Using a Randomized Experiment to Compare Mastery Learning Thresholds

Jeffrey Matayoshi
McGraw Hill ALEKS
Irvine, CA, USA
jeffrey.matayoshi@mheducation.com

Eric Cosyn
McGraw Hill ALEKS
Irvine, CA, USA
eric.cosyn@mheducation.com

Hasan Uzun
McGraw Hill ALEKS
Irvine, CA, USA
hasan.uzun@mheducation.com

Eyad Kurd-Misto
McGraw Hill ALEKS
Irvine, CA, USA
eyad.kurd-misto@mheducation.com

Many modern adaptive learning and intelligent tutoring systems implement the principles of mastery learning, where a student must demonstrate mastery of core prerequisite material before working on subsequent content within the system. Typically in such cases, a set of rules or algorithms is used to determine if a student has sufficiently mastered the concepts in a topic. In a previous work, we used a quasi-experimental design to investigate the relationship between two different mastery learning thresholds and the forgetting of the learned material. As a follow-up to this initial study, in the present work, we analyze the results from a randomized experiment—or A/B test—directly comparing these two mastery learning thresholds. These latest results seemingly agree with those from our initial study, giving evidence for the validity of the conclusions from our original quasi-experiment. In particular, we find that although students who learn with the higher mastery threshold are less likely to forget the learned knowledge, over time this difference decreases. Additionally, we build on these analyses by looking at how the relationships between the mastery thresholds change based on other factors, such as the amount of struggle students experience while learning or the subject matter being covered.

Keywords: mastery learning, forgetting, intelligent tutoring system, randomized experiment

1. INTRODUCTION

Within a *mastery learning* framework, students must demonstrate proficiency with the core prerequisite material before moving on to learn subsequent content. First articulated by Benjamin Bloom (Bloom, 1968), the principles of mastery learning are implemented in many modern adaptive learning and intelligent tutoring systems. Typically in these systems, a set of rules or algorithms determines if a student has sufficiently grasped the core material in a topic. Perhaps the most noteworthy of these systems is Bayesian knowledge tracing (BKT) and its many derivatives (Baker et al., 2008; Corbett and Anderson, 1994; Pardos and Heffernan, 2011; Yudelson, 2016). Another common set of models is the factor analysis family—examples of which include

Learning Factors Analysis (LFA) (Cen et al., 2006) and Performance Factors Analysis (PFA) (Pavlik et al., 2009)—while simpler rules and heuristics, such as requiring students to correctly answer a certain number of questions in a row (Kelly et al., 2015), are also used.

A closely related and relevant subject—both within the education field and more broadly as part of psychology and cognitive science—is that of knowledge retention and forgetting. Specifically, the Ebbinghaus forgetting curve (Averell and Heathcote, 2011; Ebbinghaus, 1913) is a well-known model of how knowledge decays over time. Many studies have examined these curves in a variety of settings, including laboratory experiments (Hanley-Dunn and McIntosh, 1984; McBride and Dosher, 1997; Paivio and Smythe, 1971; Smith, 1979), classrooms (Agarwal et al., 2012; Barzagar Nazari and Ebersbach, 2019; Goossens et al., 2016), and adaptive learning and intelligent tutoring systems (Matayoshi et al., 2018; Matayoshi et al., 2019; Wang and Beck, 2012; Xiong and Beck, 2014; Xiong et al., 2013). Furthermore, other studies have shown that accounting for forgetting (Choffin et al., 2019; Matayoshi et al., 2021; Qiu et al., 2011; Wang and Heffernan, 2011) and having personalized interventions and review schedules (Lindsey et al., 2014; Pavlik and Anderson, 2008; Settles and Meeder, 2016; Tabibian et al., 2019; Xiong et al., 2015) can be beneficial for learning systems.

When implementing a mastery learning system, the chosen threshold must ensure that a student has sufficiently practiced a topic. However, there is a delicate balance at play, as students can easily be subjected to more practice and work than necessary, an issue that has been variously referred to as "over practice" (Cen et al., 2007) or "overlearning" (Rohrer and Taylor, 2006). With these issues in mind, previous works have looked in detail at mastery learning thresholds and how to optimize them for factors such as student learning efficiency (Bälter et al., 2018; Cen et al., 2007) and classification performance (Fancsali et al., 2013; Kelly et al., 2015). Additionally, it has been argued that the choice of data and the threshold used are more important than the specific type of model being applied (Pelánek and Řihák, 2017).

In the current study, we examine the relationship between different mastery thresholds and the long-term retention of the learned material; additionally, we analyze the frequencies at which students successfully learn topics using the different thresholds. This is a continuation of the work in Matayoshi et al. (2022), where we performed a quasi-experimental analysis comparing two different mastery thresholds used in the ALEKS adaptive learning system. In the current work, we further investigate the differences between the mastery thresholds by analyzing the results from a randomized experiment (or A/B test). This experiment has three main objectives. First, given the inherent limitations of quasi-experimental studies, we want to investigate if our previous results are consistent with those from a fully randomized experiment—such verification would give us more confidence in instituting changes to the way in which these thresholds are used within the ALEKS system. Second, such a result is of interest from a methodological standpoint, as it would validate the techniques used in the earlier observational study. Third, we would like to deepen our understanding by investigating how these relationships change based on other factors, such as the amount of struggle students experience while learning, the specific topic being worked on, or the subject area being studied. While the preliminary results of this experiment were first reported in Matayoshi et al. (2024), we have expanded on that initial work with a larger data set and additional analyses.

The outline of the current paper is as follows. After reviewing several previous studies that are relevant to our current work, we continue by giving a brief background of the ALEKS system in Section 3. Next, in Section 4 we summarize the results from Matayoshi et al. (2022), and we then follow with a description of our experimental setup in Section 5. Our first analysis is

presented in Section 6, where we compare and contrast how often topics are successfully learned under the different mastery thresholds. Then, in Section 7 we study the retention and forgetting of learned knowledge where, among other things, we compare the experimental data with that from the original study in Matayoshi et al. (2022). In Section 8, we then look at how the results change based on the amount of struggle experienced by students when learning a topic, while in Sections 9 and 10 we take a closer look at the subject matter and individual topics, respectively. Lastly, we finish with a discussion of these latest results and their potential implications for learning systems.

