

## Scaling Effective Learning Strategies

### Retrieval Practice and Long-Term Knowledge Retention in MOOCs

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

**DOI**

[10.18608/jla.2018.53.3](https://doi.org/10.18608/jla.2018.53.3)

**Publication date**

2018

**Document Version**

Final published version

**Published in**

Journal of Learning Analytics

**Citation (APA)**

Davis, D., Kizilcec, R. F., Hauff, C., & Houben, G.-J. (2018). Scaling Effective Learning Strategies: Retrieval Practice and Long-Term Knowledge Retention in MOOCs. *Journal of Learning Analytics*, 5(3), 21-41. <https://doi.org/10.18608/jla.2018.53.3>

**Important note**

To cite this publication, please use the final published version (if applicable). Please check the document version above.

**Copyright**

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

**Takedown policy**

Please contact us and provide details if you believe this document breaches copyrights. We will remove access to the work immediately and investigate your claim.

# Scaling Effective Learning Strategies: Retrieval Practice and Long-Term Knowledge Retention in MOOCs

Dan Davis<sup>1</sup>, René F. Kizilcec<sup>2</sup>, Claudia Hauff<sup>3</sup>, Geert-Jan Houben<sup>4</sup>

## Abstract

Large-scale online learning environments such as MOOCs provide an opportunity to evaluate the efficacy of learning strategies in an informal learning context with a diverse learner population. Here, we evaluate the extent to which retrieval practice — recognized as one of the most effective learning strategies — facilitates long-term knowledge retention (and thus learning outcomes) among MOOC learners using an instructional intervention. We observed no effect on learning outcomes and high levels of treatment non-compliance. In addition, we conducted a series of exploratory studies into long-term recall of knowledge acquired in MOOCs. Surprisingly, both passing and non-passing learners scored similarly on a knowledge post-test, retaining approximately two-thirds of what they learned over the long term.

## Notes for Practice

- We developed and tested a system that prompts students with previous questions to encourage retrieval practice in a MOOC.
- Retrieval practice showed no effect on learning outcomes or engagement.
- Learners who passed the course scored similarly on a knowledge post-test as did learners who didn't pass.
- Both passing and non-passing learners retained approximately two-thirds of the knowledge from the course after a two-week testing delay.

## Keywords

Retrieval practice, testing effect, experiment, knowledge retention, learning outcomes.

**Submitted:** 01.06.2018 — **Accepted:** 21.07.2018 — **Published:** 11.12.2018

Corresponding author<sup>1</sup> Email: [d.j.davis@tudelft.nl](mailto:d.j.davis@tudelft.nl) Address: R EEMCS, Web Information Systems, P.O. Box 5031, 2600 GA Delft, The Netherlands

<sup>2</sup> Email: [kizilcec@cornell.edu](mailto:kizilcec@cornell.edu) Address: Cornell University, 236 Gates Hall, Ithaca NY 14853

<sup>3</sup> Email: [c.hauff@tudelft.nl](mailto:c.hauff@tudelft.nl) Address: R EEMCS, Web Information Systems, P.O. Box 5031, 2600 GA Delft, The Netherlands

<sup>4</sup> Email: [g.j.p.m.houben@tudelft.nl](mailto:g.j.p.m.houben@tudelft.nl) Address: R EEMCS, Web Information Systems, P.O. Box 5031, 2600 GA Delft, The Netherlands

## 1. Introduction

A substantial body of research in the learning sciences has identified techniques and strategies that can improve learning outcomes, but translating these findings into practice continues to be challenging. An initial challenge is access to reliable information about which strategies work, which is the focus of scientific evidence aggregation hubs like the What Works Clearinghouse<sup>1</sup>. A subsequent challenge is knowing how to successfully implement and scale up an evidence-based strategy in potentially different learning contexts. This article addresses the latter challenge, which is becoming increasingly complex with the rise in popularity of online learning for diverse learner populations. We focus on a particular strategy that has been established as one of the most effective in facilitating learning. Retrieval practice, also known as the testing effect, 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 (Koedinger, Corbett, & Perfetti, 2012) — than passively revisiting the same information, for example by going over notes or book chapters (Adesope, Trevisan, & Sundararajan, 2017; Clark & Mayer, 2016; Roediger & Butler, 2011; Henry, Roediger, & Karpicke, 2016; Lindsey, Shroyer,

<sup>1</sup> <https://ies.ed.gov/ncee/wwc/>

Pashler, & Mozer, 2014; Karpicke & Roediger, 2008; Karpicke & Blunt, 2011).

Given the wealth of scientific evidence on the benefits of retrieval practice (cf. Section 2), we test the extent to which the testing effect can be leveraged in one of today's most popular digital learning settings: Massive Open Online Courses (MOOCs). Research into both MOOC platforms and MOOC learner behaviour has found that learners take a distinctly linear trajectory (Davis, Chen, Hauff, & Houben, 2016; Wen & Rose, 2014; Geigle & Zhai, 2017) through course content. Many learners take the path of least resistance towards earning a passing grade (Zhao et al., 2017), which does not involve any backtracking or revisiting previous course units — counter to a regularly spaced retrieval practice routine.

Although 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 (Kizilcec, Pe' rez-Sanagust' in, & Maldonado, 2017). 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 self-directed autodidacts who engage in such effective learning behaviour without additional support. This motivated us to create the Adaptive Retrieval Practice System (ARPS for short), a tool that encourages retrieval practice by automatically delivering quiz questions from previously studied course units, personalized to each learner. The system is 1) automatic in that the questions appear without any required action from the learner and 2) personalized in that questions are selected based on a learner's current progress in the course. We designed ARPS in accordance with the best practices in encouraging retrieval practice based on prior studies finding it to be effective (Adesope et al., 2017). 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 but static retrieval practice recommendations).

Based on the data we collected in this experiment, we investigated the benefits of retrieval practice in MOOCs guided by the following research questions:

- RQ1** How does an adaptive retrieval practice intervention affect learners' course engagement, learning outcomes, and self-regulation compared to generic recommendations of effective study strategies in a MOOC?
- RQ2** How does a push-based retrieval practice intervention (which requires learners to act) change learners' retrieval practice behaviour in a MOOC?

We also conducted exploratory analyses on the relationship between learner demographics (age, prior education levels, and country) and their engagement with ARPS, leading to a third research question:

- RQ3** Is there demographic variation in learners' engagement in retrieval practice?

In addition to collecting behavioural 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 two research questions:

- RQ4** To what extent is long-term knowledge retention facilitated in a MOOC?
- RQ5** What is the learner experience using ARPS to engage in retrieval practice?

The remainder of the article is organized as follows: In Section 2 we review the literature on retrieval practice, spacing, and knowledge retention. In Section 3 we introduce ARPS in detail. Section 4 discusses the methodology and our study design. Results are presented in Section 5, and Section 6 offers a discussion of the main findings, limitations to consider, and implications for future research.

## 2. Related Work

The design of the current study was based on prior research in the areas of retrieval practice, spaced vs. massed practice, and long-term knowledge retention.

### 2.1. Retrieval Practice

Decades of prior research on the effectiveness of different learning strategies has found retrieval practice to be effective in supporting long-term knowledge retention (Adesope et al., 2017; Clark & Mayer, 2016; Roediger & Butler, 2011; Henry et al., 2016; Lindsey et al., 2014; Karpicke & Roediger, 2008; Karpicke & Blunt, 2011; Custers, 2010). 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. Common methods of retrieval practice include the following: 1) taking practice tests, 2) making and using flashcards, 3) making one's own questions about the topic, 4) writing and/or drawing everything one can on a topic from memory<sup>2</sup>, and 5) creating a concept

<sup>2</sup> <http://www.learningscientists.org/blog/2016/6/23-1>

map of a topic from memory. Given the online medium and affordances of the edX platform, we here adopt the practice test approach.

A key line of inquiry in the present study is the extent to which effective retrieval practice can be fostered in the MOOC setting. Retrieval practice has been evaluated in a variety of contexts in prior research. The vast majority (82%) of studies considered in the recent meta-review by Adesope et al. (2017) were conducted in laboratory settings, compared to 11% in physical class settings (7% did not report the setting). The authors observed consistently large and significant effect sizes across both contexts. Furthermore, only 22% of those studies employed a randomized trial design, and, again, among those studies that did randomly assign participants to conditions, there was a consistently large and significant effect size. This trend of significant and large effect sizes holds for studies of post-secondary learners as well as studies specifically measuring learning outcomes (as opposed to engagement or retention). The design of our experiment is informed by these findings in order to maximize the similarity between our implementation and those that have found retrieval practice to be highly effective.

Critically evaluating the efficacy of retrieval practice in large-scale digital learning environments promises to advance theory by developing a deeper understanding of how retrieval practice can be effectively deployed/designed in a digital context as well as in the highly heterogeneous population represented by MOOC learners. Prior research on intelligent tutoring systems (ITS) has evaluated the benefits of teaching effective learning strategies and the effectiveness of various prompting strategies (Roll, Aleven, McLaren, & Koedinger, 2011; VanLehn et al., 2007; Bouchet, Harley, & Azevedo, 2016). The experimental results presented by Bouchet et al. (2016), for example, found that a scaffolding condition (in which fewer SRL supports in the ITS were delivered each week) was beneficial to learners even after the supports were removed completely. However, ITS studies are typically conducted in controlled lab settings with a considerably more homogeneous population than in MOOCs (Davis, Chen, Hauff, & Houben, 2018; Kizilcec & Brooks, 2017), so it is imperative that we scale such interventions up to gain a robust understanding of their heterogeneous effects.

Adesope et al. (2017) 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 role of feedback, 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 (Adesope et al., 2017). 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.

