The Half-Life of MOOC Knowledge
A Randomized Trial Evaluating the Testing Effect in MOOCs

Davis, Dan; Kizilcec, René F.; Hauff, Claudia; Houben, Geert-Jan

DOI
10.1145/3170358.3170383

Publication date
2018

Document Version
Final published version

Published in
LAK'18 Proceedings of the 8th International Conference on Learning Analytics and Knowledge

The Half-Life of MOOC Knowledge

A Randomized Trial Evaluating Knowledge Retention and Retrieval Practice in MOOCs

Dan Davis*
Delft University of Technology
Delft, the Netherlands
d.j.davis@tudelft.nl

René F. Kizilcec
Stanford University
Stanford, CA, USA
kizilcec@stanford.edu

Claudia Hauff
Delft University of Technology
Delft, the Netherlands
c.hauff@tudelft.nl

Geert-Jan Houben
Delft University of Technology
Delft, the Netherlands
g.j.p.m.houben@tudelft.nl

ABSTRACT

Retrieval practice has been established in the learning sciences as one of the most effective strategies to facilitate robust learning in traditional classroom contexts. The cognitive theory underpinning the "testing effect" states that actively recalling information is more effective than passively revisiting materials for storing information in long-term memory. We document the design, deployment, and evaluation of an `Adaptive Retrieval Practice System` (`ARPS`) in a MOOC. This push-based system leverages the testing effect to promote learner engagement and achievement by intelligently delivering quiz questions from prior course units to learners throughout the course. We conducted an experiment in which learners were randomized to receive `ARPS` in a MOOC to track their performance and behavior compared to a control group. In contrast to prior literature, we find no significant effect of retrieval practice in this MOOC environment. In the treatment condition, passing learners engaged more with `ARPS` but exhibited similar levels of knowledge retention as non-passing learners.

CCS CONCEPTS

• **Applied computing** → *Education*;

KEYWORDS

Retrieval Practice, Testing Effect, Experiment, Knowledge Retention

ACM Reference Format:
Dan Davis, René F. Kizilcec, Claudia Hauff, and Geert-Jan Houben. 2018. The Half-Life of MOOC Knowledge: A Randomized Trial Evaluating Knowledge Retention and Retrieval Practice in MOOCs. In *LAK'18: International Conference on Learning Analytics and Knowledge, March 7–9, 2018, Sydney, NSW, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3170358.3170383

*Research supported by the *Leiden-Delft-Erasmus Centre for Education and Learning.*

LAK'18, March 7–9, 2018, Sydney, NSW, Australia

ACM ISBN 978-1-4503-6400-3/18/03…$15.00
https://doi.org/10.1145/3170358.3170383

1 INTRODUCTION

Retrieval practice is one of the most effective and well-established strategies to facilitate robust learning. Also known as the testing effect, retrieval practice is the process of reinforcing prior knowledge by actively and repeatedly recalling relevant information. This strategy is more effective in facilitating robust learning (the committing of information to long-term memory [20]) than passively revisiting the same information, for example by going over notes or book chapters [1, 6, 13, 15, 16, 21, 24].

Given the wealth of scientific evidence on the benefits of retrieval practice (cf. Section 2) and the adaptability of digital learning platforms, in this paper we explore to what extent the testing effect holds in one of today's most popular digital learning settings: Massive Open Online Courses (MOOCs). Research into both MOOC platforms and MOOC learners' behavior has found learners to take a distinctly linear trajectory [8, 12, 26] through course content. Many learners take the path of least resistance towards earning a passing grade [28] which does not involve any back-tracking or revisiting of previous course units—counter to a regularly-spaced retrieval practice routine.

Although contemporary MOOC platforms are not designed to encourage retrieval practice, prior work suggests that MOOC learners with high Self-Regulated Learning (SRL) skills tend to engage in retrieval practice of their own volition [18]. These learners strategically seek out previous course materials to hone and maintain their new skills and knowledge. However, these learners are the exception, not the norm. The vast majority of MOOC learners are not disciplined, self-directed autodidacts who engage in such effective learning behavior without additional support. This motivated us to create the `Adaptive Retrieval Practice System` (`ARPS`), a tool that encourages retrieval practice by *automatically* and *intelligently* delivering quiz questions from previously studied course units to learners. The system is automatic in that the questions appear without any required action from the learner and intelligent in that questions are adaptively selected based on a learner's current progress in the course. We deployed `ARPS` in an edX MOOC (`GeoscienceX`) in a randomized controlled trial with more than 500 learners assigned to either a treatment (`ARPS`) or a control group (no `ARPS`).

Based on the data we collect in this randomized trial, we investigate the benefits of retrieval practice in MOOCs guided by the following research questions:

RQ1 How does an adaptive retrieval practice intervention affect learners' academic achievement, course engagement, and self-regulation compared to generic recommendations of effective study strategies?

RQ2 How does a push-based retrieval practice intervention (requiring learners to act) change learners' retrieval practice behavior?

In addition to collecting behavioral and performance data inside of the course, we invited learners to complete a survey two weeks after the course had ended. This self-report data enabled us to address the following research question:

RQ3 To what extent is robust learning facilitated in a MOOC?

The primary contributions of our study show that (i) retrieval practice, in contrast to substantial prior work, may not benefit learners in a MOOC (RQ1); (ii) passing and non-passing learners who receive ARPS do not differ in their knowledge levels (as measured by ARPS) but rather in their course engagement levels (RQ2); and (iii) passing and non-passing learners do not differ in long-term knowledge retention (RQ3).

2 RELATED WORK

We now review prior research in the areas of retrieval practice, spaced vs. massed practice, and long-term knowledge retention to inform the study design. Based on the literature, we develop several research hypotheses to be evaluated in the study.

2.1 Retrieval Practice

Adesope et al. [1] conducted the most recent meta-analysis of retrieval practice. They evaluated the efficacy of retrieval practice compared to other learning strategies such as re-reading or re-watching, the impact of different problem types in retrieval practice, the mediating role of feedback, experimental context, and students' education level.

The effect of retrieval practice is strong enough overall for the authors to recommend that frequent, low-stakes quizzes be integrated into learning environments so that learners can assess knowledge gaps and seek improvement [1]. They also found that multiple choice problems not only require low levels of cognitive effort, they were the most effective type of retrieval practice problem in terms of learning outcomes compared to short answer questions. And while certainly a boon to learners (the majority of studies in the review endorse its effectiveness), feedback is actually not required or integral to effective retrieval practice. From studies that did incorporate feedback, the authors found that delayed feedback is more effective in lab studies, whereas immediate feedback is best in classroom settings. Of the 217 experiments (from the 118 articles included in the meta-analysis), 11% took place in traditional classroom settings as part of the curriculum, with the vast majority taking place in laboratory settings.