2. Related Works

As discussed in the introduction, an important goal of this study is to analyze and understand the relationship between different mastery thresholds and the long-term retention of knowledge in an adaptive learning system. While the study of mastery learning thresholds for learning systems is an active field of research, much of this work focuses on evaluating the predictive accuracy of student models (Pelánek and Řihák, 2017), as opposed to analyzing knowledge retention. That said, we are aware of a few studies that are highly relevant to our current work.

To start, Cen et al. (2007) compared two learning systems, where one determined mastery using a basic knowledge tracing model with hand-derived parameters, while the other applied an optimized Learning Factors Analysis (LFA) model. The subsequent performance of the two groups of students was then measured with a retention test two weeks after their learning was finished. On average, the optimized group spent less time learning, and while the average performance on the retention test was slightly lower for the optimized group, there was uncertainty with these measurements, as the difference was not statistically significant.

Next, the study performed by Kelly et al. (2015) compared different mastery thresholds and, in particular, looked at the differences when requiring either three or five consecutive correct answers to define mastery. While they observed large differences in performance based on these two mastery thresholds, one caveat is that the sample size was small, with only 56 total students meeting the mastery thresholds. Additionally, the study did not test long-term retention specifically, as a post-test measuring performance was given immediately after a student met the mastery threshold.

Other studies have looked at some of the factors associated with the long-term retention or forgetting of knowledge. Xiong et al. (2013) observed that the number of attempts a student took to demonstrate mastery was negatively correlated with retention one or two weeks post learning. Additionally, Matayoshi et al. (2019) found a similar relationship when comparing forgetting curves conditional on the number of learning events. However, in both studies the number of attempts was not manipulated experimentally and was instead a function of the existing definition of mastery for the learning system.

3. Background

In this section, we briefly discuss the aspects of the ALEKS system that are relevant for this study. To start, within the system a *topic* is a problem type that covers a discrete unit of an academic course. Each topic contains many examples that are known as *instances*, with these instances being carefully chosen so that they are equal in difficulty and cover similar content. Figure 1 contains a screen capture of an instance of the math topic "Introduction to solving

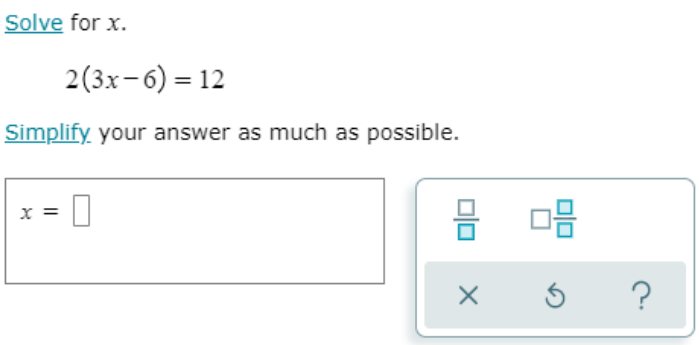


Figure 1: Screen capture of an instance of the ALEKS topic titled "Introduction to solving an equation with parentheses."

an equation with parentheses." Many *prerequisite* relationships exist between the topics in an ALEKS course. Specifically, we say that topic x is a prerequisite for topic y if x contains core material that must be learned before moving on to learn the material in y.

In order to ensure students are learning the most appropriate topics, an *initial assessment* is given at the start of an ALEKS course, with the purpose of this assessment being to measure the student's incoming knowledge. This assessment is adaptive in that it asks the student questions based on the responses to earlier questions in the assessment. After each question, for each topic in the course, the system estimates the probability that the student can answer the topic correctly (Matayoshi et al., 2022; Matayoshi and Cosyn, 2024). Then, at the very end of the assessment, based on both these probability estimates and the prerequisite relationships between the topics, the ALEKS system partitions the topics in the course into the following categories.

- Topics that are most likely known
- Topics that are most likely unknown
- All remaining topics (uncertain)

At this point, the student begins working in the ALEKS learning mode. Here, a student is presented with a topic that the system believes they are ready to learn. Additionally, the student can access a graphical list with additional topics that they are also ready to learn; however, students tend to work on the specific topic the system presents to them. The topics that are available to the student are from the unknown and uncertain categories, and they work on these one at a time, until they have either demonstrated a certain amount of mastery of the topic, or—in the event the student struggles to demonstrate this mastery—the system suggests they take a break and work on something else. To demonstrate mastery, two different thresholds—or rules—are used. The *high mastery* threshold is used for the unknown topics, while the *low mastery* threshold is used for the uncertain topics (we give precise definitions of these thresholds shortly). The idea is that, as the system does not have strong evidence that the uncertain topics are actually unknown, the student is given the benefit of the doubt, and a lower threshold is required for demonstrating mastery of these topics.

During the learning of a topic, three actions can be taken by a student: submitting a correct answer, submitting a wrong answer, or viewing an explanation page with a worked solution to

the instance. For a given topic, we define the *learning sequence* to be the sequence of actions taken by the student while working on the topic. A learning sequence for a topic starts with a score of 0. When the student first works on a topic, an example instance with a worked explanation is presented. Subsequent to this, the student receives another instance for actual practice. Whenever the student receives a new instance, they can try to answer it, or they can view the explanation page. A student is always given a new instance after a correct answer, viewing an explanation, or submitting two consecutive wrong answers.[1] Based on the student's action, the score is updated using the following rules.

(1) A single correct answer increases the score by 1; however, if the correct answer immediately follows a previous correct answer, the score increases by 2 instead of 1.

(2) An incorrect answer decreases the score by 1; however, the score does not change if it is already at 0, or it is the student's second attempt at the question following a first wrong answer.

(3) Viewing an explanation does not change the score. However, it does affect rule (1)—for example, if a student answers correctly immediately after viewing an explanation, the score increases by only 1 point, rather than 2, regardless of the student's previous responses.