Roediger and Butler (2011) also offer a synthesis of published findings on retrieval practice (from studies also carried out in controlled, homogeneous settings). From the studies reviewed, the authors offer five key points on retrieval practice for promoting long-term knowledge: 1) retrieval practice is superior to reading for long-term retention, 2) repeated testing is more effective than a single test, 3) providing feedback is ideal but not required, 4) benefits are greatest when there is lag time between learning and practicing/retrieving, and 5) retrieval practice increases the likelihood of learning transfer — the application of learned knowledge in a new context (Roediger & Butler, 2011).

Consistent with the above findings, Johnson and Mayer (2009) 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 (ranging from one week to six months) and learning transfer.

A related body of research on spaced versus massed practice has found that a higher quantity of short, regularly spaced study sessions is more effective than a few long, massed sessions (Clark & Mayer, 2016). There is overlap in the research on retrieval practice versus spaced practice. Combining the insights from both literatures, an optimal study strategy in terms of achieving long-term knowledge retention is one of a regularly spaced retrieval practice routine (Miyamoto et al., 2015; Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008; Clark & Mayer, 2016). In fact, Miyamoto et al. (2015) provide evidence for this recommendation in the MOOC setting. They 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.

## 2.2. 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 (1885) curve of forgetting. 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. Figure 1 is our own illustration/depiction of Ebbinghaus' memory curve based on his findings.

Custers (2010) conducted a 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. While the shape of the Ebbinghaus forgetting curve seems to hold across contexts, the actual proportion of knowledge remembered varies (see Figure 1).

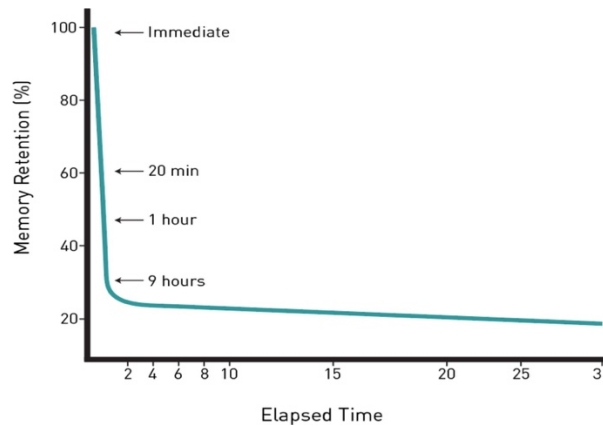


Figure 1. Rendition of Ebbinghaus’ (1885) memory curve. Recall rate as a function of lag time in days.

Considering the effect of retrieval practice on long-term retention, Lindsey et al. (2014) 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.

It is worth noting that forgetting can be viewed as an adaptive behaviour. Specifically, forgetting liberates the memory of outdated, unused information to create space for new, immediately relevant memories and knowledge (Richards & Frankland, 2017). Retrieval works against this natural, adaptive tendency to forget; by regularly reactivating and revisiting knowledge, the brain recognizes its frequent use, labels it as such, and stores it in long-term memory accordingly so that it is readily accessible going forward.

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 literature on retrieval practice and spacing effect, they also developed a support system to improve learner memory. Findings show that their support system, tuned to the “half-life regression model” of a learner’s knowledge, significantly improves learner memory (Streeter, 2015).

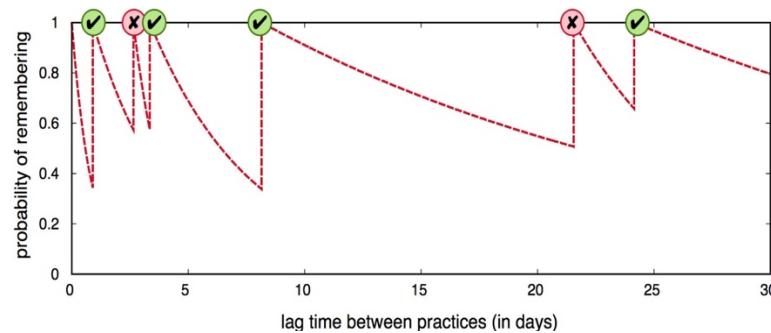


Figure 2. Example student word-learning traces over 30 days. Check marks indicate successful retrieval attempts, x’s indicate unsuccessful attempts, and the dashed line shows the predicted rate of forgetting (Settles & Meeder, 2016).

Researchers have used this model of learner knowledge at Duolingo<sup>3</sup> to explore the effect of retrieval practice on learning and knowledge retention. In Figure 2, Settles and Meeder (2016) illustrate this effect by showing how retrieval interrupts the forgetting curve and how each successful retrieval event makes the ensuing forgetting curve less steep. The design of the Duolingo platform is centred on retrieval events, as most learning activities constitute knowledge assessments. This is not the

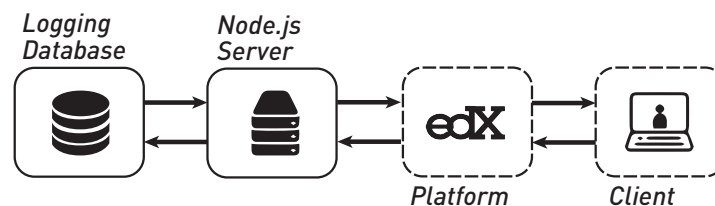
<sup>3</sup> [www.duolingo.com](http://www.duolingo.com)

case in most MOOCs where lecture videos constitute the primary instructional content. Thus, understanding how to foster effective retrieval practice in MOOCs warrants further investigation.

### 3. Adaptive Retrieval Practice System

ARPS is a client-server application (written in JavaScript/node.js) that provides automated, scalable, personalized retrieval practice questions to MOOC learners. We developed ARPS for use within the edX platform, taking advantage of the RAW HTML input affordance. This allows course creators to build custom interfaces within the platform that render along with the standard edX content (such as videos, quizzes, etc.). We designed the user interface in accordance with previous approaches to encouraging retrieval practice (Adesope et al., 2017; Clark & Mayer, 2016; Roediger & Butler, 2011; Henry et al., 2016; Lindsey et al., 2014; Karpicke & Roediger, 2008; Karpicke & Blunt, 2011; Custers, 2010).

As illustrated in Figure 3, 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 behaviour in the course. To do so, it randomly selects from a pool of questions only pertaining to content the learner has already been exposed to<sup>4</sup>. Each question is pushed to the learner in the form of a qCard, an example of which is shown in Figure 4. These qCards appear to the learner as a pop-up within the browser tab. We logged all qCard interactions: whether it was ignored or attempted, the correctness of the attempt, and the duration of the interaction.



**Figure 3.** ARPS system architecture. Solid outlines indicate back-end components; dashed lines indicate front-end components.

In every course page (see Figure 5 for the standard appearance and organization of a course page in edX) we inserted our code (a combination of JavaScript, HTML, and CSS) in a RAW HTML component. This code was either a sensor code (included in every page) or a **qCard generator** (included in some pages). The sensor code did not result in any visible/rendered component; rather, it tracked behaviour and logged learner progress throughout the course. We inserted the qCard generator code on each lecture video page.

The learners’ course progress and all of their interactions with the qCards were stored in our server, to enable real-time querying of each learner’s quiz history. By default, edX provides user logs in 24-hour intervals only, making it impossible to employ edX’s default logs for ARPS. Our system design thus effectively acts as a RESTful API for the edX platform in allowing the real-time logging of information and delivery of personalized content.

In contrast to previous interventions in MOOCs (Rosen et al., 2017; van der Zee et al., 2018; Davis, Chen, van der Zee, Hauff, & Houben, 2016; Davis et al., 2017; Kizilcec & Cohen, 2017; Kizilcec, Saltarelli, Reich, & Cohen, 2017; Yeomans & Reich, 2017; Gamage et al., 2017), we push questions to learners instead of requiring the learner to seek the questions out (e.g., by pressing a button). We adopted this 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 non-compliance that has arisen in past research (Davis, Chen, van der Zee et al., 2016; Davis et al., 2017). In the case of multiple choice (MC) questions (example problem text in Figure 4), 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. For incorrect responses, a learner receives a negative feedback symbol (an x alongside text encouraging them to make another attempt) which disappears after 4 seconds and redirects the learner to the original question to reattempt the problem<sup>5</sup>. Learners are not informed of the correct response after incorrect attempts in order to give them a chance to actively reflect and re-evaluate the problem. The qCard can also be closed without attempting the question.

<sup>4</sup> Note that learners could have engaged with a course’s content for the week without having engaged with all of its quiz questions.

<sup>5</sup> A video demonstration of the system in action is available at <https://youtu.be/25ckrrMzCr4>

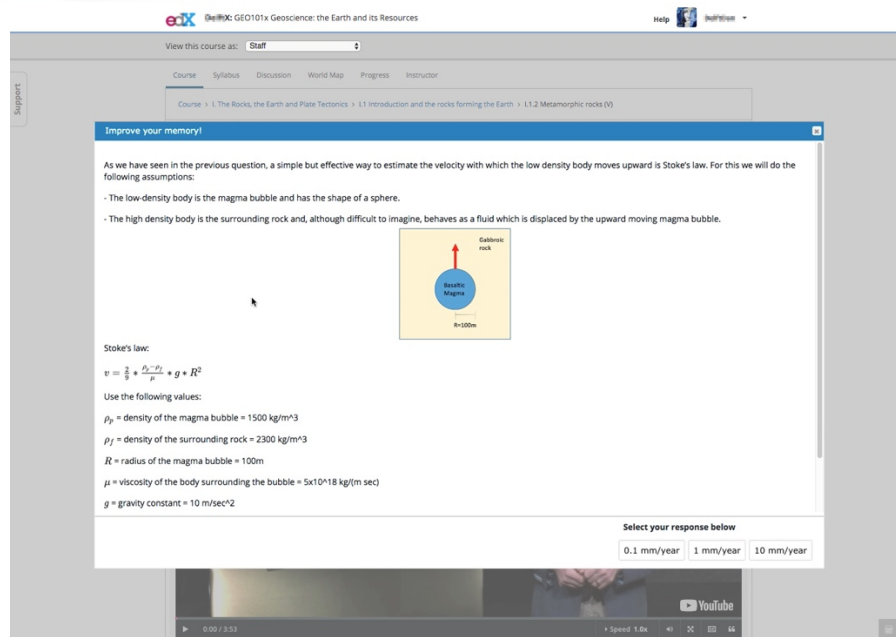


Figure 4. Example qCard in the GeoscienceX course. The main body of the qCard contains the question text, and the bar at the bottom contains the MC answer buttons. The grey “x” at the top right corner closes the window and dismisses the problem.

