learner *s*'s learning outcomes on topic *c* from assessment 1 to assessment *T*. We define $\boldsymbol{Y_s^{c(T)}} = \left(Y_{s1,}^c Y_{s2}^c, \dots, Y_{sT}^c\right)$ as a learner *s*'s hidden mastery level history on topic *c* from assessment 1 to assessment *T*. The transition covariate matrix ($\boldsymbol{W_s^{c(T)}}$) and emission covariate matrix ($\boldsymbol{Z_s^{c(T)}}$) are similarly defined. The likelihood of observing the learning outcome $\boldsymbol{X_s^{c(T)}}$ for learner *s* on topic *c* from assessment 1 to assessment *T* is $L_s^{c(T)} = P\left(\boldsymbol{X_s^{c(T)}}\right) = \sum_{\boldsymbol{Y_s^{c(T)}}} P(\boldsymbol{X_s^{c(T)}}, \boldsymbol{Y_s^{c(T)}})$. Following the Markov assumption, we can derive the likelihood as follows:

$$L_s^{c(T)} = \sum_{\boldsymbol{Y_s^{c(T)}}} P\left(\boldsymbol{X_s^{c(T)}} \middle| \boldsymbol{Y_s^{c(T)}}, \boldsymbol{Z_s^{c(T)}}\right) P(\boldsymbol{Y_s^{c(T)}} | \boldsymbol{Y_s^{c(T-1)}}, \boldsymbol{W_s^{c(T-1)}})$$

According to the dependencies in the Markov chain, we can decompose the $L_s^{c(T)}$ as the sum over all the paths.

$$L_s^{c(T)} = \sum_{Y_{s1,}^c Y_{s2}^c, \dots, Y_{sT}^c} P(Y_{s1}^c, W_{s1}^c) \prod_{t=2}^{T} P(X_{st}^c | Y_{st}^c, Z_{st}^c)\, P(Y_{st}^c | Y_{st-1}^c, W_{st-1}^c)$$

where $P(Y_{st}^c | Y_{st-1}^c, W_{st-1}^c)$ is the transition probability from the hidden state Y_{st-1}^c at assessment t-1 to the hidden state Y_{st}^c at assessment t, which is affected by the transition covariates (W_{st-1}^c), and $P(X_{st}^c | Y_{st}^c, Z_{st}^c)$ is the emission probability of hidden state Y_{st}^c at assessment t, which is affected by the emission covariates (Z_{st}^c). The joint likelihood of observing the learning outcomes for all learners and all topics from assessment 1 to assessment T is given by $L^{(T)} = \prod_S \prod_c L_s^{c(T)}$. We maximize the likelihood $L^{(T)}$ by choosing the value of each parameter.

During the experiment, we use the first two-week's learning records to train the model (~20,000 question-answer records) and apply the trained model to the remaining weeks. Following Singh et al. (2011) and Kim and Krishnan (2019), we choose the number of hidden states using the Bayesian information criterion, which indicates that the optimal number of hidden states is two, labeled as "unmastered" and "mastered", respectively.

APPENDIX B: KNOWLEDGE MAP UPDATING

We use the fuzzy association rules to discover topic dependencies and refine the knowledge map. The detailed knowledge map updating process is as follows. An assessment question answered by a learner is treated as one question record. Given that each question is mapped to one topic, we can transform a question record into a topic record. For a given topic, we pool the records for an individual learner and get a record set. For each record set, we calculate the proportion of incorrect answers as the *error rate* for this topic. For example, learner *s* answered three questions related to topic *A* and two were incorrectly answered. Thus, the error rate of topic *A* for learner *s* is $ErrorRate_s(A) = 2/3$. With *n* learners, we can calculate the *support* of *A* as the mean error rate of *A*:

$$Support(A) = \frac{1}{n}\sum_{s=1}^{n} ErrorRate_s(A).$$

Therefore, a higher support is interpreted as a higher error rate.

Extending the notion of support to a set of topics, we cannot use the Boolean logic operator because the error rates are numerical. Following Tseng et al. (2007), we calculate the support of a topic set using the fuzzy implication operator (FIO) *minimization*. Formally,

$$\text{Support}(A,B) = \frac{1}{n}\sum_{s=1}^{n} \min\left(ErrorRate_s(\text{A}), ErrorRate_s(\text{B})\right)$$

A high value of $Support(A,B)$ implies that learners frequently answer both topic-A and topic-B questions incorrectly.