If a topic is classified as unknown after the initial assessment, it uses the aforementioned high mastery threshold, in which case the student must demonstrate mastery by achieving a score of 5. For topics that are classified as uncertain after the initial assessment, a lower score of 3 is required to achieve mastery—this is the low mastery threshold. Lastly, in the event that a student gives five consecutive incorrect answers, this is considered to be a failed learning attempt, and the student is gently prompted to try another topic. These rules evolved to their current form over time based on empirical evidence, with the goal of maximizing the amount of learning without negatively impacting long-term retention. While their overall validity within the ALEKS system has been studied previously (Cosyn et al., 2021), the mastery threshold values have been investigated in isolation only recently (Matayoshi et al., 2022), and it is a goal of this experiment to understand their effects better.[2]

To test the retention of the topics after they are mastered, we make use of the ALEKS *progress assessment*. The progress assessment is a test given at regular intervals when a student has completed a certain amount of learning in the system. The purpose of the progress assessment is to focus on the student's recent learning, where it functions both as a way of confirming any recently learned knowledge, as well as a mechanism for spaced practice and retrieval practice. As spaced practice (Kang, 2016; Weinstein et al., 2018) and retrieval practice (Bae et al., 2019; Karpicke and Roediger, 2008; Roediger III and Butler, 2011; Roediger III and Karpicke, 2006a; Roediger III and Karpicke, 2006b) have been shown to help with the retention of knowledge, the progress assessment plays a key role within the ALEKS system (Matayoshi et al., 2021; Matayoshi et al., 2020). In order to evaluate student knowledge retention, we track how often students answer correctly to previously learned topics when they appear as an *extra*

[1] After a first wrong answer, the student gets a second chance to answer. If the second answer is again wrong, an explanation of the current instance is shown to the student before they are presented with a new instance to work on.

[2] Interestingly, it was shown in Doroudi (2020) that this relatively straightforward algorithm for determining mastery is theoretically equivalent to a simplified variant of BKT.

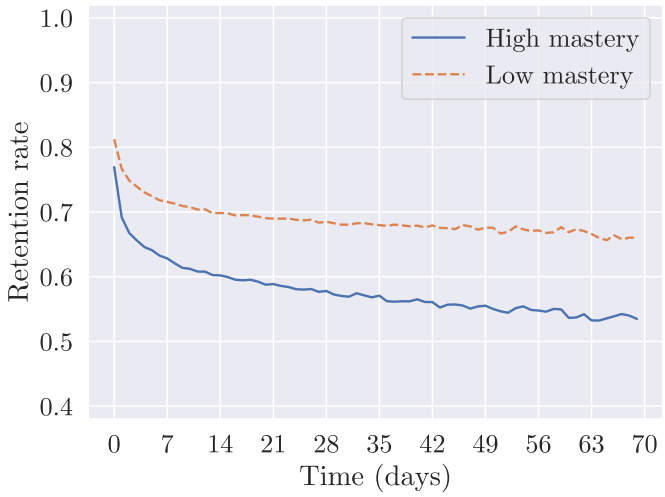


Figure 2: Forgetting curves comparing high mastery and low mastery topics based on the uncertain and unknown categories.

problem during the progress assessment. The extra problem is chosen by randomly selecting a topic, with this topic then being presented to the student as a regular question; however, the response to the question does not affect the results of the assessment. Instead, the data collected from these extra problems are used to evaluate and improve the ALEKS assessment. Thus, we define the *retention rate* to be the proportion of the time that students answer the extra problem correctly after having previously mastered the topic in the ALEKS system.

4. Previous Study: Adjusting for Selection Bias

A factor complicating our analysis is that there exists a selection bias in the assignment of the different mastery thresholds. That is, because of how the thresholds are assigned, topics using the high mastery threshold have lower probability estimates in comparison to topics that use the low mastery threshold—in general, this means that topics using the high mastery threshold tend to be more difficult. We can see this by using the data from our original study in Matayoshi et al. (2022) to look at the forgetting curves associated with each of the two categories. To generate these curves, we first find all examples where a topic was mastered before appearing as a question in the first progress assessment the student receives in the ALEKS system. Then, for each data point, we compute the time in days between the learning of the topic and its appearance on the progress assessment. Finally, we group the data points into bins of width one day, compute the correct answer rate within each bin, and plot the results.

From the forgetting curves in Figure 2, we can see that, overall, the correct rates for the topics with the low mastery threshold are noticeably higher. While this seems slightly confusing at first glance, as discussed in the previous paragraph, this is a byproduct of a selection bias. That is, the topics using the low mastery threshold, being from the uncertain category, are the ones for which the ALEKS assessment was not confident enough to classify as either known or unknown by the student—as such, it stands to reason that some proportion of these topics are likely known by the students, or that, at the very least, these topics tend to be easier for the

students to learn. In comparison, the topics that are classified as unknown by the ALEKS system are typically more difficult for the students.

Thus, while we wanted to investigate the relationship between these different amounts of practice and the retention of knowledge, due to the above issue, we did not have an accurate estimate of how large the differences could be. As such, our first step was to perform a quasi-experimental analysis—specifically, in Matayoshi et al. (2022) we used a regression discontinuity design (RDD) (Thistlethwaite and Campbell, 1960), a popular method that frequently appears in fields such as political science (Gelman et al., 2020) and econometrics (Angrist and Pischke, 2008), to analyze the differences in the mastery thresholds.

To run the RDD analysis, we leveraged the fact that a probability cutoff is employed by the ALEKS assessment to decide which mastery threshold a topic should use. The topics above this threshold are classified in the uncertain category and use the low mastery threshold; in comparison, topics below the threshold are classified in the unknown category and use the high mastery threshold. We then compared topics with probabilities close to the cutoff, in an attempt to measure the differences in retention, if any, between topics learned with the two different mastery thresholds. The results of this analysis—which are reproduced and discussed later in Section 7—suggested that the differences between the two mastery learning thresholds are not overly large.