Roediger and Butler [24] also offer a synthesis of published findings on retrieval practice. From the studies reviewed, the authors offer five key points on retrieval practice for promoting long-term knowledge: (i) retrieval practice is superior to reading for long-term retention, (ii) repeated testing is more effective than a single test, (iii) providing feedback is ideal but not required, (iv) benefits are greatest when there is lag time between learning and practicing/retrieving, and (v) retrieval practice increases the likelihood of learning transfer—the application of learned knowledge in a new context [24].

Consistent with the findings from [1, 13, 24], Johnson and Mayer [14] evaluated the effectiveness of retrieval practice in a digital learning environment focused on lecture videos. In the study, learners who answered test questions after lecture videos—pertaining to topics covered in the videos—outperformed learners who merely re-watched the video lectures in terms of both long-term knowledge retention and learning transfer [14].

2.2 Spaced vs. Massed Practice

The literature on spaced versus massed practice has shown that a higher quantity of short, regularly-spaced study sessions is more effective than a few long, massed sessions [6]. There is considerable overlap in the research on retrieval practice and that on spaced versus massed practice. As outlined in the studies above, an optimal study strategy is one of a regularly spaced retrieval practice routine [5, 6, 22].

Spaced versus massed practice has been evaluated in the MOOC setting by Miyamoto et al. [22], who analyzed learners' log data and found that learners who tend to practice effective spacing without guidance or intervention are more likely to pass the course relative to those learners who do not engage in spacing. We leveraged these insights from the learning sciences in the design of ARPS.

2.3 Expected Knowledge Retention

Scientific evaluation of the human long-term memory began at the end of the 19th century, leading to the earliest model of human memory loss/maintenance: the Ebbinghaus curve of forgetting [11]. The curve begins at time 0 with 100% knowledge uptake with a steep drop-off in the first 60 minutes to nine hours, followed by a small drop from nine hours to 31 days.

Custers [7] conducted a rigorous review of long-term retention research and found considerable evidence in support of the Ebbinghaus curve in terms of shape—large losses in short-term retention (from days to weeks) which level off for longer intervals (months to years)—but not always in terms of scale. The result of their meta-analysis shows that university students typically lose one third of their knowledge after one year, even among the highest-achieving students.

Considering the effect on retrieval practice on long-term retention, Lindsey et al. [21] conducted a similar study to the present research in a traditional classroom setting and found that their personalized, regularly spaced retrieval practice routine led to higher scores on a cumulative exam immediately after the course as well as a cumulative exam administered one month after the course. In their control condition (massed study practice), learners scored just over 50% on the exam, whereas those exposed to the retrieval practice system scored 60% on average. For the control group, this marked an 18.1% forgetting rate, compared to 15.7% for those with retrieval practice. They also found that the positive effect of retrieval practice was amplified with the passing of time.

Duolingo, a popular language learning platform with hundreds of thousands of daily users, has developed their own forgetting

curve to model the "half-life" of knowledge—theirs operates on a much smaller time scale, with a 0% probability of remembering after seven days. Based on the retrieval practice and spacing effect literature, they also developed a support system to improve learners' memory. Findings show that their support system, tuned to the "half-life regression model" of a learner's knowledge, significantly improves learners' memory [25].

It is worth noting, however, that forgetting is viewed as an adaptive behavior: forgetting liberates the memory of outdated, unused information to free up space for new, immediately relevant memories and knowledge [23]. Retrieval works adjacently to this phenomenon in that by regularly reactivating and revisiting knowledge, the brain does not tag it as unused and forgettable, but rather recognizes its relevance and, accordingly, stores it in long-term memory.

Based on the existing literature in retrieval practice, spaced versus massed practice, and knowledge retention over time, we arrive at the following hypotheses to test in a MOOC setting with regard to the GeoscienceX course:

- **H1** Push-based interventions will lead to higher levels of retrieval practice than static interventions.
- **H2** Learners who are exposed (i.e. learners in the treatment group) to ARPS will show higher rates of course completion, engagement, and self-regulation than those who are not.
- **H3** Learners will retain approximately two thirds of newly-learned knowledge from the course over the long term.

3 ADAPTIVE RETRIEVAL PRACTICE SYSTEM

The Adaptive Retrieval Practice System (ARPS) is a client-server application (written in JavaScript/node.js)[1] that provides *automated*, *scalable* and *personalized* retrieval practice questions to MOOC learners on a continuous basis. We developed ARPS specifically for use within the edX platform in taking advantage of the RAW HTML input affordance. This allows course teams/instructors to build custom interfaces within the platform that render along with the standard edX content (such as videos, quizzes, etc.).

The ARPS back-end keeps track of the content a MOOC learner has already been exposed to through client-side sensor code that logs a learner's progress through the course and transmits it to the back-end. Once the back-end receives a request from the ARPS front-end (a piece of JavaScript running in a learner's edX environment on pages designated to show retrieval practice questions), it determines which question to deliver to a learner at a given time based on that learner's previous behavior in the course by randomly selecting from a personalized pool of questions only pertaining to content the learner has already been exposed to. Each question is pushed to the learner in the form of a qCard, an example of which is shown in Figure 3. These qCards appear to the learner as a pop-up within the browser window. We log all qCard interactions—whether it was ignored or attempted, the correctness of the attempt, and the duration of the interaction.

In contrast to previous interventions in MOOCs [9, 10, 17, 19, 27], we push questions to learners instead of requiring the learner to seek the questions out. We adopted this push-based design in order to allow learners to readily engage with the intervention with minimal interruption to the course experience. This design also addresses the issue of treatment noncompliance that has arisen in past research [9, 10]. ARPS is seamlessly integrated in the course, requiring as few additional interactions as possible. In the case of Multiple Choice (MC) questions (example problem text in Figure 1), the entire interaction requires just a single click: the learner selects their chosen response and if correct, receives positive feedback (a ✓ mark accompanied by encouraging text), and the qCard disappears. Incorrect responses invoke negative feedback (a **x** symbol alongside text encouraging the learner to make another attempt) which disappears after 4 seconds and returns the learner to the original question so they can try the problem again.

> A body with a low density, surrounded by material with a higher density, will move upwards due to buoyancy (negative density difference). We analyze the situation of a basaltic magma generated at a depth of 10 km and surrounded by gabbroic rocks. Will the magma move downward, remain where it is or move upward?

Figure 1: Example of an easy (less than 5% of incorrect responses) Multiple Choice question in GeoscienceX.

> Suppose an earthquake occurred at a depth of 10 kilometers from the surface that released enough energy for a P-wave to travel through the center of the Earth to the other side. This is for the sake of the exercise, because in reality sound waves tend to travel along the boundaries and not directly through the Earth as depicted. Assume the indicated pathway and the given thicknesses and velocities. How many seconds does it take for the seismic P-wave to reach the observatory on the other side of the Earth?