Following the rule of Bayesian posterior, we can derive the confidence level of association ($A \rightarrow B$) as follows:

$$Confidence(\text{A} \rightarrow \text{B}) = \frac{support\ (A,B)}{support(\text{A})}$$

A high value of confidence (A → B) implies that a large portion of learners who incorrectly answered questions on topic A also incorrectly answered questions on topic B. With high support and confidence, we can consider A to be a dependency for B. To ensure that A plays a significant role in B's accuracy, we follow the literature to only keep the rules whose *lift*, measured as $\frac{Confidence\ (A \rightarrow B)}{support(B)}$, is larger than 1.

APPENDIX C: PSEUDO ALGORITHM OF THE SCHEDULING ENGINE

Pseudo Algorithm for Refining Topic List

 # This function modifies the weak topic list for a specific learner according to the knowledge map so that related weak topics are placed next to each other

Function **GetRefinedTopicList** (WeakTopicList, KM) → RefinedTopicList {

 # Inputs:

 # WeakTopicList: Store the focal learner's weak topics detected by HMM

 # KM: Stored the knowledge map updated by the fuzzy association rule

 # Returns:

 # RefinedTopicList: an ordered list of refined weak topics for the learner.

RefinedTopicList = [] ;

Sort WeakTopicList by the descending order of the probability of being unmastered;

for (i=0; i< WeakTopicList.length(); i++) {

```
        WeakTopic= WeakTopicList[i] ;
        if (WeakTopic not in RefinedTopicList ) {
            Append WeakTopic at the end of RefinedTopicList;
        }
        TopicDependencyList = look up all the topics that the WeakTopic depends on in KM;
        foreach TopicDependency in TopicDependencyList {
            if (TopicDependency in WeakTopicList) {
                insert TopicDependency into RefinedTopicList just before WeakTopic;
            }
        }
    }
    Return RefinedTopicList;
}
```

Pseudo Algorithm for Schedule Exercises

```
# This function schedules the exercises for the next learning session
Function ScheduleExercises () → RankedExerciseList{
# RankedExerciseList: a list of exercises for the focal learner in the next learning session
    RefinedTopicList = GetRefinedTopicList (WeakTopicList, KM) ;
    ExerciseList= look up all the exercises in the database that have not been assigned to the focal
    learner;
    foreach Exercise in ExerciseList {
        Exercise.TopicRankSum = 0 ;
        ExerciseWeakTopics = look up all the topics in the RefinedTopicList covered by the Exercise;
        foreach ExerciseWeakTopic in ExerciseWeakTopics {
            TopicRank = The position of ExerciseWeakTopic in the RefinedTopicList ;
            Exercise.TopicRankSum = Exercise.TopicRankSum + TopicRank ;
        }
        Exercise.NumberWeakTopics = ExerciseWeakTopics.length() ;
    }
    RankedExerciseList = sort ExerciseList by the descending order of NumberWeakTopics and the
    ascending order of TopicRankSum;
    Return the top N Exercises in the RankedExerciseList ;
}
```