Given that our earlier study was observational, there was some level of concern that the results might not be completely valid—for example, perhaps there were additional confounding variables that we failed to control for, or maybe the assumptions of the RDD were not completely satisfied. Furthermore, even if the RDD was completely valid, a basic limitation of the RDD procedure is that we only looked at topics near the probability cutoff; as such, it is possible that the results could change for a larger range of topics. Thus, for these reasons, we wanted to follow up on our earlier study with a fully randomized experiment, the details of which we describe in the next section.

5. Experimental Setup

Beginning in April 2023, for all ALEKS products, the assignment of the mastery threshold was altered according to the following procedure. Whenever a student starts learning a topic, 5% of the time the topic is randomly assigned the high mastery threshold, another 5% of the time the topic is randomly assigned the low mastery threshold, and the remaining 90% of the time there is no change—that is, the topic is not part of the experiment, and it instead uses the mastery threshold normally assigned by the system following the initial assessment. To run our analyses, we collected data from April 2023 through November 2024. Although we do not have access to demographic information on ALEKS users, overall, the program is used at a wide variety of colleges and K–12 schools, mainly in the U.S., with a total user base of over 7 million students. ALEKS products cover subjects such as mathematics, chemistry, and statistics, with mathematics being the most popular, followed by chemistry. Finally, appropriate consents are collected and notice provided to all our users via our Terms of Service and Privacy Notice, which specify the use of the anonymized data for product improvements and research purposes.

While our modeling procedure and analysis of the data closely follow the methodology used in Matayoshi et al. (2022), for completeness we next describe this methodology in detail. To compare the differences in retention between the mastery thresholds, we apply a linear regression; as our outcome variable is binary, this model is sometimes referred to as a linear probability

model. While using a generalized linear model—such as logistic regression—is usually recommended with a binary outcome variable, we opt for a linear regression here so that it is easier for us to interpret the coefficients. In theory, the use of a linear model with a binary outcome variable could lead to biased estimates; however, arguments have been made that this bias is typically low. In particular, Angrist and Pischke (2008) present theoretical and empirical arguments along these lines. Another potential weakness of a linear probability model, that it can lead to invalid probability estimates less than zero or greater than one, does not apply here. Previous works analyzing forgetting in the ALEKS system (Cosyn et al., 2021; Matayoshi et al., 2022; Matayoshi et al., 2018; Matayoshi et al., 2019; Matayoshi et al., 2020) showed the probability estimates of a correct answer to be bounded away from zero and one. That is, as these topics have been learned relatively recently, students should have a non-zero probability of answering correctly; at the same time, due to careless errors and slips it is unlikely they can answer correctly all the time, or even a large majority of the time, as these topics are typically on the edge of the student's current knowledge. In any event, fitting logistic regression models yielded results that are consistent with those from the linear regression models—for our main results from Section 7, comparisons with the corresponding logistic regression plots are included in the appendix.

To handle the fact that students can appear multiple times in our data, data points associated to the same student are considered a "group" or "cluster", and in each of our analyses we then fit a marginal model using a generalized estimating equation (GEE) (Hardin and Hilbe, 2012; Heagerty and Zeger, 2000; Liang and Zeger, 1986). GEE models are commonly applied in epidemiological studies and analyses containing repeated measurements—as such, they are well-suited for our current work. When using a GEE model, the type of correlation structure must be specified for the data within each group. An advantage of GEE models is that, even if this structure is misspecified, the parameter estimates are statistically consistent, and only the efficiency of these estimates is compromised (Hardin and Hilbe, 2012; Liang and Zeger, 1986). In all cases, we use an *exchangeable structure*, which assumes that there is some common dependence between all the data in a group (Hardin and Hilbe, 2012; Heagerty and Zeger, 2000; Szmaragd et al., 2013). All of these models are fit using the GEE class in the `statsmodels` (Seabold and Perktold, 2010) Python library.[3]

Next, to facilitate comparisons with the work in Matayoshi et al. (2022), we use the same predictor variables, defined as follows.

- x_1: 1 for high mastery; 0 for low mastery
- x_2: Initial assessment probability estimate
- x_3: Initial assessment score = (number of known topics) / (number of topics in course)
- x_4: Categorical variable encoding ALEKS product
- x_5: Categorical variable encoding first action in learning sequence (correct, incorrect, or explanation)
- x_6: Categorical variable encoding time (in weeks) since topic was learned (see Table 1)
- x_7: Interaction between mastery and time ($x_1 \times x_6$)

[3] Alternatively, we could use a mixed-effects model with a separate random intercept for each student. However, in the specific case of linear regression, such a formulation is equivalent to the GEE models we use here (Hardin and Hilbe, 2012).

Table 1: Categorical variable for time (x_6).

Category	Description
1	Less than 7 days after learning
2	Between 7 and 14 days after learning
⋮	
9	Between 56 and 63 days after learning
10	More than 63 days after learning

Our main focus is on the variables x_1 and x_7, as we are interested in estimating the average difference in retention between the groups using the mastery thresholds. The remaining predictors are control variables, as we attempt to adjust for factors such as the estimated difficulty of the topic (x_2), starting knowledge in the course (x_3), variation between students using the different ALEKS products (x_4), and initial amount of struggle experienced by the students while learning the topics (x_5).

As discussed in Matayoshi et al. (2022), the time since the topic was learned is technically a *post-treatment* variable—that is, it is measured after the "treatment" occurs, where the treatment corresponds to the successful learning of the topic with the high mastery threshold. When there is a suspected causal link between the post-treatment variable and the treatment, the estimate of the coefficient for the treatment variable could be biased by including the post-treatment variable in the regression (Acharya et al., 2016; Rosenbaum, 1984). Fortunately, because the extra problems are chosen randomly, we do not believe there is any reason to suspect a causal link between the time variable and the type of mastery threshold. Nonetheless, for our main regression results, which are presented in Section 7, we employ the following procedure to investigate this issue further. After first running our analysis including the categorical variable for time, we then re-run our analysis using the two-step regression procedure known as the *sequential g-estimator* (Joffe and Greene, 2009; Vansteelandt, 2009). This procedure allows us to make an estimate of β, the coefficient of the treatment, that adjusts for possible bias from the inclusion of the post-treatment variable (Acharya et al., 2016; Goetgeluk et al., 2008; Joffe and Greene, 2009; Vansteelandt, 2009; Vansteelandt et al., 2009). As with the results from our previous study, we do not see any substantial differences between the estimates using the sequential g-estimator and the estimates from our standard regression. Specifically, for the regression models in Section 7, the maximum difference between the coefficients from these results and the corresponding coefficients fit using the sequential g-estimator is less than 0.0008 in absolute value. Thus, to simplify the exposition, in the rest of this study we report only the results from the models fit without using the sequential g-estimator.