Figure 2: Example of a difficult (5% correct response rate) Numerical Input question in GeoscienceX.

We also enabled one other question type[2] to appear in qCards: Numeric Input (NI) problems (an example is shown in Figure 2). These problems require the learner to calculate a solution and enter the answer in a text box. While requiring more effort than a single click response, we included this problem type to allow for a comparison between the two.

4 STUDY DESIGN

We now describe the MOOC we deployed ARPS in as well as the design of our empirical study.

4.1 Participants

A total of 2,324 learners enrolled in the course titled *Geoscience: the Earth and its Resources* (or GeoscienceX), which was offered on the edX.org platform between May 23, 2017 and July 26, 2017. The course consists of 56 lecture videos and 217 graded quiz questions. Of the 132 total problems from the 217 in the course question bank deemed suitable for use with qCards (multi-step problems were excluded so that each qCard could be answered independently), 112 were Multiple Choice and 20 were Numerical Input problems.

[1] The code is available at https://github.com/dan7davis/Lambda.

[2] Additional question types that are supported by the edX platform can easily be added to ARPS; in this paper we focus exclusively on MC and NI questions as those are the most common question types in the MOOC we deployed ARPS in.

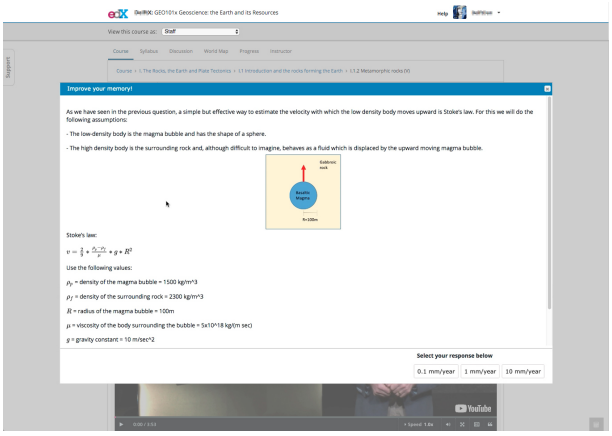

Figure 3: Example `qCard` in the `GeoscienceX` course. The main body of the `qCard` contains the question text, and the bar at the bottom contains the MC answer buttons. The grey "x" at the top right corner closes the window and dismisses the problem.

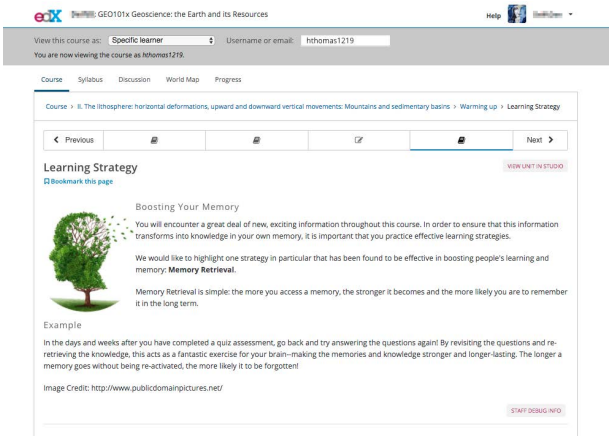

Figure 4: Page shown to learners in the control condition at the beginning of each course week describing how to practice an effective memory retrieval routine.

Based on self-reported demographic information (available for 1,962 learners), 35% of participants were women and the median age was 27. This course drew learners from a wide range of educational backgrounds: 24% held at least a high school diploma, 7% an Associate's degree, 42% a Bachelor's degree, 24% a Master's degree, and 3% a PhD. Learners were not provided any incentive beyond earning a course certificate for participating in the study.

We define the study sample as the 1,047 learners who entered the course at least once (out of the 2,324 who initially enrolled): 524 assigned to the control condition and 523 to the treatment condition.

A post-course survey & quiz (cf. Section 4.2) was sent to all 102 learners who engaged with the ARPS system (9 complete survey responses—8.8% response rate) and the 150 highest performing learners in the control condition in terms of final grade (11 complete responses—7.3%).

4.2 Procedure

This study was designed as a randomized controlled trial in the `GeoscienceX` course. Upon enrolling in the course, learners were randomly assigned to one of two conditions for the duration of the course:

- **Control condition**: A lesson on effective study habits was added to the weekly introduction section. The lesson explained the benefits of retrieval practice and offered an example of how to apply it (Figure 4).
- **Treatment condition**: ARPS was added to the course to deliver quiz questions (via a `qCard`) from past weeks. The same weekly lesson on study habits as in the control condition was provided to help learners understand the value of the tool. In addition, information on how the adaptive retrieval system works and that responses to the `qCard` do not count towards learners' final grade was provided. The `qCards` were delivered before each of the 49 course lecture videos (from Weeks 2–6) across the six course weeks. A button at the bottom of each lecture video page enabled learners to receive a new `qCard` on demand after the initial one to keep practicing.

To assess how well learners retained their knowledge from the course, we sent a post-course survey to the most active learners in the course (in terms of time spent in the platform) two weeks after the course had ended. The survey contained a random selection of ten assessment questions from the `GeoscienceX` course. Learners in the treatment condition additionally received eight questions about their experience with ARPS. We evaluated the results of this post-course assessment with respect to differences between the two cohorts in long-term knowledge retention.

4.3 Measures

In order to measure and compare the behavior of learners in both the control and treatment conditions, we consider the following measures of in-course events (tracked and logged on the edX platform):

- **Final grade** (a score between 0 and 100);
- **Course completion** (binary indicator: pass, no-pass);
- Course activities:
  - Video interactions (play, pause, fast-forward, rewind, scrub);
  - Quiz submissions (number of submissions, correctness);
  - Discussion forum posts;
  - Duration of time in course;
- ARPS interactions:
  - Duration of total `qCard` appearance;
  - Response submissions (with correctness);
  - `qCard` interactions (respond, close window).

The following data were collected in the post-course survey:

- Course survey data
  - **Post-Exam Quiz Score** (between 0-10);
  - Learner intentions (e.g., to complete or just audit);
  - Prior education level (highest degree achieved).

We have selected the three bolded variables as our *primary outcome variables* for this study for the following reasons: (i) a learner's

Table 1: Course outcomes in the control and treatment group.

Condition	Subset	N	Non-Zero Grade	Passing Rate	Grade Quantiles
Control	All	524	31%	8%	[0, 0, 2]
Treatment	All	523	34%	7%	[0, 0, 2]
Treatment	Complier	102	76%	34%	[2, 19, 74]
Treatment	Noncomplier	421	23%	0.2%	[0, 0, 0]

final grade is the best available indicator of their performance in the course in terms of their short-term mastery of the materials and (ii) the Post-Exam Quiz score measures how well learners retained the knowledge weeks after finishing the course.