An Example of Related-interleaving Across Sessions

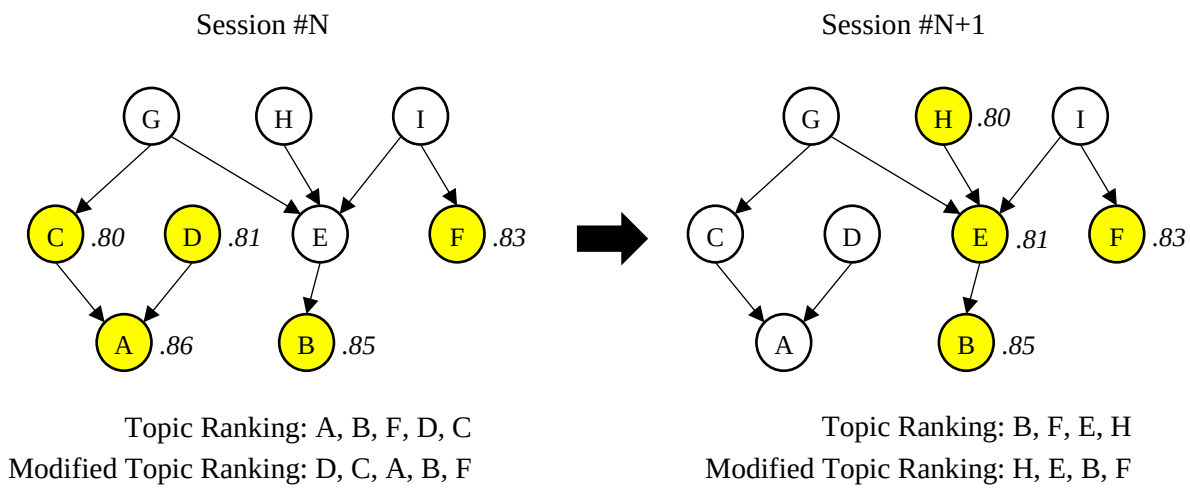


Figure C1: An Example of Related-interleaving Across Sessions

Notes*:* As an illustrative example, the knowledge map models the relatedness among 9 topics (A to I). We use shaded circles to indicate a learner's unmastered topics and white circles to indicate the mastered topics at the moment. The number beside an unmastered topic, as derived by the HMM, is the probability that the learner has *not* mastered it. Based on the modified ranking of weak topics in Session #N (D, C, A, B, F), we assign the learner exercises that cover *three* unmastered topics that are directly *connected* in the knowledge map (C, D, A). In Session #N+1, the system detects that the learner has mastered the practiced topics but becomes unfamiliar with E and H as time goes by. Consequently, the personalized list of weak topics becomes B, E, F, and H. According to the updated weak topics in modified topic ranking, we assign the learner exercises that cover *three* unmastered topics that are directly *connected* in the knowledge map (H, E, B) for Session #N+1.

APPENDIX D: SYSTEM DESCRIPTION

When a learner logs in the system, he/she can click the "personalized exercise" module and find all the personalized exercises assigned to him/her, including both the finished and unfinished ones, and the associated assignment dates and due dates, as shown in Figure D1.

	Student name	Class ID	Status	Avaliable from	Due by	Submit time
Select	*****, Chen	Grade 8, Class 3	Unfinished	2017-08-28	2017-08-29	
Select	*****, Chen	Grade 8, Class 3	Unfinished	2017-08-28	2017-08-29	
Select	*****, Chen	Grade 8, Class 3	Finished	2017-08-26	2017-08-27	2017-08-27 20:17:17
Select	*****, Chen	Grade 8, Class 3	Finished	2017-08-26	2017-08-27	2017-08-26 09:14:25
Select	*****, Chen	Grade 8, Class 3	Unfinished	2017-08-24	2017-08-25	
Select	*****, Chen	Grade 8, Class 3	Finished	2017-08-24	2017-08-25	2017-08-24 21:41:17
Select	*****, Chen	Grade 8, Class 3	Finished	2017-08-22	2017-08-23	2017-08-23 16:40:12
Select	*****, Chen	Grade 8, Class 3	Finished	2017-08-22	2017-08-23	2017-08-22 16:28:06

Figure D1. Demo of a Learner-Specific List of Exercises

When a learner clicks one exercise, the system displays an article and the corresponding assessment questions. After the learner submits the answers, the system automatically grades the answers and provides real-time feedback, including the correct answers and the reasoning processes, as illustrated in Figure D2. The learner can find all the learning records, along with the feedback, in the "learning records" section.


Answers and reasoning processes of this article

Article 38

Imagine you are walking to school when suddenly you notice a wallet in the street. After picking it up, you realize that it's full of $10, $20, and $50 bills. They add up to $500! What would you do?

Would you hand the wallet over to your headmaster as soon as you arrive at school? That's just what KemoyGourzang did.

"I have lost money before," Kemoy, a 10-year-old boy, told the reporter, "I knew if I had lost my wallet, I would have wanted it back."

Joy-Ann Morgan, the headmaster, immediately called the owner of the wallet, using the ID inside. The man did not realize he had lost it. Morgan said he was happy to learn that an honest kid had found it.

"This is what we want to teach our students," Morgan explained.

Kemoy has learned that doing the right thing pays off. His school district gave him a "Good Citizen" prize in honor of what he did. A person who has good citizenship is responsible and does the right thing. Kemoy has also received gifts from complete strangers, money, gift cards, and even a pair of sports shoes.

As for the owner of the wallet, he thanked Kemoy when he went to the school to get his wallet back. He also gave the student $100 as a reward. "He told me I was the most honest person he had ever met," Kemoy said, "It makes me really happy."

1 **How much money was there in the wallet?**

A:10
B:20
C:50
D:500
Correct answer:D
Your answer: D
Reference:They add up to $500!

2 **What did Kemoy do after he picked the wallet up?**

A: He handed it over to the headmaster.
B:He told his friends about the money.
C:He called the owner of the wallet.
D:He asked a reporter for help.
Correct answer:A
Your answer: A
Reference:Would you hand the wallet over to your headmaster as soon as you arrive at school? That's just what KemoyGourzang did.

3 **Kemoy received _____ for what he did from the school district.**

A:a pair of sports shoes
B:a gift card
C:a "Good Citizen" prize
D:a $100 bill
Correct answer:C
Your answer:B
Reference:His school district gave him a "Good Citizen" prize in honor of what he did.

Figure D2. Snapshot of the Answer Page

APPENDIX E: RANDOMIZATION AND MANIPULATION CHECK

To ensure that the subjects were randomly assigned without significant differences in their prior performance across the three experimental groups (i.e., non-interleaving, unrelated-interleaving, and related-interleaving), we compare the pre-experiment exam scores (*FinalScore*) across the three groups.