6. Comparing Learning Ratios

For our first analysis, we compare the mastery thresholds by measuring the difference in their *learning ratios*, which we define to be the proportions of topics worked on by students that are eventually mastered. To perform this analysis, we partition the data points based on their learning outcomes, using the following categories.

- Learn: topic successfully mastered

- Fail: topic failed by submitting five consecutive incorrect answers
- Incomplete: at least one answer submitted, but the topic is neither learned nor failed
- No response: an instance of the topic is viewed—and possibly an explanation page, as well—but no answers are submitted

An additional detail is that the student's score on a topic is reset whenever a progress assessment is taken. Because of this, we categorize the topics based on their status (a) at the time a progress assessment is taken and (b) at the end of our data collection period.[4] Using these categorizations, Table 2 shows the summary statistics for the two mastery thresholds, using a data set containing almost 33 million data points.[5] We can see that the high mastery threshold, with a learning ratio of 0.789, is more difficult in comparison to the low mastery threshold, which has a learning ratio of 0.809. Furthermore, the high mastery threshold has slightly larger proportions in the Fail, Incomplete, and No Response categories. Finally, Table 2 also shows the mean and median times for the learning sequences, with these times being larger for the high mastery threshold. Overall, these results are sensible, as we expect it to be more difficult and require more time for students to learn topics with the high mastery threshold. However, the differences are relatively small, suggesting that the mastery thresholds are perhaps not too different.

Table 2: Comparison of low mastery and high mastery groups. Sample sizes are shown in parentheses.

Mastery threshold	Outcome				Duration (minutes)	
	Learn	Fail	Inc.	No resp.	Mean	Median
High (16,441,259)	0.789	0.024	0.073	0.115	6.6	3.2
Low (16,474,845)	0.809	0.022	0.059	0.110	5.1	2.4

To investigate further, we divide the topics into groups based on the first action in each learning sequence. Specifically, in Table 3 we show the outcome statistics conditional on whether the student's first action was the submission of a correct answer (C*), the submission of a wrong answer (W*), or a viewing of the explanation page (E*). (Note that, in Table 3, we dismiss the learning sequences—less than 10% of the total—where the student viewed an example instance, but did not perform any further action.) The idea is that learning sequences starting with a correct answer are more likely to end successfully; in comparison, sequences starting with a wrong answer or a viewing of the explanation are less likely to be successful. Thus, we can compare the different mastery thresholds when students are struggling less (the C* sequences) and when they are struggling more (the W* and E* sequences). Across these different categories, we can see that the proportions vary greatly—for example, the learning ratios range from 0.757 to 0.977. However, as with the overall statistics, within each conditional category the differences between

[4]Note that a topic could appear multiple times for one student. For example, this happens if a student works on a topic, takes a progress assessment, and then returns to work on the same topic at a later point in time.

[5]Based on 5,000 cluster bootstrap samples—with the data from each student representing a single "cluster"—the 95% confidence intervals for the outcome proportions are all less than 0.0006 in width. Similarly, the 95% confidence intervals for the duration values are all less than 0.02 in width.

Table 3: Comparison of low mastery and high mastery groups, partitioned by the first action taken by the student. Sample sizes are shown in parentheses.

Mastery threshold		Outcome				Duration (minutes)	
		Learn	Fail	Inc.	No resp.	Mean	Median
C*	High (8,183,729)	0.962	0.006	0.031	0.0	5.5	3.0
	Low (8,103,625)	0.977	0.005	0.018	0.0	4.0	2.0
W*	High (4,647,850)	0.770	0.061	0.169	0.0	9.0	5.0
	Low (4,769,277)	0.799	0.057	0.145	0.0	7.3	4.1
E*	High (1,994,418)	0.757	0.029	0.079	0.135	10.5	5.5
	Low (2,062,859)	0.774	0.026	0.066	0.134	8.4	4.3

the high and low mastery topics are again relatively small but significant.[6] The largest differences are for the W* sequences where, presumably, students are more likely to be struggling with the material.

7. Forgetting and Retention Analysis

We now turn our focus to analyzing the forgetting and retention of the learned topics, with the goal of measuring and understanding the differences between the mastery thresholds. There are two main factors that give the high mastery threshold an advantage when comparing the retention of learned knowledge. First, given that the high mastery threshold requires a higher score before a topic is learned, this means students tend to get more practice on the topics, which should benefit retention—or, at the very least, should not harm retention. Next, in Section 6 we saw that the high mastery threshold has a slightly lower learning ratio when compared to the low mastery threshold, which introduces a selection bias when comparing the thresholds. Specifically, students who learn a topic with the high mastery threshold tend to be stronger students, or have a better understanding of the material; in both cases, we might again expect slightly higher retention rates for these students.

Attempting to precisely separate these issues when comparing the mastery thresholds would be a difficult task, and in what follows we do not attempt to do so. From a purely scientific perspective, this is suboptimal, as it would be interesting to precisely understand the relationship between the amount of practice and the retention of knowledge. However, we are still able to analyze the overall differences between the mastery learning thresholds, without worrying about which of the above factors specifically are responsible for these differences. Importantly, we believe that this approach is more useful for designing and improving adaptive learning and intelligent tutoring systems, as it follows the behavior of the mastery thresholds in an actual production system.