5 RESULTS

This section presents the findings from each of the five analyses we conducted: (i) estimating the causal effect of the intervention based on the randomized controlled experiment (RQ1), (ii) examining how learners interacted with ARPS (RQ2), (iii) modeling how learners' knowledge changed over time (RQ3), (iv) estimating the rate of learners' long-term knowledge retention (RQ3), and (v) understanding learners' experience with ARPS from a qualitative angle using survey responses. Each subsection concludes with a statement synthesizing its key finding.

5.1 Effect of Encouraging Retrieval Practice

The goal of the randomized experiment is to estimate the *causal effect* of retrieval practice (RQ1). By comparing learners in the control and treatment group, we can estimate the effect of the encouragement to engage in retrieval practice with ARPS. However, many learners who were encouraged did not actually engage in retrieval practice, which is a form of treatment noncompliance. Specifically, of the 523 learners assigned to the treatment, only 102 interacted at least once with a qCard (i.e. complied with the treatment). For this reason, in order to estimate the effect of retrieval practice itself, we also analyze the experiment as an encouragement design.[3]

The primary outcome measure is the final course grade, which determines certificate eligibility (the passing threshold is 60%). Table 1 contains summary statistics for grade and certification outcomes in the control group and the treatment group, overall and separately for treatment compliers and noncompliers. First, we estimate the Intent-to-treat Effect (ITT), which is the difference in average outcomes between the treatment and control groups. We find that the ITT is not significant for certification (log odds ratio= $-0.215, z = -0.920, p = 0.357$), getting a non-zero grade (logOR= $0.143, z = 1.08, p = 0.280$), and the continuous grade itself (Kruskal-Wallis $\chi^2_{df=1} = 0.592, p = 0.442$).

[3]The study was pre-registered at www.osf.io/4py2h. Due to the small sample size and compliance rate, we adjusted our analytic approach. Specifically, we analyze the experiment as an encouragement design beyond estimating average treatment effects, and we did not apply the specified sample exclusion criteria because they could inadvertently bias the causal inference.

Next, we use an instrumental variable approach (Two-stage Least Squares) to estimate the effect of retrieval practice for those who used it (i.e. a Local Average Treatment Effect, or LATE) [2]. For a binary instrument Z, outcome Y, and compliance indicator G, we can compute the Wald estimator:

$$\beta^{IV} = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(G|Z=1) - E(G|Z=0)}$$

The LATE is not significant either for certification ($\beta^{IV} = -0.078, z = -0.893, p = 0.371$), getting a non-zero grade ($\beta^{IV} = 0.160, z = 1.11, p = 0.267$), and the continuous grade itself ($\beta^{IV} = -0.066, z = -0.889, p = 0.374$).

Finally, we estimate the per-protocol effect, which is the difference in average outcomes between treatment compliers and control compliers (i.e. the entire control group). We find large differences in terms of certification (logOR= $1.74, z = 6.66, p < 0.001$), getting a non-zero grade (logOR= $2.00, z = 7.94, p < 0.001$), and the continuous grade itself (Kruskal-Wallis $\chi^2_{df=1} = 99, p < 0.001$). However, the per-protocol estimates do not have a causal interpretation because different subpopulations are compared: all learners in the control group versus those highly motivated learners who comply in the treatment group. For instance, note that treatment compliance is strongly correlated with receiving a higher grade (Spearman's $r = 0.56, p < 0.001$).

In addition to estimating effects based on the final course grade, the pre-registration also specifies a number of process-level analyses (RQ2). In particular, we hypothesized that learners who receive the treatment would exhibit increased self-regulatory behavior in terms of (i) revisiting previous course content such as lecture videos, (ii) self-monitoring by checking their personal progress page, and (iii) generally persisting longer in the course. No evidence in support of the hypothesized behavior was found, neither in terms of the ITT (Kruskal-Wallis $\chi^2_{df=1}$s$< 0.68, p$s> 0.41) nor in terms of the LATE ($|z|$s$< 0.98, p$s> 0.32). Focusing on learners in the treatment group, we also hypothesized that learners who attempt qCards at a higher rate would learn more and score higher on regular course assessments, which is supported by the data (Spearman's $r = 0.42, p < 0.001$). In summary (and in contrast to previous studies on the topic [1, 6, 13, 15, 16, 21, 24]):

> The causal analysis yields no evidence that ARPS raised learning, performance, or self-regulatory outcomes in this course.

This may be due to the low sample size or rate of compliance in this study. We also observed a selection effect into using ARPS among highly motivated learners in the treatment group. Among those learners, increased engagement with qCards was associated with higher grades, though this pattern could be due to self-selection (e.g., more committed learners both attempt more qCards and put more effort into course assessments). To better understand how different groups of learners used ARPS and performed on subsequent learning assessments, we conducted a series of exploratory analyses.

5.2 Engaging with Retrieval Cues

5.2.1 Question-by-Question Analysis. Figure 5 illustrates learners' responses for every question delivered by ARPS, which indicates

Table 2: Summary statistics for the mean value of the measures listed in Section 4.3 for analyses including all learners in both conditions who logged at least one session in the platform. The vertical line separates standard course behavior measures and those collected by ARPS.

Group	N=	Final Grade	Passing Rate	Video Interactions	Quiz Submissions	Forum Posts	Time in Course	Time with qCards	qCards Seen
Control	524	9	8%	6.52	34.77	0.26	4h47m	–	–
Treatment	523	8	7%	5.83	30.88	0.29	3h40m	23m9s	7.71

which questions learners tended to struggle with (or ignore). The figure reveals that the choice to attempt or ignore a qCard is strongly associated with a learner's eventual passing or failing of the course. Moreover, it shows a steady decrease in learner engagement over time, not only among non-passing learners, but also among those who earned a certificate. Thus, attrition in MOOCs is not limited to those who do not pass the course; even the highest-achieving learners show a tendency of slowing down after the first week or two (also observed in [28]).

From Figures 6 and 7, we observe that passing and non-passing learners do not appear to differ in their rate of giving incorrect responses (which would indicate misconceptions or a lack of understanding the materials). Instead, they differ in their choice to ignore the problems all together. When removing the instances of ignored qCards and focusing only on attempted problems (right-hand side of Table 3), we observe a significant albeit small difference (6% difference, $\chi^2 = 9.63$, $p = 0.002$) between the proportion of correct or incorrect responses between passing and non-passing learners (cf. Table 3). In other words, passing and non-passing learners both perform about the same on these quiz problems—and yet, with no discernible difference in their assessed knowledge, only some go on to earn a passing grade and course certificate.