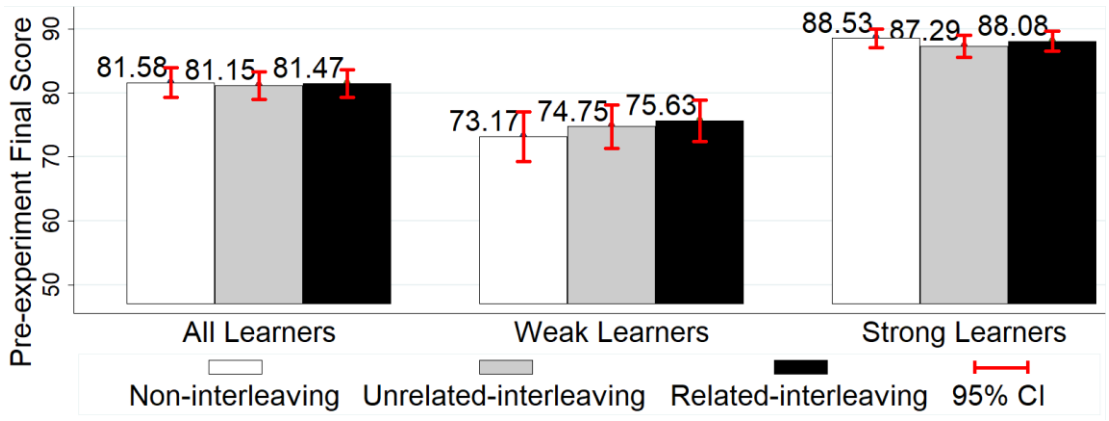


Figure E1. Randomization Check: Pre-experiment Exam Score Comparison

The results show no significant difference (Figure E1, all learners). Furthermore, the average scores of weak and strong learners in the three groups are not significantly different (see Figure E1, Weak Learners and Strong Learners). In addition, the numbers of weak and strong learners are also equally distributed across the three groups.

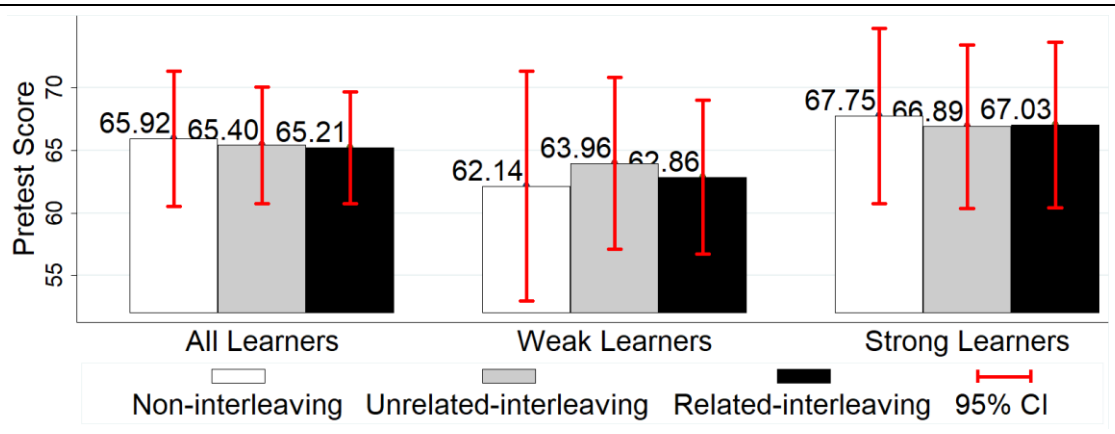


Figure E2. Randomization Check: Pretest Score Comparison

We also compare the pretest scores (*Pretest*) across the three groups. The results show no significant difference (Figure E2, all learners). Furthermore, the average scores of weak and strong learners in the three groups are not significantly different (see Figure E2, Weak Learners and Strong Learners). Then we compare the gender (*Female*) distribution across the three groups and find no significant difference (see Figure E3).