With these caveats in mind, we begin by extracting extra problem data from April 2023 through November 2024. After processing the data to remove any extra problems that are not part of the experiment, we are left with slightly less than two million data points. Next, because we want to include the student's performance on the initial assessment as one of our control

[6]Based on 5,000 cluster bootstrap samples, the 95% confidence intervals for the outcome proportions are all less than 0.0015 in width, while the 95% confidence intervals for the duration values are all less than 0.06 in width.

Table 4: Comparison of learning sequence statistics for topics in the high mastery and low mastery groups.

	High mastery (692,780)	Low mastery (711,945)
Correct answers	4.2 (3)	2.7 (2)
Wrong answers	2.2 (1)	1.8 (1)
Explanations	1.0 (0)	0.9 (0)
Total	7.4 (5)	5.4 (4)

variables, we remove students for whom we do not have initial assessment data.[7] This leaves us with 1,404,725 data points from 806,279 unique students.

7.1. Forgetting Curves

Using this set of data, in Table 4 we show statistics describing the differences in the learning sequences between the two mastery thresholds. In addition to the average number of actions of each type—correct answer, wrong answer, or viewing the explanation—we have also included the median values in parentheses. Overall, the learning sequences for the high mastery topics include about two extra learning actions, on average, with the majority of these extra actions being correct answers. Also, note that there are slightly more data points from the low mastery threshold; this is expected, as any successful learning sequence under the higher mastery threshold would succeed first under the low mastery threshold.

Next, in Figure 3 we show the forgetting curves for the two different mastery thresholds. To generate the curves, we group the data into bins based on the number of days between the time the topic was learned and its appearance as an extra problem. Next, for each bin, we compute the correct answer rate when the topic appears as the extra problem, and we then plot the resulting values to get the curves. While the curves start with a gap between them, this gap appears to decrease slightly as the time value increases.

7.2. Regression Analysis

To investigate the differences in retention further, our next step is to apply the regression analysis described in Section 5. The resulting coefficients are displayed in Figure 4a and, for comparison, the original results from the RDD analysis in Matayoshi et al. (2022) are then displayed in Figure 4b. In both plots, each (blue) dot shows the estimated average retention difference between the two mastery thresholds for the given time category, while the dashed lines show the 95% confidence interval for each point estimate. We submit that, overall, the results are roughly consistent between the two analyses. That is, the greatest estimated differences occur at the shorter time intervals, with the general trend being that these differences decrease as the time value increases. Furthermore, these estimated differences, the majority of which are less than

[7]While some of these initial assessments may be missing due to technical issues, the majority of the missing assessments are due to students being transferred between ALEKS courses—in many such cases, rather than being given an initial assessment, students are instead given credit for the topics they already demonstrated knowledge of in their previous course.

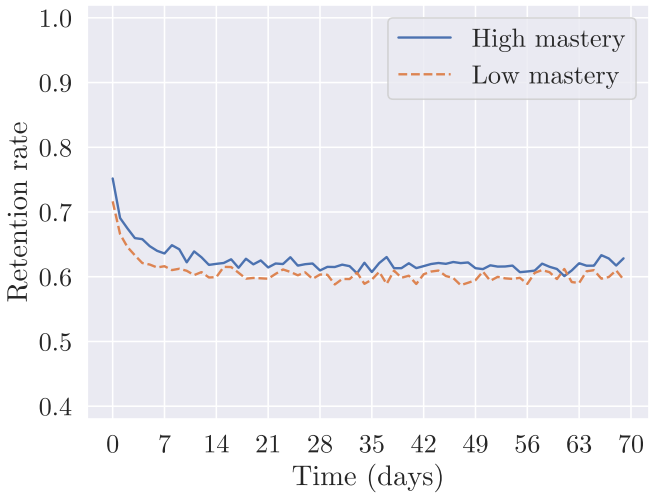


Figure 3: Forgetting curves comparing high mastery and low mastery topics from the randomized experiment.

0.02, are small in comparison to the overall retention rates shown in Figure 3. However, there are some contrasts in these trends, as the estimates in Figure 4a do not decrease quite as sharply, and they also appear to converge to a non-zero value; on the other hand, the original estimates in Figure 4b appear to be converging towards zero. As the ALEKS system has undergone modifications and improvements since the RDD analysis was performed, it is possible that some of these differences are due to these changes to the system.

7.3. Regression Analysis With Matched Data

We next run a type of matched analysis. Starting with the full set of 1,404,725 data points, we find all the students who have learned at least one topic each using the high mastery threshold and the low mastery threshold. Then, we use all of the data points from this group of students. After performing this procedure, we have 601,544 data points from a total of 186,734 unique students. Of these data points, 299,482 use the high mastery threshold, while 302,062 use the low mastery threshold.

The resulting coefficient estimates for the matched data are shown in Figure 4c, with the corresponding results from Matayoshi et al. (2022) displayed in Figure 4d. As with the full data set, we can see that the estimated differences are highest for the small time values, with these differences decreasing as the time value increases. Additionally, the estimated differences are once again relatively small, with most being around 0.02 or less in absolute value. Overall, the main features of the two plots appear to be similar.

7.4. Regression Analysis Using all Questions

While the plots from the regression analyses on the data from the randomized experiment—i.e., Figure 4a and Figure 4c—both show a decreasing trend, that trend is not as pronounced as in Matayoshi et al. (2022), as the plots tend to fluctuate up and down. Thus, we next attempt to investigate this trend further with a slight modification to our data set. Recall that, in our