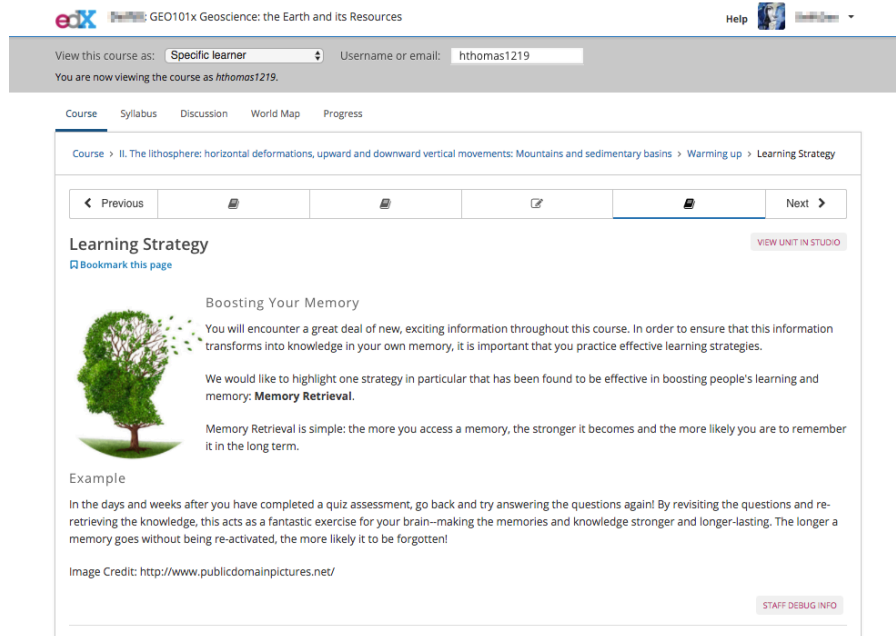


Figure 5. Page shown to learners in the control condition at the beginning of each course week describing how to practice an effective memory retrieval routine.

Apart from MC questions, we also enabled one other question type<sup>6</sup> to appear in qCards: the numeric input (NI) type. These questions require the learner to calculate a solution and enter the answer. While requiring more effort than a single click response, we included this question type to allow for a comparison between the two. Even though MC questions were most common in GeoscienceX, this is not the case across all MOOCs.

The following is an example of an easy (less than 5% of incorrect responses) multiple choice question in GeoscienceX:

<sup>6</sup> 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 in which we deployed ARPS.

*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?*

The following is an example of a difficult (5% correct response rate) numerical input 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?*

These two sample problems are representative of the set of questions delivered in ARPS in that they prompt the learner to mathematically simulate an outcome of a situation based on given circumstances.

## 4. Study Design

In this section, we describe the design and context of our empirical study.

### 4.1. Participants

A total of 2,324 learners enrolled in the MOOC Geoscience: The Earth and its Resources (or GeoscienceX), which was offered on the edX platform between 23 May 2017 and 26 July 2017<sup>7</sup>. The course consists of 56 lecture videos and 217 graded quiz questions. We manually determined 132 of those 217 questions to be suitable for use with qCards (multi-step problems were excluded so that each qCard could be answered independently), 112 were MC and 20 were NI questions.

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 with any incentive beyond earning a course certificate for participating in the study.

We define the study sample as the 1,047 enrolled learners who entered the course at least once: 524 assigned to the control condition and 523 to the treatment condition. The random assignment was based on each learner's numeric edX ID (modulo operation).

We collected a stratified sample for the post-course survey and knowledge retention quiz. We sent an invitation email to all 102 learners who attempted a problem within the ARPS system — nine completed the survey, an 8.8% response rate — and the 150 learners with the highest final grades in the control condition — eleven completed the survey, a 7.3% response rate.

### 4.2. Procedure

This study was designed as a randomized controlled trial. Upon enrolling in GeoscienceX, 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 5).
- **Treatment condition:** ARPS was added to the course to deliver quiz questions 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 about how ARPS works and that responses to the qCard do not count towards learners' final grades 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 a stratified sample of learners in the course (as previously defined) two weeks after the course had ended. The survey contained a random selection of ten assessment questions from GeoscienceX. 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.

<sup>7</sup> The study was pre-registered at <https://osf.io/4py2h/>



### 4.3. Measures

To compare the behaviour of learners in the control and treatment conditions, we considered the following in-course events:

- **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)

We also collected the following data in the post-course survey:

- **Post-exam quiz score** (between 0–100%)
- Learner intentions (e.g., to complete or just audit)
- Prior education level (highest degree achieved)

The three variables printed in bold — final grade, course completion, and post-exam quiz score — serve as primary outcome variables in our study for the following reasons: 1) a learner’s final grade is the best available indicator of their performance in the course in terms of their short-term mastery of the materials, and 2) the post-exam quiz score measures how well learners retained the knowledge two weeks after the end of the course.

## 5. Results

This section presents results from the five sets of analyses we conducted: 1) estimating the causal effect of the retrieval practice intervention (RQ1), 2) examining how learners interacted with ARPS (RQ2), 3) exploring demographic variation in the use of ARPS (RQ3), 4) modelling how learner knowledge evolves and decays over time (RQ4), and, 5) understanding learners’ experience with ARPS from a qualitative angle using survey responses (RQ5).

Group	N=	<b>Final Grade</b>	<b>Passing Rate</b>	<b># Video Interactions</b>	<b># Quiz Submissions</b>	<b># Forum Posts</b>	<b>Time in Course</b>	<b>Time with qCards</b>	<b>qCards Seen</b>
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

**Note:** 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 differences observed between the control and treatment groups are not statistically significant.

### 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 groups, we can estimate the effect of encouragement to engage in retrieval practice with ARPS with the following pre-registered confirmatory analyses. Table 1 provides summary statistics for learners in the two experimental conditions, showing no significant differences between conditions, according to tests of equal or given proportions and Kruskal-Wallis tests (all  $p > 0.05$ ). However, many learners who were encouraged did not engage in retrieval practice, which is a form of treatment non-compliance. 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. Due to the small sample size and low compliance rate, however, we adjusted our analytic approach. Specifically, we analyze the experiment as an encouragement design beyond estimating average treatment effects, and we do not apply the pre-registered sample exclusion criteria because they could inadvertently bias the causal inference.

The primary outcome measure is the final course grade, which determines certificate eligibility (the passing threshold is 60%). Table 2 contains summary statistics for grade and certification outcomes in the control group and the treatment group, overall and separately for treatment compliers and non-compliers. 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  $\chi^2_{df=1} = 0.592$ ,  $p = 0.442$ ).

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; Angrist, Imbens, & Rubin, 1996). 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 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 conduct a per-protocol analysis, which considers only those learners who completed the protocol for their allocated treatment. We estimate the per-protocol effect as the difference in average outcomes between treatment compliers and control compliers, which is the entire control group by design. 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 analysis does not yield causal estimates due to self-selection bias; in particular, we compare everyone in the control group with those highly motivated learners who comply in the treatment group. In fact, treatment compliance is strongly correlated with receiving a higher grade (Spearman’s  $r = 0.56, p < 0.001$ ).

**Table 2. 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	Non-complier	421	23%	0.2%	[0, 0, 0]

**Note:** Compliers are learners who complied with the intervention; non-compliers are those who ignored the intervention.

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 behaviour in terms of 1) revisiting previous course content such as lecture videos, 2) self-monitoring by checking their personal progress page, and 3) generally persisting longer in the course. No evidence in support of the hypothesized behaviour was found, neither in terms of the ITT (Kruskal-Wallis  $\chi^2$ -values < 0.68,  $p$ -values > 0.41) nor in terms of the LATE ( $|z|$ -values < 0.98,  $p$ -values > 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 (Adesope et al., 2017; Clark & Mayer, 2016; Roediger & Butler, 2011; Henry et al., 2016; Lindsey et al., 2014; Karpicke & Roediger, 2008; Karpicke & Blunt, 2011) we find that:

A causal analysis yields no evidence that encouraging retrieval practice raised learning, performance, or self-regulatory outcomes in this course.

We also observe 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 assessments). To better understand how different groups of learners used ARPS and performed on subsequent learning assessments, we conduct a series of exploratory analyses.

### 5.2. Navigational Behaviour Following Retrieval Cues

With the following exploratory analyses, we now aim to understand the extent to which engagement with the qCards (i.e., compliance with the experimental manipulation) led to changes in learner navigational behaviour through the course (RQ2). Specifically, we examine whether or not there were marked changes in behaviour immediately following the appearance of a qCard — as the qCard is from a prior unit in the course, we expected this to stimulate learners to skip backwards in the course and revisit old content to refresh their memories.