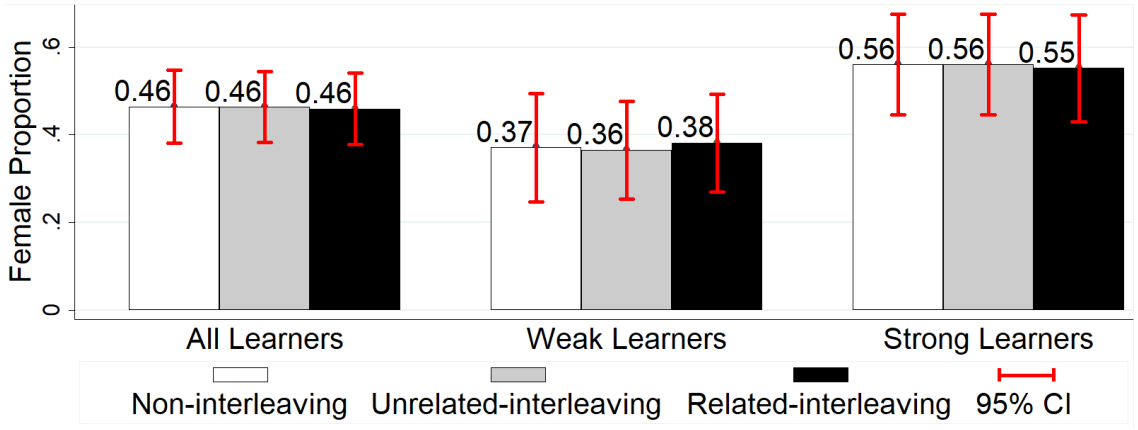


Figure E3. Randomization Check: Learner Gender

Finally, we conduct the randomization checks in each of the 17 different classes. Table E1 reports the p-value of the pairwise comparison across the three groups by using the subsamples of each class. Table E1 also reports the p-value of ANOVA analysis for the differences between three groups in each of the 17 classes. Panel A, Panel B, and Panel C compare the pre-experiment final score, pretest score, and gender, respectively. In general, 16 out of 17 classes show no significant difference across the three groups in terms of the pre-experiment final score, pretest score, and gender. One exception is Class 15. Class 15 shows no significant difference in terms of the pre-experiment final score and gender, but a marginal significance in terms of the pretest score across the three groups. To resolve this issue, we conduct robustness checks by removing all the students in Class 15 and find consistent results. Thus, we deem randomization successful.

Table E1: The p-value of the pairwise comparison across three experimental groups in each of the 17 classes

ClassID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Panel A: Pre-experiment final score																	
Unrelated-Interleaving - Non-Interleaving	0.96	0.55	0.66	0.84	0.77	0.92	0.81	0.26	0.42	0.95	0.66	0.89	0.95	0.72	0.70	0.91	0.75
Related-Interleaving - Unrelated-Interleaving	0.63	0.81	0.58	0.28	0.50	0.90	0.61	0.42	0.75	0.13	0.79	0.65	0.51	0.98	0.44	0.90	0.91
Related-Interleaving - Non-Interleaving	0.63	0.69	0.88	0.21	0.77	0.83	0.75	0.71	0.57	0.16	0.50	0.75	0.52	0.72	0.69	0.83	0.84
Three Group Comparison	0.85	0.83	0.85	0.40	0.79	0.98	0.87	0.50	0.70	0.27	0.78	0.89	0.75	0.92	0.74	0.98	0.95
Panel B: Pretest Score																	
Unrelated-Interleaving - Non-Interleaving	0.86	0.49	0.56	0.52	0.63	0.55	0.70	0.50	0.70	0.67	0.51	0.74	0.68	0.59	0.85	0.85	0.88
Related-Interleaving - Unrelated-Interleaving	0.30	0.75	0.72	0.56	0.23	0.50	0.73	0.16	0.35	0.36	0.24	0.74	0.41	0.26	0.05	0.61	0.76
Related-Interleaving - Non-Interleaving	0.14	0.28	0.39	1.00	0.31	0.87	0.94	0.47	0.67	0.22	0.50	0.53	0.62	0.50	0.09	0.72	0.86
Three Group Comparison	0.30	0.54	0.68	0.77	0.40	0.74	0.91	0.36	0.64	0.46	0.48	0.81	0.71	0.52	0.11	0.87	0.95
Panel C: Gender Distribution																	
Unrelated-Interleaving - Non-Interleaving	1.00	1.00	0.87	0.42	0.78	0.25	0.92	0.36	0.60	1.00	0.43	0.29	0.33	0.85	1.00	1.00	0.88
Related-Interleaving - Unrelated-Interleaving	0.70	0.85	0.85	0.42	0.70	1.00	0.95	0.20	1.00	1.00	0.62	0.79	0.69	0.85	0.55	1.00	0.28
Related-Interleaving - Non-Interleaving	0.73	0.85	0.70	0.97	0.54	0.25	0.97	0.68	0.57	1.00	0.75	0.48	0.56	1.00	0.52	1.00	0.31
Three Group Comparison	0.91	0.98	0.93	0.64	0.82	0.41	0.99	0.42	0.82	1.00	0.72	0.55	0.62	0.97	0.76	1.00	0.48