Table 3: qCard problem response (left) and correctness (right). Significant differences at the $p < 0.001$ (between passing and non-passing) are indicated with †.

	Attempted†	Ignored†	Correct	Incorrect
Non-passing	0.47	0.53	0.76	0.24
Passing	0.73	0.27	0.82	0.18

5.2.2 First Question Response. To further explore the predictive power of a learner's choice to either attempt or ignore the qCards, we next analyzed each learner's first interaction with a qCard. Figure 8 (left) shows the passing rate of learners segmented according to their first interaction with a qCard. Learners who attempted rather than ignored the first qCard had a 47% chance of passing the course. In contrast, learners who ignored the first qCard delivered to them only had a 14% chance of passing. Figure 8 (right) additionally illustrates the relationship between the result of the first qCard attempt and passing the course. There were notably few learners who responded incorrectly, but their chance of passing the course was still relatively high at 33% compared to those who simply ignored the qCard.

To evaluate whether the response of a learner's second qCard problem adds any predictive value, we replicated the analysis shown in Figure 8 for the responses to the first two qCards delivered to each learner. No difference in predictive value was observed by considering the second consecutive response—learners who answered their first two consecutive qCards correctly had a 53% chance of earning a passing grade.

From these analyses we conclude that initial adoption of ARPS appears to depend partly on learners' motivation to complete the course.

5.2.3 Response Duration. We next explore how much time learners spent interacting with qCards and how time spent predicts the outcome of the interaction. Figure 9 shows the proportion of correct, incorrect, and ignored responses as a function of time elapsed with a visible qCard. We find that the decision to ignore the qCard happened very quickly, with a median duration of 7 seconds (from the time the qCard appeared to the time the learner clicked the "x" button to close it). For situations where learners did attempt to answer the question, the amount of time they spent did not have any association with the correctness of their response; the median duration for correct and incorrect responses was 18 seconds and 16 seconds, respectively.

From the question-by-question, first question response, and response duration analyses, we conclude:

> There is no significant difference in assessed knowledge between passing and non-passing learners; the key difference lies in a learner's willingness to engage with the retrieval practice questions.

5.3 Modeling Knowledge Over Time

One of the contributions of ARPS is the data set that it generates: by tracking learners' responses to these periodic, formative, and ungraded questions throughout the entire course, we have a longitudinal account of learners' evolving knowledge state throughout the entire process of instruction. In this section we explore how learners' knowledge (as measured by performance with the qCards) deteriorates over time (RQ3).

Figure 10 shows the cumulative week-by-week performance of both passing and non-passing learners. As qCards could only be delivered with questions coming from *prior* course weeks, the x-axis begins with Week 2, where only questions from Week 1 were delivered. This continues up to Week 6 where questions from Weeks 1–5 could be delivered.

The left (Passing) graph in Figure 10 illustrates the forgetting curve of the passing learners in GeoscienceX. We observe a statistically significant decrease in performance between Weeks 2 and 6 (correct response rate dropping from 67% to 49% respectively;

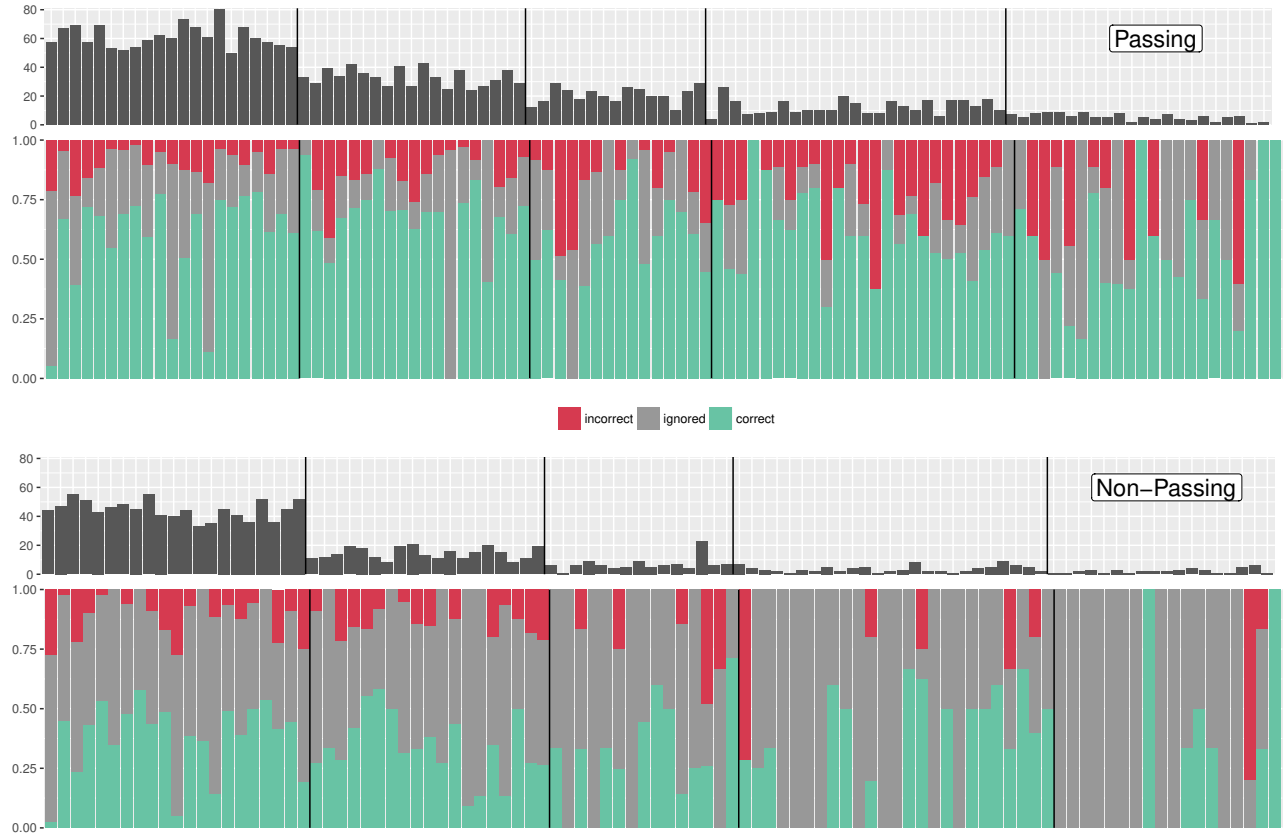

Figure 5: A question-by-question breakdown of every learner interaction with the `qCards`. The top two figures represent the behavior of passing learners—the upper image shows the number of learners being served that question, and the lower shows how they interacted with it—and the bottom two show that of non-passing learners. Questions are shown in order of appearance in the course (left to right), and the solid vertical line indicates a change in course week. Best viewed in color.