Once again, we divide learners in the treatment condition into compliers and non-compliers. For the two groups, we then isolate every page load immediately following the appearance of a qCard (i.e., where did learners navigate to after being exposed to a qCard?), measured by the proportion of times where learners either continue forwards or track backwards to an earlier point in the course. There is no significant difference between compliers and non-compliers in terms of their forward or back-seeking behaviour immediately following exposure to a qCard ( $\chi^2 = 1.5, p = 0.22$ ). Both groups navigated backwards

after qCards 8–10% of the time. This indicates that qCards did not have a significant effect on learners’ navigation and did not motivate them to revisit previous course topics to retrieve relevant information.

### 5.3. Engagement with Retrieval Cues

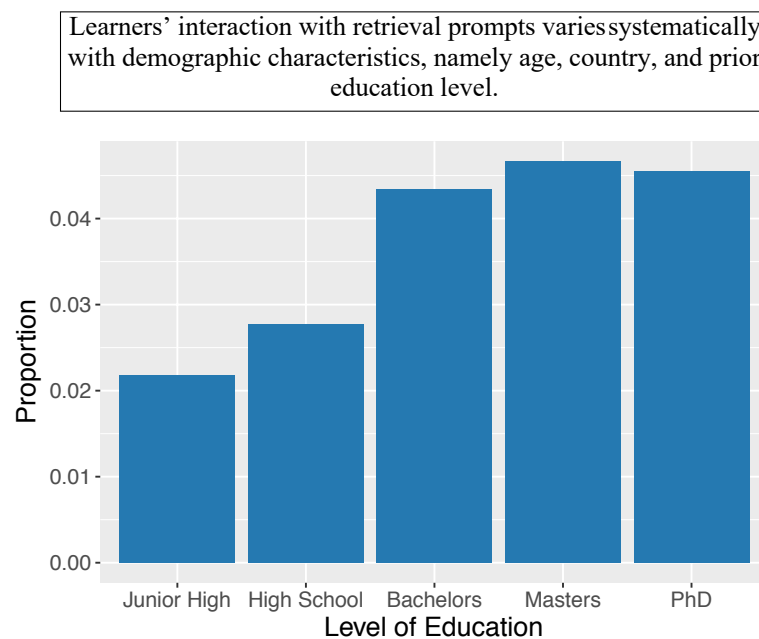
As prior work has found significant demographic variation in learner behaviour and achievement in MOOCs (Kizilcec, Saltarelli, et al., 2017; Kizilcec & Halawa, 2015; Davis et al., 2017), we explore such variation in the use of ARPS (RQ3).

#### 5.1.1 Demographic Analysis

We here consider learner characteristics that have been identified as important sources of variation: age, prior education level, and the human development index (HDI) of the country in which a learner is located (HDI = the United Nations’ combined measure of a country’s economic, health, and educational standing; Sen, 1994). The following analyses explore whether the retrieval practice system was similarly accessible and helpful to all groups of learners, as we intended it to be in the design phase.

We find that learners with a higher level of education are more likely to interact with ARPS (i.e., intervention compliers), as illustrated in Figure 6. This suggests that the design of the system may be less accessible or usable for learners with lower levels of education. Next, we test for age differences between compliers and non-compliers. The average complier is 38 while the average non-complier is only 32 years old, a statistically significant difference ( $p= 0.003$ ,  $t=2.99$ ). Finally, we examine whether learners from more developed countries would be more or less likely to engage with qCards. We find that compliers came from significantly more developed countries (higher HDI) than non-compliers, even when adjusting for prior education levels (ANCOVA  $F = 4.93$ ,  $p =0.027$ ).

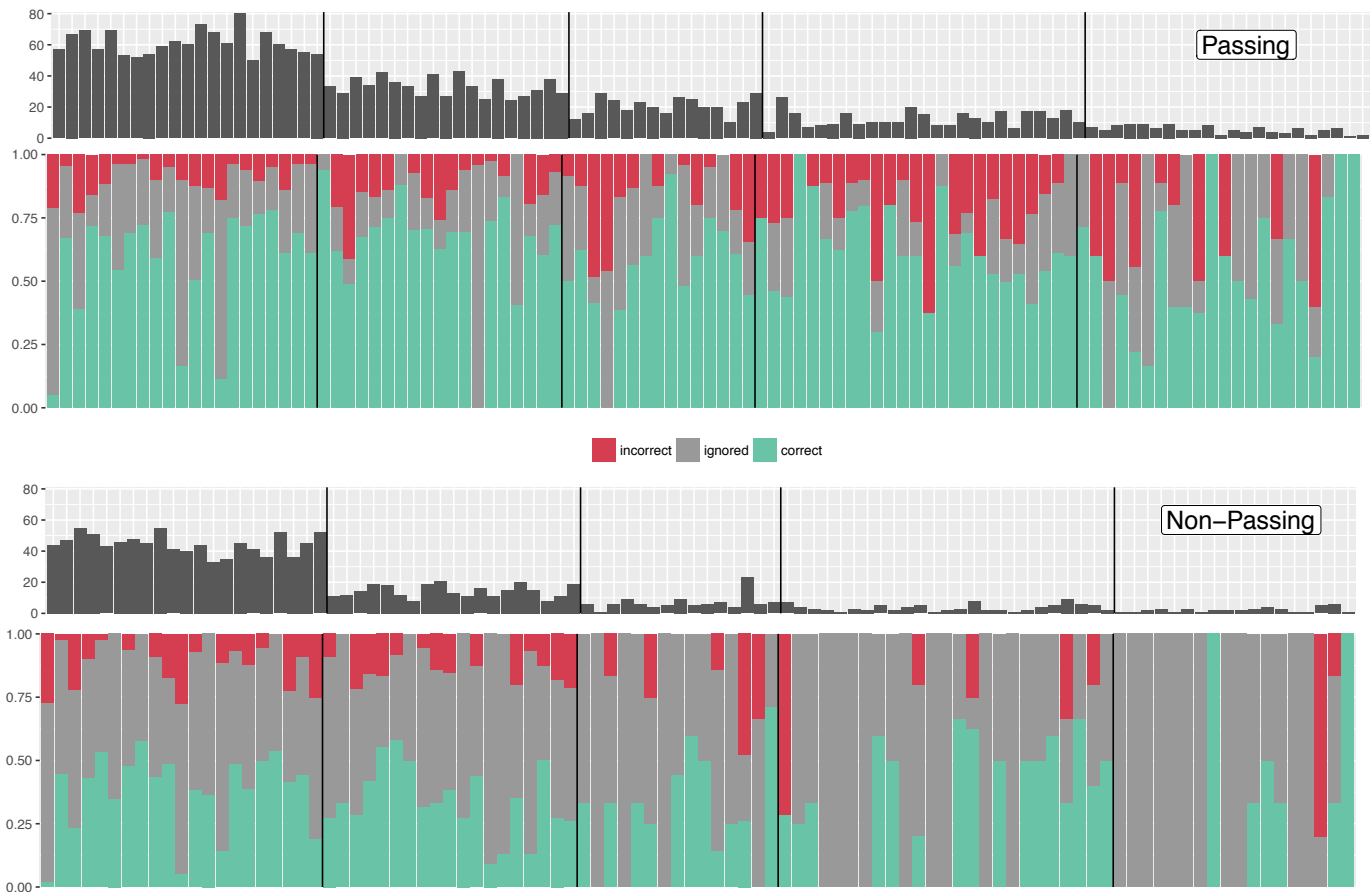
One goal of open education (and ARPS) is to make education more accessible to historically disadvantaged populations. We found, however, that ARPS was not used equally by different demographic groups. Younger and less educated learners from developing countries are less likely to use the system and thus miss an opportunity to benefit from the testing effect facilitated by ARPS.



**Figure 6.** Compliance vs. prior education level: proportion of all learners of each level of prior education to have complied with the intervention (engaged with at least one qCard).

#### 5.3.1. Question-level Analysis

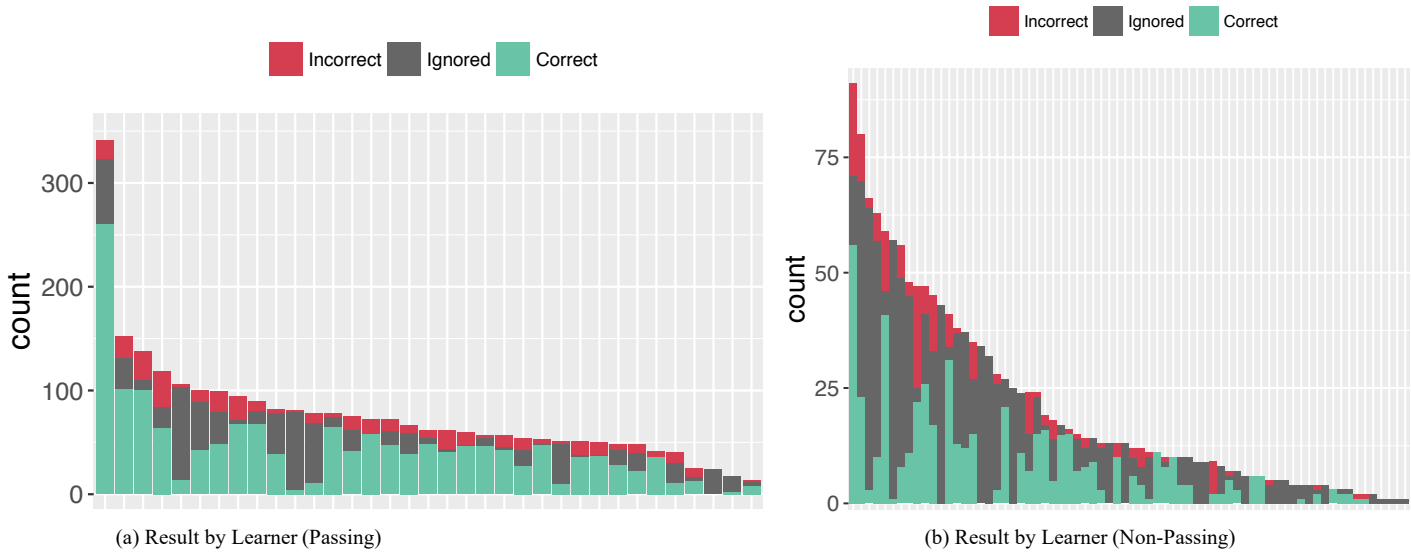
Figure 7 illustrates learner responses for every question delivered by ARPS, indicating which questions learners tended to struggle with (or ignore). The figure reveals that the choice to attempt or ignore a qCard is strongly associated with a learner’s eventual passing or failing of the course. Moreover, 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 Zhao et al., 2017, and Kizilcec, Piech, & Schneider, 2013).