To check the manipulation in terms of whether the topics were successfully interleaved, we retrieve learners' exercise records during the experiment. Each learner in the non-interleaved group completed 34.3 exercises and each exercise covered 1.16 topics on average, indicating that the learners were supplied with exercises covering only one unmastered topic most of the time. Each learner in the unrelated-interleaving group completed 34.9 exercises, with each exercise covering 3.42 topics on average. Each learner in the related-interleaved group completed 35.6 exercises, with each exercise covering 3.38 topics on average. As depicted in Figure E4, the learners in the unrelated-interleaving and related-interleaving groups received exercises covering significantly more topics than those in the non-interleaving group ($p < 0.001$). This pattern holds for both strong learners and weak learners. Hence, the manipulation of interleaving is considered successful.

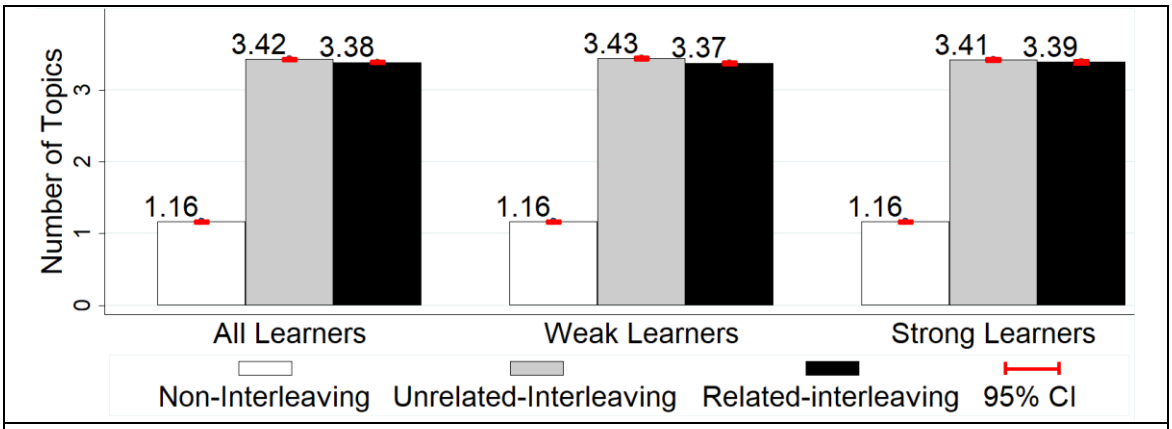


Figure E4. Manipulation Check: Average Number of Topics Covered per Exercise

We also check the manipulation of topic relatedness by comparing the number of topic dependencies covered by each exercise between the unrelated-interleaving group and the related-interleaving group. Learners in the unrelated-interleaving group received exercises covering 0.47 topic dependencies on average, whereas learners in the related-interleaved group received exercises covering 2.29 topic dependencies on average. As depicted in Figure E5, the learners in the related-interleaving group received exercises with significantly more topic relationships ($p < 0.001$). This pattern also holds for both strong learners and weak learners. Hence, the manipulation of topic relatedness is considered successful.

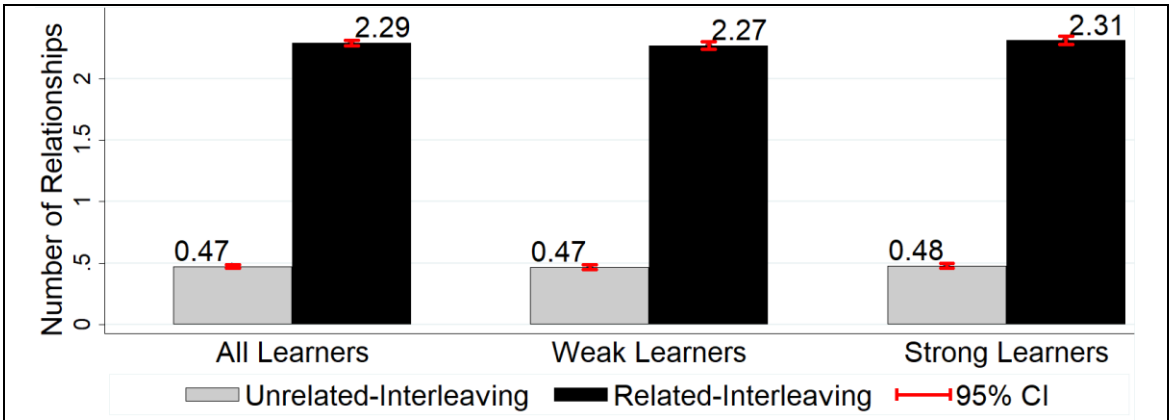


Figure E5. Manipulation Check: Average Number of Topic Dependencies in Each Exercise

APPENDIX F: EXPLORING THE EFFECTS OF RELATED-INTERLEAVING ON COGNITIVE LOAD

To provide further evidence of whether our findings can be explained by cognitive load theory, we explore whether related-interleaving induces more cognitive load on schema building during learning sessions than unrelated-interleaving and non-interleaving. Although we do not directly observe learners' schema-building load, the literature suggests a few indirect indicators of schema-building load, as detailed below.