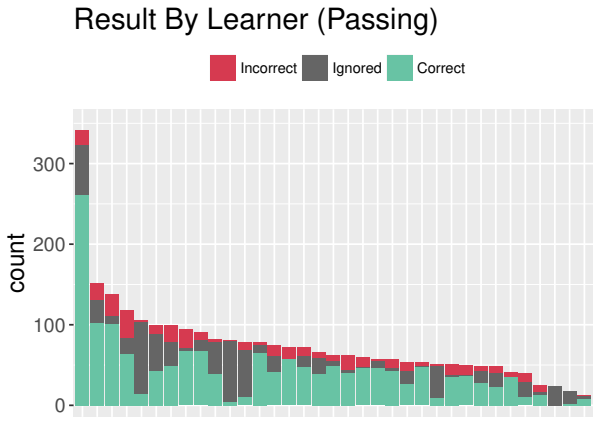

Figure 6: Each bar corresponds to one passing learner. Only one learner took advantage of the "infinite quizzing" capability by frequently using the "Generate new `qCard`" button. Best viewed in color.

$\chi^2 = 32.8, p < 0.001$). While the proportion of ignored responses remains steadily low, the proportion of correct responses drops by 18% (nearly identical to the forgetting rate found in [21]). The rate of incorrect responses increased from 4% to 25% ($\chi^2 = 87.8$, $p < 0.001$).

On the right (Non-Passing) graph in Figure 10 we observe that the choice to ignore `qCards` was common through the entire course duration, with a slight increase in the later weeks. We also observe a significant decrease in correct response rates for non-passing learners ($\chi^2 = 15.7, p < 0.001$). However, unlike passing learners who exhibited a significant increase in incorrect responses, there is no significant change for non-passing learners. The change, instead, is in the rate of ignored responses, increases from 47% in Week 2 to 61% in Week 6.

We identify two main contributing factors to this decline in performance over time. First, the amount of assessed content increases each week; in Week 6 there are five course weeks worth of content to be assessed, whereas in Week 2 there is only content from Week 1 being assessed. Second, people simply forget more with the passing of time [23]; each passing week moves the learner temporally farther away from when the content was initially learned.

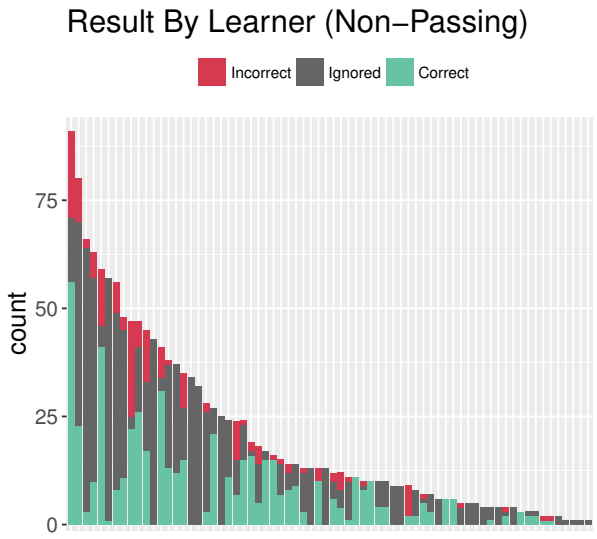

Figure 7: Each bar corresponds to one non-passing learner. Best viewed in color.

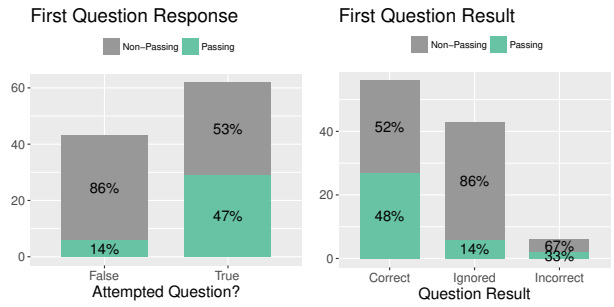

Figure 8: The likelihood of course completion based on learners' response (left) and result (right) to the first `qCard` they were shown. "True" indicates both correct or incorrect responses, and "False" indicates the `qCard` was ignored. Best viewed in color.

We next explore the relationship between testing delay and learners' memory and performance on `qCards`. In Figure 11, the x-axis represents the difference between a learner's current week and the week from which the `qCard` came. For example, if a learner was currently watching a lecture video in Week 5 and the `qCard` delivered was a question from Week 2, that would be a difference of three. While Figure 10 shows how the amount of content covered/assessed is related to performance, Figure 11 illustrates how the testing delay is related to performance.

We observe very similar trends as above for both passing and non-passing learners. For passing learners there is a 23% drop in correct response rates from 1 Week Elapsed to 5 Weeks Elapsed (65% to 42%, $\chi^2 = 23.6, p < 0.001$). Also significant is the 13% increase in incorrect response rate (8% to 21%, $\chi^2 = 17.5, p < 0.001$). The increase in ignored question frequency is not significant for passing learners, though it is large and significant for non-passing learners: between 1 Week Elapsed and 5 Weeks Elapsed, ignored questions

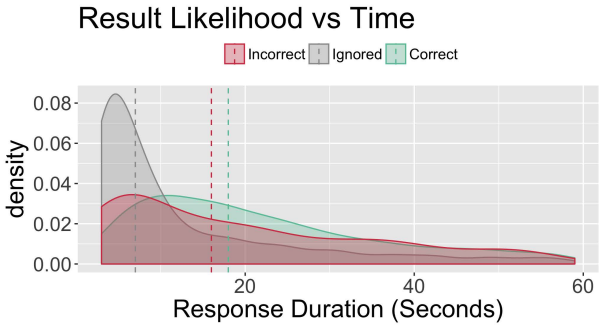

Figure 9: Kernel Density Estimation plot showing the relationship between time elapsed with the `qCard` visible and the result of a learner's response. The median time for each result is indicated with the dashed vertical line. Best viewed in color.

increased by 22% from 50% to 72% ($\chi^2 = 4.9, p = 0.025$). Overall, for non-passing learners, we observe increased ignoring, decreased correct problem attempt rates, and steady incorrect problem attempt rates.

This pattern shows that non-passing learners are able to recognize, attempt, and correctly answer `qCard` problems that are more proximate to their current stage in the course. This suggests a high level of self-efficacy especially among the non-passing learners; they are able to identify questions that they likely do not know the answer to and choose to ignore them.

Another encouraging finding from this analysis is that of learners' short-term knowledge retention. As partially illustrated by Figure 11, considering problems that were attempted with 1 Week Elapsed, passing learners answer 88% of problems correctly. Non-passing learners also show good performance with 79% correct (note that the required passing grade for the course was 60%).

From the above findings on learner knowledge as a function of both time and course advancement, we conclude:

> Learner quiz performance deteriorates with the introduction of more course concepts/materials and the passing of time.