**Figure 7.** A question-by-question breakdown of every learner interaction with the qCards. The top two figures represent the behaviour of passing learners — the upper image shows the number of learners being served that question, and the lower shows the proportion of ignored, incorrect, and correct responses — 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 (from 1–5). Best viewed in colour.

From Figure 8, we observe that passing and non-passing learners do not differ in their rate of giving incorrect responses, which would indicate misconceptions or a lack of understanding of the materials. Instead, they differ in their choice to ignore the problems altogether. 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 percentage points,  $\chi^2 = 9.63$ ,  $p = 0.002$ ) between the proportion of correct or incorrect responses between passing and non-passing learners. 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. Figure 8a shows that there is a single learner who answered more than 300 qCard questions. This is the only learner who took advantage of the ARPS functionality that allowed learners to continue answering qCard questions after engaging with the original (automatically appearing) one, which enabled a Duolingo-like approach where a learner can engage with an extended sequence of assessments (or retrieval cues) from past course weeks<sup>8</sup>. This functionality was previously not available in edX, as quiz questions can only be attempted a finite number of times — most commonly once or twice.

<sup>8</sup> We found that this learner followed a pattern of regular weekly engagement in the course over the span of 1.5 months and that their high level of ARPS usage was not a result of gaming the system or a single massed practice session, but rather an indication of a highly disciplined pattern of spaced learning.



**Figure 8.** Each bar corresponds to one learner. Only one learner took advantage of the “infinite quizzing” capability by frequently using the “Generate new qCard” button. Best viewed in colour.

**Table 3.** ARPS Problem Response (left) and Correctness (right)

	Attempted <sup>†</sup>	Ignored <sup>†</sup>	Correct	Incorrect
Non-passing	47%	53%	76%	24%
Passing	73%	27%	82%	18%

**Note:** Significant differences ( $p < 0.001$ ) between passing and non-passing learners in each column are indicated with †.

### 5.3.2. First Question Response

To 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 9a shows the passing rate of learners segmented according to their first interaction with a qCard. Learners who attempted the first qCard instead of ignoring it had a 47% chance of passing the course. In contrast, learners who ignored their first qCard only had a 14% chance of passing. Figure 9b additionally illustrates the relationship between the result of the first qCard attempt and course completion. 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.

While Figure 9a shows passing rates conditional on attempting a problem, Figure 9b shows passing rates conditional on the result of the first attempt. The latter adds no predictive power; getting the first question correct is still associated with about a 50% chance of passing the course.

To evaluate whether the response of a learner’s second qCard problem adds any predictive value, we replicate the analysis shown in Figure 9 for the responses to the first two qCards delivered to each learner. No difference in predictive power 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.

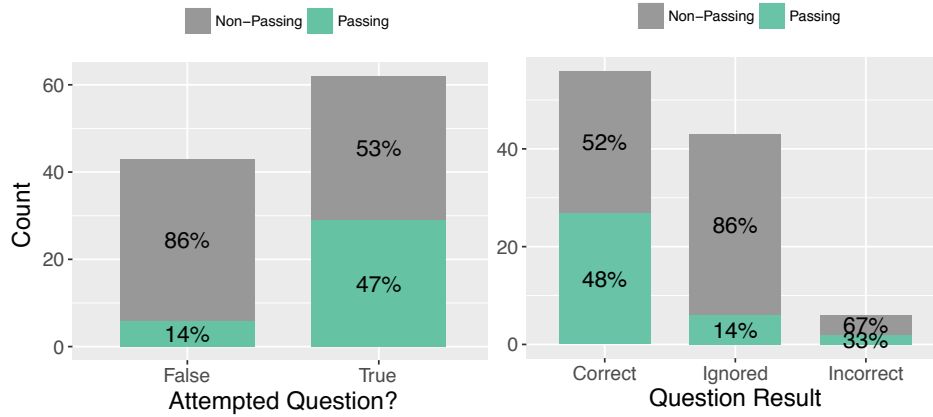
We conclude that initial adoption of ARPS appears to depend partly on learner motivation to complete the course.

### 5.3.3. Response Duration

We next explore how much time learners spent interacting with qCards and how the amount of time they spent predicts the outcome of the interaction. Figure 10 shows the proportion of correct, incorrect, and ignored responses as a function of time elapsed with a qCard. We find that the decision to ignore the qCard happened 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). When learners did attempt to answer the question, the amount of time they spent was not predictive of the correctness of their response, with a median duration for correct and incorrect responses of 18 and 16 seconds, respectively.

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

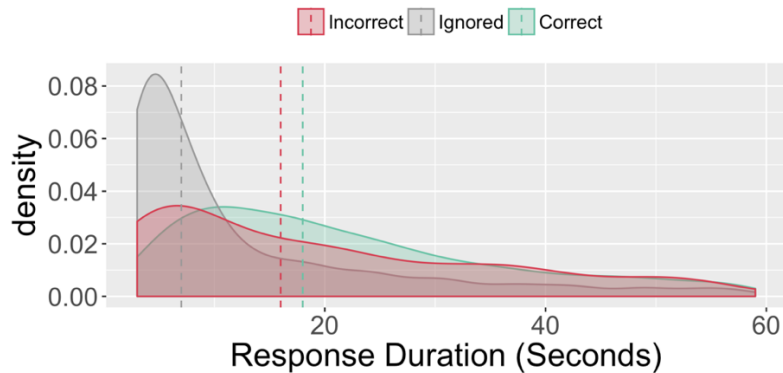
There is no significant difference in assessed knowledge between passing and non-passing learners; the key difference lies in a learner’s willingness to engage with the retrieval practice questions.



a) First question response: The likelihood of course completion based on the learner’s choice to engage with the first qCard they were shown.

(b) First question result: The likelihood of course completion based on the learner’s result of engaging with the first qCard they were shown.

**Figure 9.** The predictive value of learner engagement with the first qCard to which they were exposed. “True” indicates both correct and incorrect responses, and “False” indicates the qCard was ignored. Best viewed in colour.



**Figure 10.** Result likelihood vs. time: Kernel Density Estimation plot showing the relationship between time 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 colour.

### 5.4. Modelling Knowledge Over Time

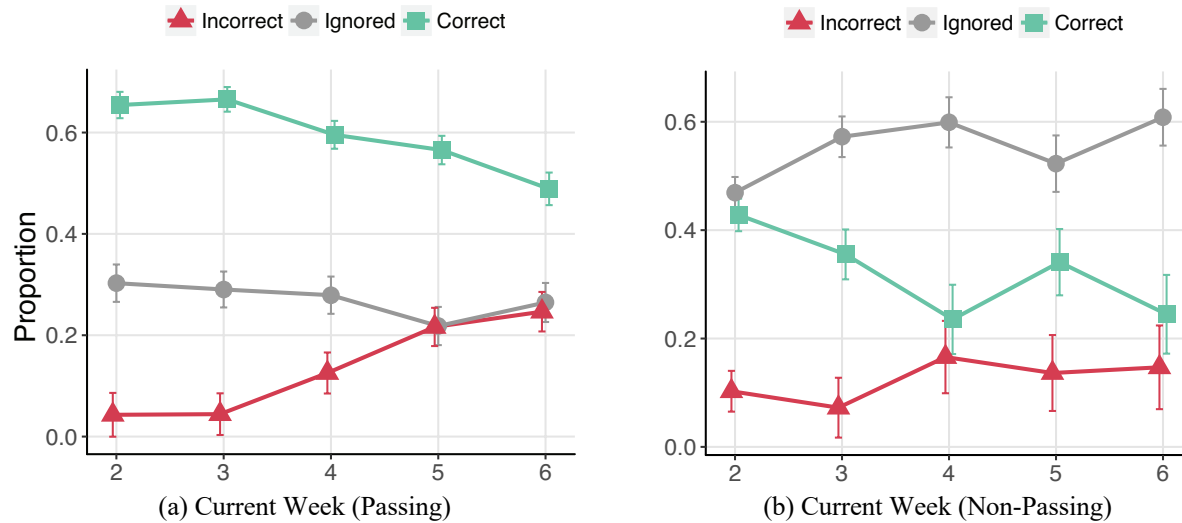
One of the contributions of ARPS is the data set that it generates: by tracking learner responses to these periodic, formative, and ungraded questions throughout the entire course, we have a longitudinal account of learners’ evolving knowledge state throughout the entire process of instruction. In this section, we explore how learner knowledge (as measured by performance with the qCards) deteriorates over time (RQ4).

Figure 11 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.