First, CLT suggests that when learners engage more cognitive resources in schema building, they achieve a higher level of knowledge acquisition (Ausubel et al. 1968). Accordingly, knowledge acquisition scores are the "most common method of investigating cognitive load" (Brunken et al. 2003; Orru and Longo 2019). In our study, because we use the HMM to estimate the learner's *topic mastery* after each learning session, we can leverage HMM estimates of *topic mastery* as a measure of knowledge acquisition scores. Second, similar to knowledge acquisition scores, researchers have also used learners' *accuracy* of doing exercises during a learning session as an indirect measure of their schema-building load (Martin 2014; Orru and Longo 2019). We, therefore, obtain learners' accuracy in answering questions in each learning session.

Table F1: Effect of Related-interleaving on Topic Mastery and Session Accuracy

	Topic Mastery	Session Accuracy
	(1)	(2)
Interleaving	-0.004	0.006
	(0.00)	(0.01)
Interleaving×*Relatedness*	0.043***	0.026***
	(0.00)	(0.01)
FinalScore	0.005***	0.006***
	(0.00)	(0.00)
Pretest	0.005***	0.002***
	(0.00)	(0.00)
Female	0.050***	0.067***
	(0.00)	(0.01)
Constant	-1.176***	-0.044
	(0.05)	(0.10)
Class fixed effect	YES	YES
Date fixed effect	YES	YES
Topic fixed effect	YES	NO
N	79772	11269
R^2	0.229	0.196

Notes: *Interleaving* represents the effect of unrelated-interleaving relative to non-interleaving. *Relatedness* is meaningful in the interleaving condition, but not in the non-interleaving condition. Therefore, *Interleaving*×*Relatedness* is equivalent to *Relatedness* and captures the effect of related-interleaving relative to unrelated-interleaving. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors are in parentheses.

We first compare learners' *mastery* of each topic in each learning session across three experimental groups (see Table F1, column 1), controlling for the pre-experiment final exam score (*FinalScore*), the pretest score (*Pretest*), and gender (*Female*). We also include the class-fixed effect to incorporate class-level heterogeneity, the date-fixed effect to control for time heterogeneity, and the topic-fixed effect to incorporate topic-level heterogeneity. Then, we compare learners' *accuracy* during the learning session across three experimental groups (see Table F1, Column 2).

Figure F1 Column 1 shows that learners under the related-interleaving condition have 3.8% and 4.3% higher probabilities of mastering the topics learned in the session, compared with learners under the non-interleaving and unrelated-interleaving conditions, respectively. Column 2 shows that learners under the related-interleaving condition had 3.1% and 2.6%

increases in session accuracy compared with learners under the non-interleaving and unrelated-interleaving conditions, respectively. Overall, evidence based on the two indirect measures indicates that related-interleaving leads to increased schema-building load during learning sessions, which is consistent with the CLT predictions.