5.4 Long-Term Knowledge retention

Long-term knowledge retention is the primary learning outcome affected by highly-spaced retrieval practice, which is typically evaluated in either a final, cumulative exam in a course, or a post-exam with some lag time between learners' exposure to the material and assessment [7, 21]. As the `GeoscienceX` course only featured weekly quizzes, we took a random selection of ten quiz questions from throughout the six end-of-week quizzes and created a post-course knowledge assessment. Delivered to learners in a survey format two weeks after the course had ended, we compared the performance between the two experimental conditions.

A χ^2 test revealed no significant difference in long-term knowledge retention between the control condition and learners in the treatment condition who interacted with the intervention at least once (RQ3). The mean score for learners in the control and treatment conditions was 6.2 ($SD = 1.9$) and 6.6 ($SD = 1.8$), respectively, out of a possible 10 points ($N = 20, t(17.6) = -0.45, p = 0.66$).

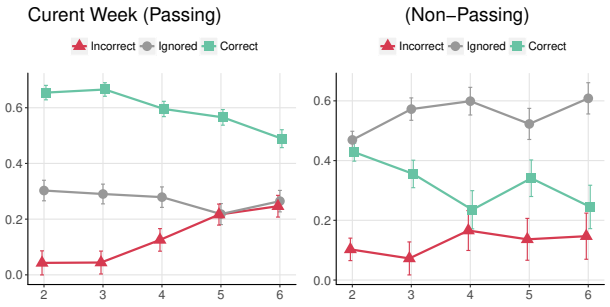

Figure 10: Week-by-week results of learners' interaction with the qCards. The x-axis represents the course week w, and the y-axis represents the proportion (%) of correct, incorrect, or ignored responses (with qCards showing queries from course weeks 1 to $w-1$). Error bars show standard error. Best viewed in color.

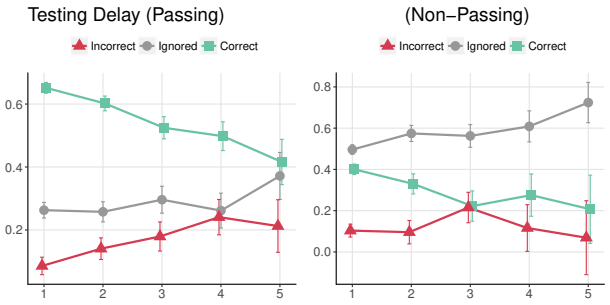

Figure 11: The x-axis represents the number of elapsed weeks between the course week where the topic was introduced and the course week in which the qCard was delivered (testing delay), and the y-axis represents the proportion of each response. Error bars show standard error. Best viewed in color.

Results from these analyses are consistent with prior literature [7, 21] on long-term knowledge retention in finding that, regardless of experimental condition and whether or not a learner passed the course:

> Approximately two thirds of course knowledge is retained over the long-term.

5.5 Learner Experience

To evaluate learners' experience with ARPS, we adapted the System Usability Survey [4] for the context of the present research. The scale was included in the post-course survey and learners indicated a cumulative usability score of 73.9 (SD = 12.2) on the SUS scale. According to [3], this is categorized as "acceptable usability" corresponding to a "C" grade. This means that the system's usability falls into the third quartile of SUS scores overall [3]—especially positive given that this was deployed not as a production system but as a research prototype.

To gain deeper insight into learners' experience and find out which specific aspects of the system could be improved, we also offered learners the opportunity to describe their experience with ARPS in two open response questions. One prompted them to share which aspects of ARPS they found to be the most enjoyable and another asked about frustrating aspects of ARPS.

One learner explained how the type of problem delivered was a key factor in their use of ARPS:

> "It [would] be better if only conceptual questions [were] asked for [the] pop quiz, it's troublesome if calculation is required. If calculation is required, I would prefer that the options are equations so that we can choose the right equation without evaluating them."

Other learners reported similar sentiments and also shared insights that indicate a heightened level of self-awareness induced by the qCards. Learners shared their perspectives talking about how the system helped *"...remind me [of] things that I missed in the course"* and how it gave them *"the chance to see what I remembered and what I had learned."* These anecdotes are encouraging as for these learners the system was able to encourage a deliberate activation of previously-learned concepts which may have otherwise been forgotten.

Upon seeing the learner feedback about how the problem type affected the learner's experience, we conducted a follow-up analysis to see if there was any indication that other learners felt the same way (as expressed through their interaction with ARPS). Figure 12 reveals that, indeed, this learner was not alone in their sentiment; we find that there was a 69% likelihood of learners attempting a MC qCard problem type compared to 41% attempt rate for NI problems. A χ^2 test shows this difference to be statistically significant ($p < 0.001$). Given that the question type (mostly evaluations of mathematical equations) is consistent across both problem types (MC and NI), we can conclude that these differences are indeed an effect of the problem type. This finding supports our initial design decision for a hyper-efficient interaction process—learners are far more likely to attempt a problem which only requires a single click selecting from a list of answers than one that requires two extra processes: they must first generate an answer from scratch and then type it out. From the data we are unable to identify which of these two extra processes contributes more to the problems being ignored, so we consider them in tandem.

6 CONCLUSION

Decades of prior research on the effectiveness of different learning strategies has found retrieval practice to be effective at supporting long-term knowledge retention [1, 6, 7, 13, 15, 16, 21, 21, 24]. However, how to effectively support retrieval practice in digital learning environments has not yet been thoroughly examined. The vast majority of prior work was conducted in offline learning environments, including university laboratory settings. Critically evaluating the efficacy of retrieval practice in digital learning environments promises to advance theory by developing a deeper understanding of how retrieval practice can be effective in a digital context as well as in a highly heterogeneous population that is embodied by MOOC learners.

In the current research we evaluated an Adaptive Retrieval Practice System in a MOOC to address the emerging issue of supporting learning strategies at large scale and to bridge retrieval

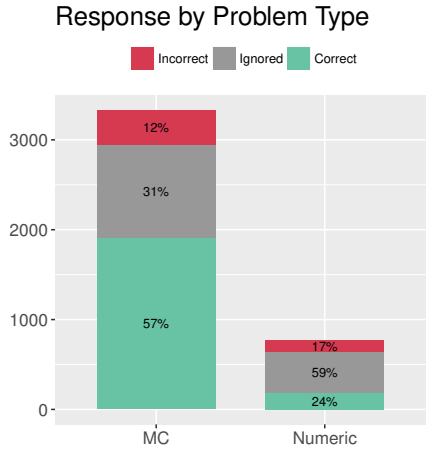

Figure 12: Breakdown of qCard interaction results across the two problem types. Best viewed in color.

practice theory into the digital learning space. We found noncompliance to be a major limitation in our evaluation of the system and its effectiveness. Many learners did not engage with the intervention, which limits our ability to draw causal inferences about the effect of retrieval practice on learners' achievement and engagement in the course.