Figure 11a illustrates the forgetting curve for passing learners in GeoscienceX. We observe a statistically significant decrease in performance between weeks 2 and 6 (the correct response rate drops from 67% to 49%;  $\chi^2 = 32.8, p < 0.001$ ). While the proportion of ignored responses remains steadily low, the proportion of correct responses drops by 18 percentage

points (nearly identical to the forgetting rate found in Lindsey et al., 2014). The rate of incorrect responses increased from 4% to 25% ( $\chi^2 = 87.8, p < 0.001$ ).

Figure 11b illustrates the forgetting curve for non-passing learners. We observe that the choice to ignore qCards was common throughout the duration of the course, 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 is in the rate of ignored responses, which increases from 47% in week 2 to 61% in week 6.



**Figure 11.** Week-by-week results of learner 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 colour.

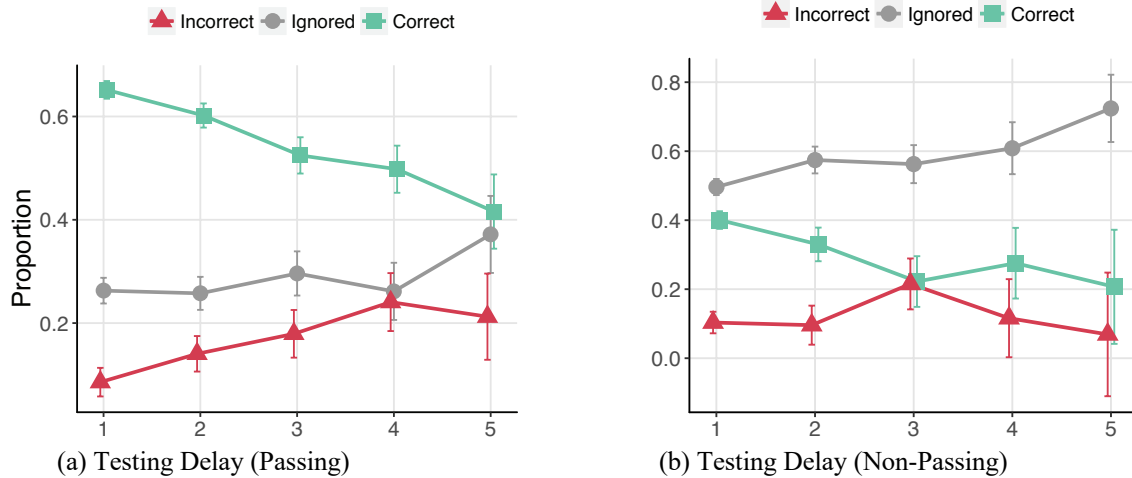
We identify two main contributing factors to this decline in performance over time. First, the amount of assessed content increases each week; in week 6 there are five course weeks' worth of content to be assessed, whereas in week 2 there is only content from week 1 being assessed. Second, people simply forget more with the passing of time (Richards & Frankland, 2017); each passing week moves the learner temporally farther away from when the content was initially learned and therefore, without regular retrieval, increases the likelihood of this knowledge being marked as non-essential and thus forgotten.

We next explore the relationship between testing delay and learners' memory and performance on qCards. In Figure 12, 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 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 11 shows how the amount of content covered/assessed is related to performance, Figure 12 illustrates how the delay in testing relates to performance.

We observe very similar trends for both passing and non-passing learners. For passing learners, there is a drop in the correct response rate from 1 week elapsed to 5 weeks elapsed (65% to 42%,  $\chi^2 = 23.6, p < 0.001$ ). Similarly, there is an increase in the 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 increased 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 non-passing learners; they are able to identify questions that they likely do not know the answer to and choose to ignore them.

Another promising finding from this analysis concerns learners' short-term knowledge retention. As illustrated by Figure 12, passing learners correctly answer 88% of problems that were attempted with 1 week elapsed. Non-passing learners also exhibit high recall levels with 79% correct (note that the required passing grade for the course was 60%). This, in tandem with the rapid rates of forgetting, suggests that there is room for a learning intervention to help learners commit new knowledge to long-term memory. We expected ARPS to accomplish this, but it was ultimately ineffective at promoting long-term knowledge retention.



**Figure 12.** 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 colour.

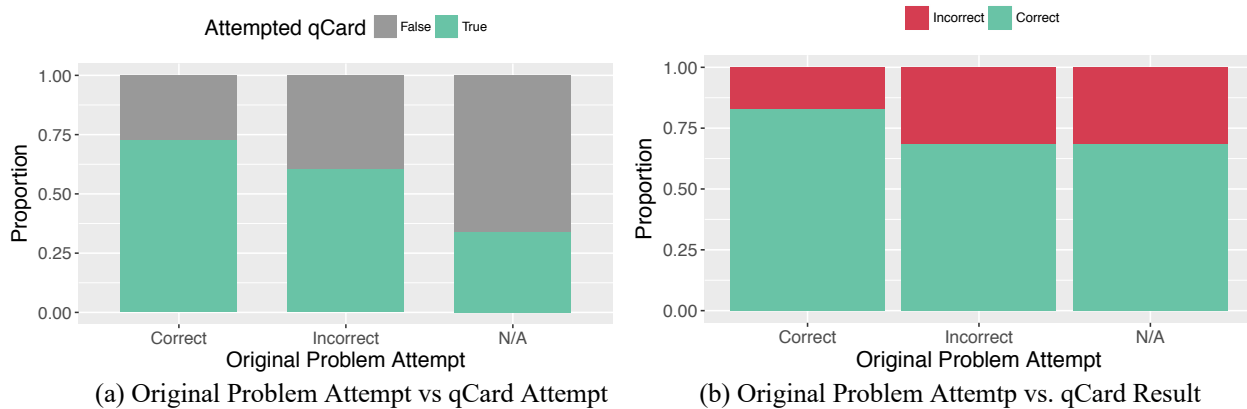
From the above findings for assessed knowledge as a function of both time and course advancement, we conclude

Knowledge recall deteriorates with the introduction of more course concepts/materials and the passing of time.

#### 5.4.1. Considering the Initial Knowledge State

The questions shown on qCards are sourced directly from the questions that appear earlier in the course as created by the course instructor. The results illustrated in Figures 11 and 12 show learner performance on qCards without taking into consideration learner performance on the original graded question that appeared earlier in the course. We may not expect a learner to correctly answer a question later if they never got it right in the first place. With Figures 13–15, we explore the learner’s engagement and performance with the qCards related to their performance the first time they answered the question.

Figure 13a shows the likelihood of a learner attempting a qCard based on their performance on the original question in the course. When a qCard is shown to a learner who initially answered the question correctly, the learner will attempt the qCard 73% of the time. However, if the question was initially answered incorrectly, only 61% reattempt it. And for questions that were ignored or never answered, those qCards were only attempted 34% of the time. These percentage values are all significantly different from each other ( $\chi^2 > 22, p < 0.0001$ ). This indicates that a learner’s performance on the original problem in the course is significant for their likelihood of engaging with a qCard showing that same question.



**Figure 13.** qCard performance vs. original problem attempt. N/A means not attempted. Best viewed in colour.

Figure 13b shows the likelihood of a learner either answering the qCard correctly, incorrectly, or ignoring it based on the learner’s performance on the original quiz question in the course. When faced with qCards showing questions that the



learner initially answered correctly, they get it correct again 61% of the time. In contrast, if the initial question was answered incorrectly or ignored, the chance of getting it right in a qCard later on drops to 42% and 23%, respectively. These percentages are significantly different from each other ( $\chi^2 > 40.0, p < 0.0001$ ). This indicates that a learner’s initial performance on the question significantly predicts their likelihood of answering the corresponding qCard correctly. Although learners do not perform well on the qCards in general, they demonstrate consistent knowledge of mastered concepts over time.

We next replicate the analyses done in Figures 11 and 12, except this time we again take into consideration learner performance on the original problem in the course. These results are shown in Figures 14 and 15.

We find from Figure 14a that for passing learners, with each successive course week there is a steady decrease in the likelihood of answering qCards correctly, even when they had answered the original question correctly. We likewise observe a weekly increase in the proportion of incorrect answers and a slight decline in the proportion of ignored qCards. This trend is amplified for questions that were originally ignored. In week 2, qCard questions that were originally ignored were answered correctly 79% of the time and this proportion drops to 7% in week 6. This indicates the importance of merely attempting problems in the course, as the week 6 rate of correct qCard responses to questions that were originally answered incorrectly is 40% — in other words, learners are 5.7 times more likely to answer a qCard correctly when they attempted the original question and got it wrong than when they ignored/did not attempt the original question.

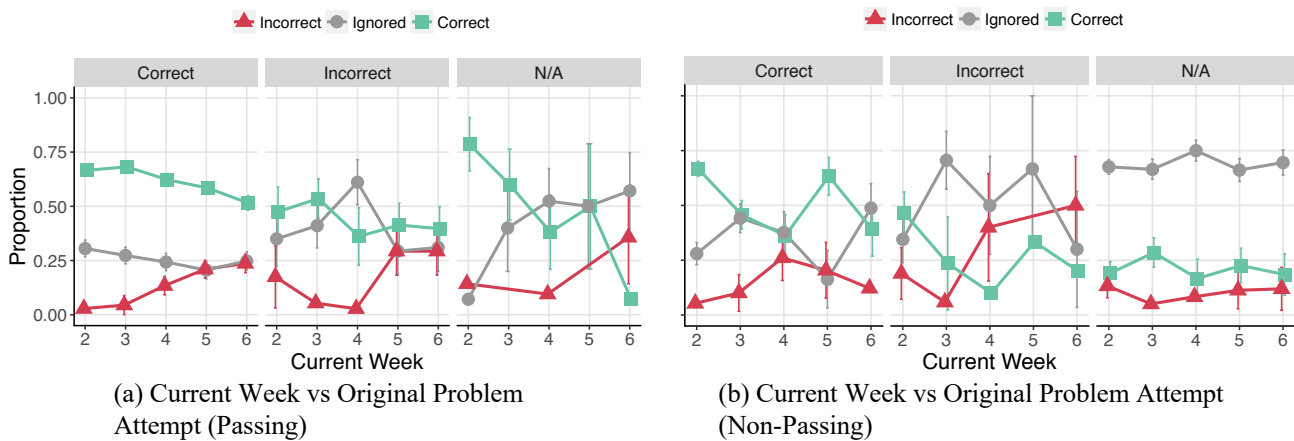


Figure 14. qCard performance vs. current course week and original problem attempt. Error bars show standard error. Best viewed in colour.