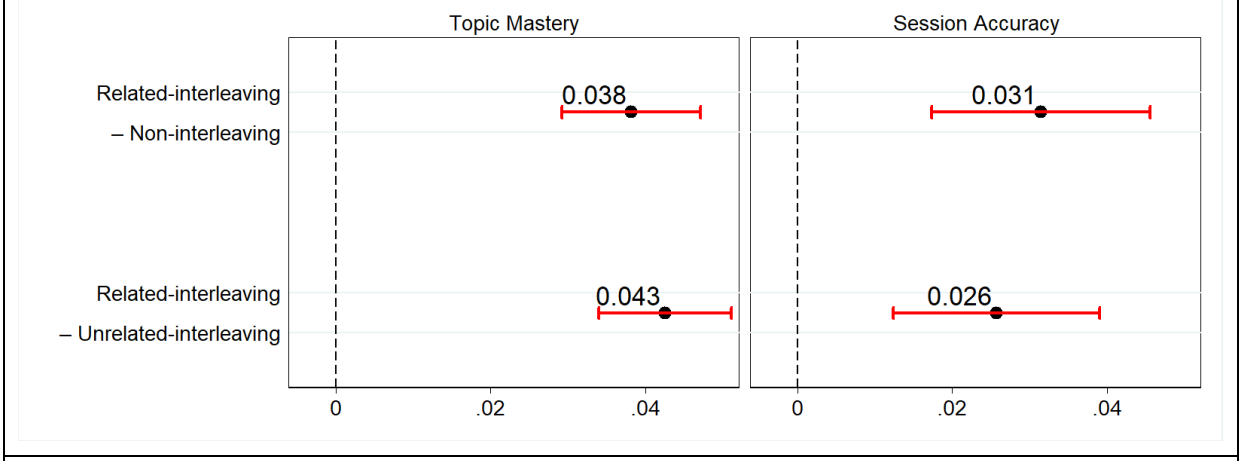


Figure F1. Test the Schema Building by Topic Mastery and Session Accuracy
Notes: A dot represents the mean value, and the error bar represents the 95% confidence interval.

APPENDIX G: SENSITIVITY ANALYSIS

In the above analyses, we categorize the learners into strong or weak based on a median split of their pre-experiment exam scores. We conduct a sensitivity analysis by varying the definition of strong learners from the top 40% to the top 60% in the class based on their final exam scores. We repeat the analyses and report the results in Table G1. Our results show that the three-way interaction (*Interleaving*×*Relatedness*×*StrongLearner*) is consistently negative. Thus, our findings are not sensitive to the selection of the cutting point for the division. Furthermore, the effect of relatedness for weak learners (*Interleaving*×*Relatedness*) increases from 14.56 to 24.32 when strong learners are defined from the top 40% to the top 60% of the class. In other words, the positive effect of topic relatedness is more pronounced for learners in the remaining 40% of the class. For these learners, increasing topic relatedness is more beneficial because it helps reduce the basic processing load caused by interleaved learning.

Table G1: Sensitivity Analyses of the Heterogeneous Effect for Different Types of Learners

	40%	45%	50%	55%	60%
	(1)	(2)	(3)	(4)	(5)
Interleaving	-15.92***	-16.02**	-19.86***	-21.31***	-28.17***
	(4.59)	(4.98)	(5.38)	(5.88)	(6.19)
Interleaving×Relatedness	14.56***	17.42***	18.81***	21.40***	24.32***
	(4.01)	(4.28)	(4.48)	(4.63)	(4.87)
Interleaving×StrongLearner	25.25***	21.57**	25.18***	26.60***	34.59***
	(6.98)	(6.96)	(6.98)	(7.24)	(7.41)
Interleaving×Relatedness×StrongLearner	-11.59+	-15.50*	-15.86*	-19.22**	-22.58***
	(6.94)	(6.74)	(6.56)	(6.45)	(6.42)
StrongLearner	-6.87	-4.42	-7.13	-4.16	-8.71
	(5.04)	(5.16)	(5.35)	(5.72)	(5.88)
Pretest	0.35***	0.35***	0.36***	0.36***	0.38***
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)
Female	7.01*	6.83*	6.84*	6.81*	6.76*
	(2.81)	(2.84)	(2.82)	(2.76)	(2.71)
Constant	58.55***	56.23***	58.46***	56.28***	58.32***
	(7.28)	(7.36)	(7.48)	(7.57)	(7.63)
Class fixed effect	YES	YES	YES	YES	YES
N	306	306	306	306	306
R^2	0.343	0.335	0.343	0.362	0.384

Notes: $^{+}$ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors are in parentheses.

REFERENCE

Ausubel, D. P., Novak, J. D., and Hanesian, H. 1968. *Educational Psychology: A Cognitive View*. New York: Holt, Rinehart and Winston.

Brunken, R., Plass, J. L., and Leutner, D. 2003. "Direct Measurement of Cognitive Load in Multimedia Learning," *Educational Psychologist* (38:1), pp. 53-61.

Kim, Y., and Krishnan, R. 2019. "The Dynamics of Online Consumers' Response to Price Promotion," *Information Systems Research* (30:1), pp. 175-190.

Martin, S. 2014. "Measuring Cognitive Load and Cognition: Metrics for Technology-Enhanced Learning," *Educational Research and Evaluation* (20:7-8), pp. 592-621.

Orru, G., and Longo, L. 2019. "The Evolution of Cognitive Load Theory and the Measurement of Its Intrinsic, Extraneous and Germane Loads: A Review," *Human Mental Workload: Models and Applications,* L. Longo and M.C. Leva (eds.), Cham: Springer International Publishing, pp. 23-48.

Singh, P. V., Tan, Y., and Youn, N. 2011. "A Hidden Markov Model of Developer Learning Dynamics in Open Source Software Projects," *Information Systems Research* (22:4), pp. 790-807.

Tseng, S., Sue, P., Su, J., Weng, J., and Tsai, W. 2007. "A New Approach for Constructing the Concept Map," *Computers & Education* (49:3), pp. 691-707.