We acknowledge the following limitations of the present study: (i) the qCards could potentially act as a distraction and make a learner more inclined to disengage, and (ii) despite the course being designed by trained course developers, there is a possibility that the assessment items used may not effectively measure the psychometric properties of learning, which would threaten the validity of our claim that retrieval practice does not improve learning outcomes.

Despite the lack of causal findings, the data collected from ARPS allowed us to offer multiple insights into the online learning process as it pertains to the persistence and transience of knowledge gains. By examining learner behavior and engagement with the intervention, we were able to track their performance on the same problem or topic and observe how their performance is affected by both the passage of time and introduction of new course materials.

We observed an encouraging trend of learners showing high levels of short- and medium-term knowledge retention, which is indicative of the early stages of learning. To what extent this newly gained knowledge is integrated into long-term memory warrants further research in the context of large online courses. Despite the null results from our causal analyses (Section 5.1), the wealth of evidence showing that retrieval practice is one of the most effective strategies to support knowledge retention makes this approach ripe for further investigation in online learning settings. The key to developing this theory further, however, is to design systems and interfaces that foster high levels of engagement to collect more causal evidence.

REFERENCES

[1] Olusola O Adesope, Dominic A Trevisan, and Narayankripa Sundararajan. 2017. Rethinking the Use of Tests: A Meta-Analysis of Practice Testing. *Review of Educational Research* (Feb. 2017).

[2] Joshua D Angrist, Guido W Imbens, and Donald B Rubin. 1996. Identification of causal effects using instrumental variables. *Journal of the American statistical Association* 91, 434 (1996), 444–455.

[3] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies* 4, 3 (2009), 114–123.

[4] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[5] Nicholas J Cepeda, Edward Vul, Doug Rohrer, John T Wixted, and Harold Pashler. 2008. Spacing Effects in Learning: A Temporal Ridgeline of Optimal Retention. *Psychological Science* 19, 11 (Nov. 2008), 1095–1102.

[6] Ruth C Clark and Richard E Mayer. 2016. *E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning.* John Wiley & Sons.

[7] Eugène JFM Custers. 2010. Long-term retention of basic science knowledge: a review study. *Advances in Health Sciences Education* 15, 1 (2010), 109–128.

[8] Dan Davis, Guanliang Chen, Claudia Hauff, and Geert-Jan Houben. [n. d.]. Gauging MOOC Learners' Adherence to the Designed Learning Path. In *Proceedings of the 9th International Conference on Educational Data Mining*. International Educational Data Mining Society (IEDMS), 54–61.

[9] Dan Davis, Guanliang Chen, Tim Van der Zee, Claudia Hauff, and Geert-Jan Houben. 2016. Retrieval practice and study planning in moocs: Exploring classroom-based self-regulated learning strategies at scale. In *European Conference on Technology Enhanced Learning*. Springer, 57–71.

[10] Dan Davis, Ioana Jivet, René F Kizilcec, Guanliang Chen, Claudia Hauff, and Geert-Jan Houben. 2017. Follow the successful crowd: raising MOOC completion rates through social comparison at scale. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. ACM, 454–463.

[11] Hermann Ebbinghaus. 1885. *Über das gedächtnis: untersuchungen zur experimentellen psychologie.* Duncker & Humblot.

[12] Chase Geigle and ChengXiang Zhai. 2017. Modeling Student Behavior With Two-Layer Hidden Markov Models. *JEDM-Journal of Educational Data Mining* 9, 1 (2017), 1–24.

[13] III Henry L Roediger and Jeffrey D Karpicke. 2016. The Power of Testing Memory: Basic Research and Implications for Educational Practice. *Perspectives on Psychological Science* 1, 3 (June 2016), 181–210.

[14] Cheryl I Johnson and Richard E Mayer. 2009. A testing effect with multimedia learning. *Journal of Educational Psychology* 101, 3 (2009), 621.

[15] Jeffrey D Karpicke and Janell R Blunt. 2011. Retrieval practice produces more learning than elaborative studying with concept mapping. *Science* 331, 6018 (2011), 772–775.

[16] Jeffrey D Karpicke and Henry L Roediger. 2008. The critical importance of retrieval for learning. *science* 319, 5865 (2008), 966–968.

[17] René F Kizilcec and Geoffrey L Cohen. 2017. Eight-minute self-regulation intervention raises educational attainment at scale in individualist but not collectivist cultures. *Proceedings of the National Academy of Sciences* (2017), 201611898.

[18] René F Kizilcec, Mar Pérez-Sanagustín, and Jorge J Maldonado. 2017. Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses. *Computers & education* 104 (2017), 18–33.

[19] René F Kizilcec, Andrew J Saltarelli, Justin Reich, and Geoffrey L Cohen. 2017. Closing global achievement gaps in MOOCs. *Science* 355, 6322 (2017), 251–252.

[20] Kenneth R Koedinger, Albert T Corbett, and Charles Perfetti. 2012. The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science* 36, 5 (2012), 757–798.

[21] Robert V Lindsey, Jeffery D Shroyer, Harold Pashler, and Michael C Mozer. 2014. Improving studentsâĂŹ long-term knowledge retention through personalized review. *Psychological science* 25, 3 (2014), 639–647.

[22] Yohsuke R Miyamoto, Cody A Coleman, Joseph Jay Williams, Jacob Whitehill, Sergiy Nesterko, and Justin Reich. 2015. Beyond time-on-task: The relationship between spaced study and certification in MOOCs. *Journal of Learning Analytics* 2, 2 (Dec. 2015), 47–69.

[23] Blake A Richards and Paul W Frankland. 2017. The Persistence and Transience of Memory. *Neuron* 94, 6 (2017), 1071–1084.

[24] Henry L Roediger and Andrew C Butler. 2011. The critical role of retrieval practice in long-term retention. *Trends in cognitive sciences* 15, 1 (2011), 20–27.

[25] Matthew Streeter. 2015. Mixture Modeling of Individual Learning Curves. *International Educational Data Mining Society* (2015).

[26] Miaomiao Wen and Carolyn Penstein Rosé. 2014. Identifying latent study habits by mining learner behavior patterns in massive open online courses. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 1983–1986.

[27] Michael Yeomans and Justin Reich. 2017. Planning prompts increase and forecast course completion in massive open online courses.. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. 464–473.

[28] Yue Zhao, Dan Davis, Guanliang Chen, Christoph Lofi, Claudia Hauff, and Geert-Jan Houben. 2017. Certificate Achievement Unlocked: How Does MOOC Learners' Behaviour Change?. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*. ACM, 83–88.