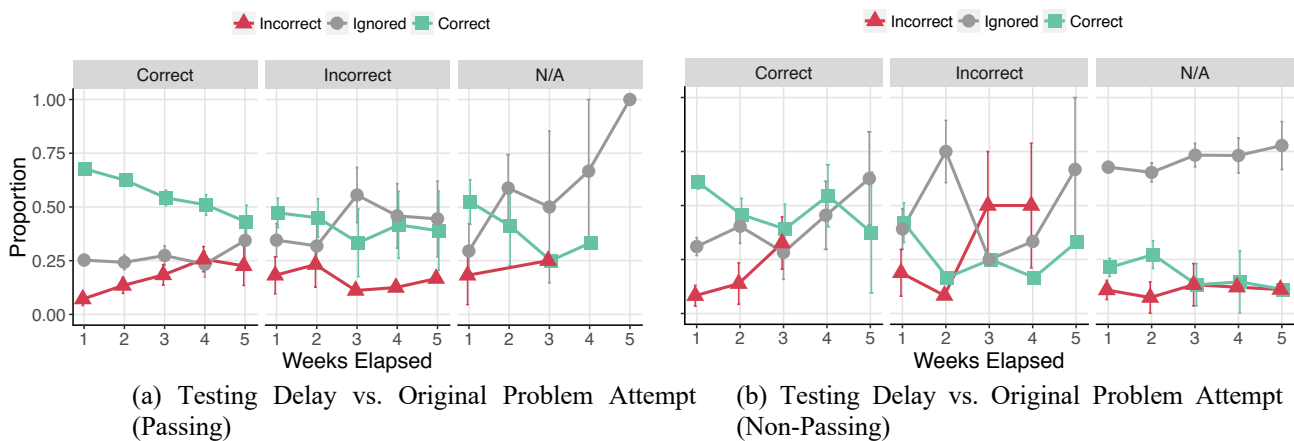


Figure 15. qCard performance vs. testing delay and original problem attempt. Error bars show standard error. Best viewed in colour.

Figure 15 shows the effect of testing delay and original problem performance on qCard performance. For passing learners, the most prominent trend to consider is the steady decline in performance over time for questions that were originally answered correctly. In conducting this follow-up analysis, we expected to find that taking learners’ original performance on individual questions into consideration would reduce the apparent rate of forgetting; however, the trend does hold. For qCard questions with a 1-week testing delay, passing learners answer them correctly 68% of the time. Compare this to a 5-week testing delay

where qCards are answered correctly only 39% of the time (a chi-Squared test indicates that this difference is statistically significant:  $p < 0.0001$ ,  $\chi^2 = 21.8$ ). With this result, we identify more evidence of the trend of MOOC learners' knowledge (as ascertained by quiz question performance) to steadily deteriorate over time, even if they had initially exhibited some level of mastery.

Well-spaced retrieval practice is known to improve long-term knowledge retention, which is typically evaluated by a post-test with some lag time between exposure to the learning materials and the assessment (Lindsey et al., 2014; Custers, 2010). As the GeoscienceX course used weekly quizzes for assessment, we took a random selection of ten quiz questions from throughout the six end-of-week quizzes and created a post-course knowledge assessment that was delivered to learners in a survey format two weeks after the course had ended.

We compare performance on the post-test between the two experimental conditions and found no significant difference (RQ4). 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 ( $N = 20$ ,  $t(17.6) = 0.45$ ,  $p = 0.66$ ). Notably, our findings for long-term knowledge retention are consistent with prior literature (Lindsey et al., 2014; Custers, 2010), which finds 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 learner experience with ARPS, we provided a stratified sample of learners with a mix of rating and open-ended questions in a post-course survey (RQ5). Using an adapted version of the system usability scale (SUS; Brooke, 1996), we found that the average usability score of ARPS is 73.9 ( $\sigma = 12.2$ ) out of 100. According to Bangor, Kortum, and Miller (2009), this is categorized as “acceptable usability” and the system's usability falls into the third quartile of SUS scores overall. For a research prototype that was not developed as a production system, this is a notably positive rating. We do acknowledge, however, that there is room for improvement in the system's usability, and that this improvement could cause more learners to engage and benefit. We find evidence for survey response bias in terms of course performance. Learners who responded to the survey achieved higher grades in the course (median = 82%) than learners who did not respond (median = 4%; Kruskal-Wallis  $\chi^2 = 26.7$ ,  $p < 0.0001$ ).

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 less desirable aspects of ARPS.

One learner explained how the type of problem delivered was a key factor in their use of ARPS, and how, for qCard questions they would prefer not to have to do any calculations:

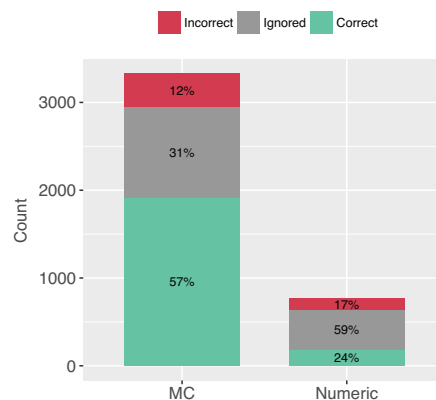
*It [would] be better if only conceptual questions [were] asked for [the] pop quiz, it's troublesome if calculation is required. If calculation is required, I would prefer that the options are equations so that we can choose the right equation without evaluating them.*

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 that may have otherwise been forgotten. One learner also reported that ARPS helped to “*reinforce my learning from previous chapters,*” and another learner reported, “*It helps me to review important information in previous sections.*” So while the challenge of compliance and engagement remains the priority, it is encouraging that those who did choose to engage with ARPS did indeed experience and understand its benefits.

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 16 reveals that, indeed, this learner was not alone in their sentiment. We found a 69% likelihood of learners attempting a MC qCard problem type compared to 41% attempting NI problems ( $p < 0.001$ ). 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 highly efficient interaction process — learners are far more likely to attempt a problem that only requires a single click to select from a list of answers than one that requires two extra steps — generating an answer from scratch and then typing it out. Nevertheless, we cannot identify from the data which of these two steps contributes more to questions being ignored.

## 6. Discussion

In this study we have evaluated the extent to which the effectiveness of retrieval practice directly translates from traditional classroom and laboratory settings to MOOCs. We replicated the implementation of retrieval practice from past studies as closely as possible in this new context. The design therefore aligns with the best practices in the retrieval practice literature as outlined by Adesope et al. (2017). While Adesope et al. (2017) showed consistent positive effects on learning outcomes in randomized experiments in post-secondary physical classroom settings, no significant effects were found in the present study in the context of a MOOC. This indicates that traditional instructional design practices for encouraging retrieval practice in a traditional classroom or laboratory setting do not directly translate to this online context. We therefore highlight the need for further research to empirically investigate how evidence-based instructional/learning strategies can be successfully implemented in online learning contexts. In the case of retrieval practice, we find that directly translating what works in traditional contexts to MOOCs is especially ineffective for younger and less educated learners who joined the MOOC from developing countries. More research is needed to understand how to adapt learning strategies for novel learning contexts in a way that elicits the same cognitive benefits as have been observed in classroom and lab settings.



**Figure 16.** Response by problem type: Breakdown of qCard interaction results across the two problem types. Best viewed in colour.

We evaluated an Adaptive Retrieval Practice System (ARPS) to address the emerging issue of supporting learning strategies at large scale (Davis et al., 2018) and to bridge retrieval practice theory into the digital learning space. As was the case with most MOOC studies in the past (Rosen et al., 2017; van der Zee et al., 2018; Davis, Chen, van der Zee et al., 2016; Davis et al., 2017; Kizilcec & Cohen, 2017; Kizilcec, Saltarelli et al., 2017; Yeomans & Reich, 2017; Gamage et al., 2017), we found non-compliance 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 learner achievement and engagement in the course.

Despite the lack of causal findings, the data collected from ARPS allowed us to gain multiple insights into the online learning process as it pertains to the persistence and transience of knowledge gains. By examining learner performance and engagement with the retrieval prompts, we were able to track changes in knowledge retention for the same concept over time and with the introduction of new course materials. We summarize our key findings:

- We find no evidence that encouraging retrieval practice raised learning, performance, or self-regulatory outcomes.
- Demographic factors play a significant role in learner compliance with retrieval cues, with older and more highly educated learners being most likely to engage.
- 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.
- Learner quiz performance deteriorates with the introduction of more course concepts/materials and the passing of time.
- Approximately two-thirds of course knowledge is retained over the long-term.

We observed an encouraging trend in learners’ 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 analysis (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. To this end, researchers will need to find better

ways of encouraging learners to use activities designed to support the learning process by designing interfaces that foster high levels of engagement.

## Declaration of Conflicting Interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

Dan Davis' research is supported by the Leiden Delft Erasmus Centre for Education and Learning.