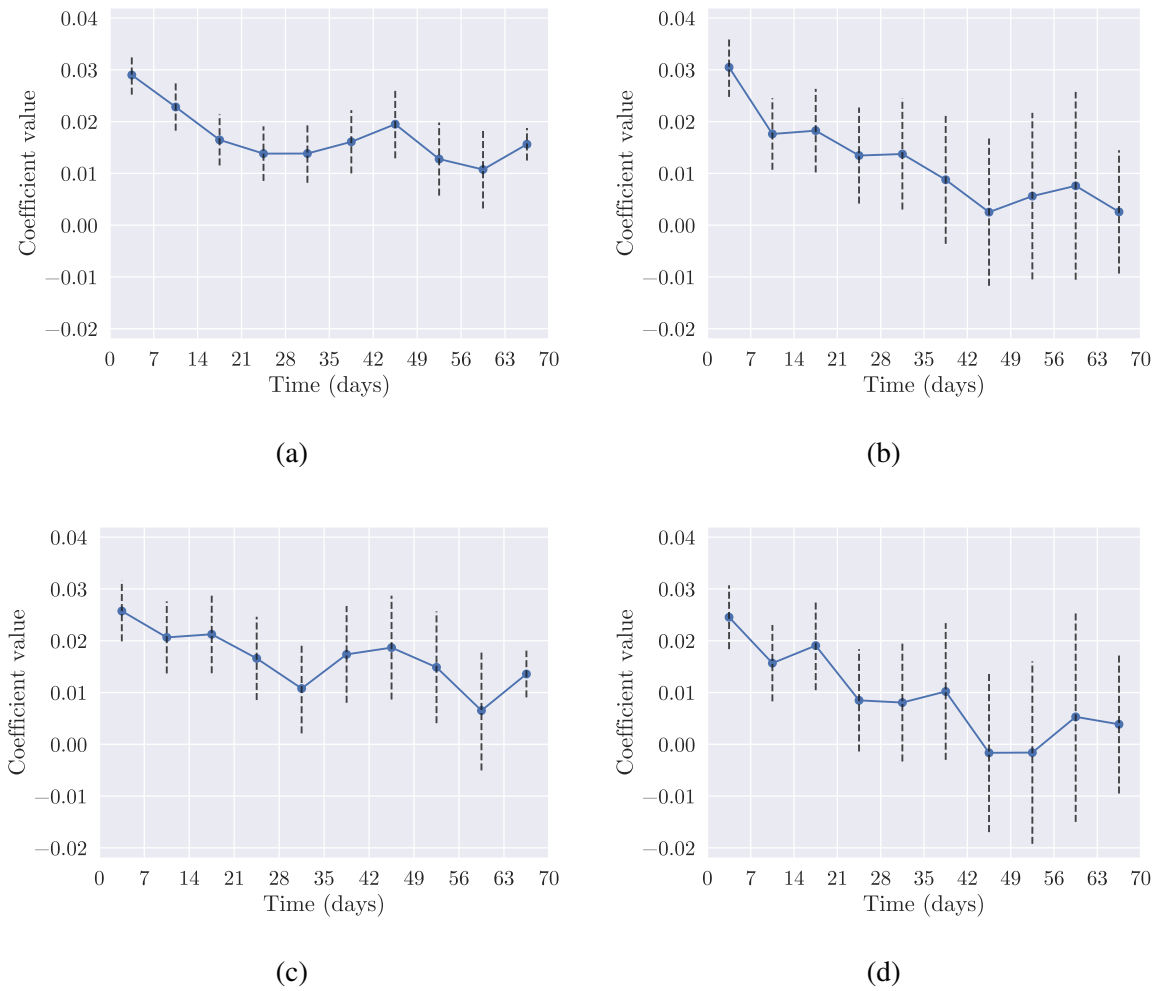


Figure 4: Coefficient estimates of the retention rate differences. The plot in (a) contains the estimates from the randomized experiment, while (b) contains the results from the observational study in Matayoshi et al. (2022). Then, (c) and (d) contain the corresponding results using the matched data; that is, (c) contains the estimates using the matched data from the randomized experiment, while (d) has the results from the observational study and matched data in Matayoshi et al. (2022).

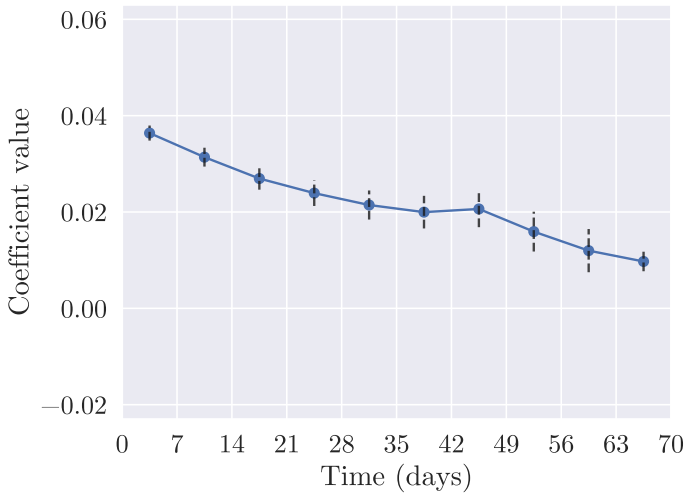


Figure 5: Coefficient estimates of the retention rate differences, using all questions from progress assessments—i.e., not just the randomly chosen extra problem—that are part of the experiment.

previous regression analyses, we used only progress assessment extra problems to check the retention of a learned topic. Using the extra problems gives a clean measure of retention, as these problems are chosen randomly and are not affected by the student's performance on the progress assessment. In comparison, the regular progress assessment questions are not only affected by the student's performance on that assessment, but also by their performance when learning the topic. Specifically, the progress assessment uses a neural network model to decide which topics to focus on (Matayoshi et al., 2019; Matayoshi et al., 2021), and it seems plausible that this could bias the resulting estimates of the retention differences between the mastery thresholds.

However, for this particular analysis, we are interested in the overall trend of the estimates, rather than their specific values. The advantage of adding the regular assessment questions is that this gives us a significantly larger amount of data, with the trade-off being the potential biases discussed in the previous paragraph. While we do not feel comfortable making strong inferences when adding the data from the regular assessment questions, we do believe there is still useful information to be gained. To that end, we compile a new data set, with this data set containing both the extra problems and the regular assessment questions. After applying the matching procedure from Section 7.3, the data set is composed of 5,740,871 data points from 952,806 unique students. The results are shown in Figure 5, where the strong downward trend of the estimates is consistent with, and yields further evidence for, the hypothesis that the difference in retention between the two mastery thresholds declines as the time values increase.