## References

- Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research, 87*(3), 659–701. <http://dx.doi.org/10.3102/0034654316689306>
- Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91*(434), 444–455. <http://dx.doi.org/10.1080/01621459.1996.10476902>
- Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies, 4*(3), 114–123.
- Bouchet, F., Harley, J. M., & Azevedo, R. (2016). Can adaptive pedagogical agents' prompting strategies improve students' learning and self-regulation? In A. Micarelli, J. Stamper, & K. Panourgia (Eds.), *Proceedings of the 13th International Conference on Intelligent Tutoring Systems (ITS 2016)*, 7–10 June 2016, Zagreb, Croatia (Volume 9684, pp. 368–374). New York: Springer. [http://dx.doi.org/10.1007/978-3-319-39583-8\\_43](http://dx.doi.org/10.1007/978-3-319-39583-8_43)
- Brooke, J. (1996). SUS: A quick and dirty usability scale. *Usability Evaluation in Industry, 189*(194), 4–7.
- Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridge of optimal retention. *Psychological Science, 19*(11), 1095–1102.
- Clark, R. C., & Mayer, R. E. (2016). *E-learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning*. Hoboken, NJ: John Wiley & Sons.
- Custers, E. J. (2010). Long-term retention of basic science knowledge: A review study. *Advances in Health Sciences Education, 15*(1), 109–128.
- Davis, D., Chen, G., Hauff, C., & Houben, G.-J. (2016). Gauging MOOC learners' adherence to the designed learning path. In T. Barnes et al. (Eds.), *Proceedings of the 9th International Conference on Educational Data Mining (EDM2016)*, 29 June–2 July 2016, Raleigh, NC, USA (pp. 54–61). International Educational Data Mining Society.
- Davis, D., Chen, G., Hauff, C., & Houben, G.-J. (2018). Activating learning at scale: A review of innovations in online learning strategies. *Computers & Education*, submitted.
- Davis, D., Chen, G., van der Zee, T., Hauff, C., & Houben, G.-J. (2016). Retrieval practice and study planning in MOOCs: Exploring classroom-based self-regulated learning strategies at scale. *Proceedings of the 11th European Conference on Technology Enhanced Learning (EC-TEL 2016)*, 13–16 September 2016, Lyon, France (pp. 57–71). *Lecture Notes in Computer Science*. Springer.
- Davis, D., Jivet, I., Kizilcec, R. F., Chen, G., Hauff, C., & Houben, G.-J. (2017). Follow the successful crowd: Raising MOOC completion rates through social comparison at scale. *Proceedings of the 7th International Conference on Learning Analytics and Knowledge (LAK '17)*, 13–17 March 2017, Vancouver, BC, Canada (pp. 454–463). New York: ACM. <http://dx.doi.org/10.1145/3027385.3027411>
- Ebbinghaus, H. (1885). *Über das Gedächtnis: untersuchungen zur experimentellen psychologie*. Duncker & Humblot.
- Gamage, D., Whiting, M., Rajapakshe, T., Thilakarathne, H., Perera, I., & Fernando, S. (2017). Improving assessment on MOOCs through peer identification and aligned incentives. *Proceedings of the 4th ACM Conference on Learning @ Scale (L@S 2017)*, 20–21 April 2017, Cambridge, MA, USA (pp. 315–318). New York: ACM Press. <http://dx.doi.org/10.1145/3051457.3054013>
- Geigle, C., & Zhai, C. (2017). Modeling student behavior with two-layer hidden Markov models. *JEDM: Journal of Educational Data Mining, 9*(1), 1–24. <http://dx.doi.org/10.1145/3051457.3053986>
- Henry, L., Roediger, I., & Karpicke, J. D. (2016). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181–210. <http://dx.doi.org/10.1111/j.1745-6916.2006.00012.x>
- Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology, 101*(3), 621. <http://dx.doi.org/10.1037/a0015183>
- Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 331*(6018), 772–775. <http://dx.doi.org/10.1126/science.1199327>

- Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, 319(5865), 966–968. <http://dx.doi.org/10.1126/science.1152408>
- Kizilcec, R. F., & Brooks, C. (2017). Diverse big data and randomized field experiments in MOOCs. *Handbook of Learning Analytics*, 211–222. Society for Learning Analytics Research (SoLAR). <http://dx.doi.org/10.18608/hla17.018>
- Kizilcec, R. F., & Cohen, G. L. (2017). Eight-minute self-regulation intervention raises educational attainment at scale in individualist but not collectivist cultures. *Proceedings of the National Academy of Sciences*, 114(17), 4348–4353.
- Kizilcec, R. F., & Halawa, S. (2015). Attrition and achievement gaps in online learning. *Proceedings of the 2<sup>nd</sup> ACM Conference on Learning @ Scale (L@S 2015)*, 14–18 March 2015, Vancouver, BC, Canada (pp. 57–66). New York: ACM. <http://dx.doi.org/10.1145/2724660.2724680>
- Kizilcec, R. F., Pe´rez-Sanagust´ın, M., & Maldonado, J. J. (2017). Self-regulated learning strategies predict learner behavior and goal attainment in massive open online courses. *Computers & Education*, 104, 18–33. <http://dx.doi.org/10.1016/j.compedu.2016.10.001>
- Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. *Proceedings of the 3<sup>rd</sup> International Conference on Learning Analytics and Knowledge (LAK ’13)*, 8–12 April 2013, Leuven, Belgium (pp. 170–179). New York: ACM
- Kizilcec, R. F., Saltarelli, A. J., Reich, J., & Cohen, G. L. (2017). Closing global achievement gaps in MOOCs. *Science*, 355(6322), 251–252. <http://dx.doi.org/10.1126/science.aag2063>
- Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The knowledge–learning–instruction framework: Bridging the science–practice chasm to enhance robust student learning. *Cognitive Science*, 36(5), 757–798.
- Lindsey, R. V., Shroyer, J. D., Pashler, H., & Mozer, M. C. (2014). Improving students’ long-term knowledge retention through personalized review. *Psychological Science*, 25(3), 639–647.
- Miyamoto, Y. R., Coleman, C. A., Williams, J. J., Whitehill, J., Nesterko, S., & Reich, J. (2015). Beyond time-on-task: The relationship between spaced study and certification in MOOCs. *Journal of Learning Analytics*, 2(2), 47–69. <http://dx.doi.org/0.2139/ssrn.2547799>
- Richards, B. A., & Frankland, P. W. (2017). The persistence and transience of memory. *Neuron*, 94(6), 1071–1084. <http://dx.doi.org/10.1016/j.neuron.2017.04.037>
- Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, 15(1), 20–27. <http://dx.doi.org/10.1016/j.tics.2010.09.003>
- Roll, I., Alevin, V., McLaren, B. M., & Koedinger, K. R. (2011). Improving students’ help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction*, 21(2), 267–280. <http://dx.doi.org/10.1016/j.learninstruc.2010.07.004>
- Rosen, Y., Rushkin, I., Ang, A., Federicks, C., Tingley, D., & Blink, M. J. (2017). Designing adaptive assessments in MOOCs. *Proceedings of the 4<sup>th</sup> ACM Conference on Learning @ Scale (L@S 2017)*, 20–21 April 2017, Cambridge, MA, USA (pp. 233–236). New York: ACM Press. <http://dx.doi.org/10.1145/3051457.3053993>
- Sen, A. (1994). Human development index: Methodology and measurement. United Nations Development Programme.
- Settles, B., & Meeder, B. (2016). A trainable spaced repetition model for language learning. *Proceedings of the 54<sup>th</sup> Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, 7–12 August 2016, Berlin, Germany (Vol. 1: Long papers, pp. 1848–1858). Stroudsburg, PA: Association for Computational Linguistics.
- Streeter, M. (2015). Mixture modeling of individual learning curves. In O. C. Santos et al. (Eds.), *Proceedings of the 8<sup>th</sup> International Conference on Educational Data Mining (EDM2015)*, 26–29 June 2015, Madrid, Spain (pp. 45–52). International Educational Data Mining Society.
- van der Zee, T., Davis, D., Saab, N., Giesbers, B., Ginn, J., van der Sluis, F., ... Admiraal, W. (2018). Evaluating retrieval practice in a MOOC: How writing and reading summaries of videos affects student learning. *Proceedings of the 8<sup>th</sup> International Conference on Learning Analytics and Knowledge (LAK ’18)*, 5–9 March 2018, Sydney, NSW, Australia (pp. 216–225). New York: ACM. <http://dx.doi.org/10.1145/3170358.3170382>
- VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science*, 31(1), 3–62. <http://dx.doi.org/10.1080/03640210709336984>
- Wen, M., & Rose, C. P. (2014). Identifying latent study habits by mining learner behavior patterns in massive open online courses. *Proceedings of the 23<sup>rd</sup> ACM International Conference on Information and Knowledge Management (CIKM ’14)*, 3–7 November 2014, Shanghai, China (pp. 1983–1986). New York: ACM. <http://dx.doi.org/10.1145/2661829.2662033>
- Yeomans, M., & Reich, J. (2017). Planning prompts increase and forecast course completion in massive open online courses. *Proceedings of the 7<sup>th</sup> International Conference on Learning Analytics and Knowledge (LAK ’17)*, 13–17 March 2017, Vancouver, BC, Canada (pp. 464–473). New York: ACM. <http://dx.doi.org/10.1145/3027385.3027416>
- Zhao, Y., Davis, D., Chen, G., Lofi, C., Hauff, C., & Houben, G.-J. (2017). Certificate achievement unlocked: How does MOOC learners’ behaviour change? In *Adjunct Publication of the 25<sup>th</sup> Conference on User Modeling, Adaptation and*

*Personalization* (UMAP 2017), 9–12 July 2017, Bratislava, Slovakia (pp. 83–88). New York: ACM.  
<http://dx.doi.org/10.1145/3099023.3099063>