8. Retention and Student Struggle

In this section, we take a closer look at the relationship between the mastery thresholds and the student's first action in the learning sequence. Recall that the categorical variable x_5 encodes this information—that is, whether the student's learning sequence starts with a correct answer (C*), a wrong answer (W*), or a viewing of the explanation page (E*). Using our full set of 1,404,725

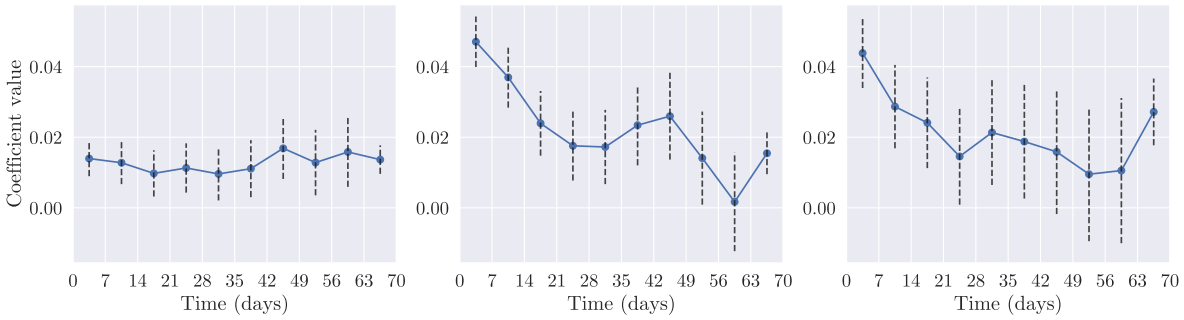
Table 5: Comparison of learning sequence statistics for topics in the high mastery and low mastery groups, partitioned by the first learning action.

		C*	W*	E*
High mastery	N	383,598	208,571	100,611
	Actions	5.1 (3)	10.1 (8)	10.5 (8)
	Retention	0.68	0.58	0.56
Low mastery	N	388,669	218,628	104,648
	Actions	3.3 (2)	7.8 (6)	8.3 (6)
	Retention	0.67	0.55	0.53

data points, partitioned by the mastery threshold and the first learning action, Table 5 shows the sample size for each category; the average total number of learning actions per learning sequence (with the median in parentheses); and the average retention rate, which is defined as the average correct answer rate when the topic later appears as the extra problem in a progress assessment.

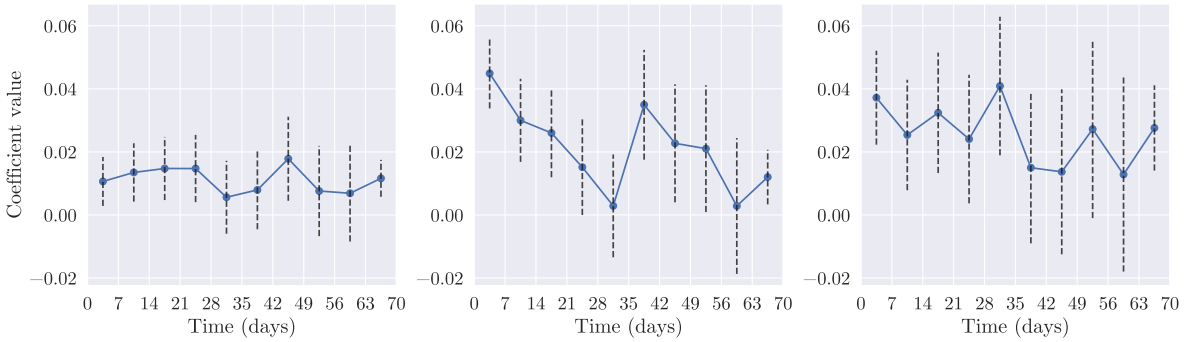
From these statistics, we can see that students with W* and E* learning sequences typically require more learning actions to master the topics, in comparison to the C* sequences. Additionally, the average retention rates are systematically lower for the W* and E* sequences, which means these topics are less likely to be answered correctly when they appear as an extra problem, again in comparison to the topics learned with C* sequences. Overall, this suggests that students with sequences of W* and E* tend to struggle more when learning the topics. While this makes sense for topics in the W* category, this is perhaps slightly surprising for the topics in the E* category. That is, it seems reasonable for some students to access the explanation not because they are struggling, but simply to perform their due diligence and prepare themselves fully before learning the topic. However, based on these statistics, such behavior does not appear to be the norm.

Next, to analyze the mastery thresholds for these different types of sequences, we use a more complex model with additional interactions between mastery and the student's first action ($x_1 \times x_5$); the student's first action and time ($x_5 \times x_6$); and mastery, time, and the student's first action ($x_1 \times x_5 \times x_6$). The results are shown in Figure 6, with each (sub)plot showing the estimated difference in retention for the specific subset of the learning sequences; similarly, in Figure 7 we show the analogous plot for the matched data. Comparing the plots across the different starting actions, it appears that the coefficient estimates are smallest for the C* sequences, indicating that the average difference in retention between the high and low mastery thresholds is smallest for this set of sequences; additionally, the coefficient estimates for the C* sequences are relatively consistent across the different time values. Contrast this with the estimates for the W* and E* sequences, which both start relatively high and then decrease as the time value increases. Thus, it is interesting and informative to see that the estimated differences are larger for the W* and E* groups, as this suggests that, at least initially, the extra practice from the higher mastery threshold has a larger effect for struggling students.



(a) C* (b) W* (c) E*

Figure 6: Coefficient estimates based on the student's first action.



(a) C* (b) W* (c) E*

Figure 7: Coefficient estimates based on the student's first action, using the matched data.

9. Subject Area Analysis

In this section, our goal is to check for possible differences in the results based on the subject area of the ALEKS course. To do this, we use the full set of extra problem data—i.e., not only the matched data—and group the courses into the following categories, where we have included the sample sizes in parentheses. (About 4% of our data are not contained in any of these categories and are excluded from this analysis.)

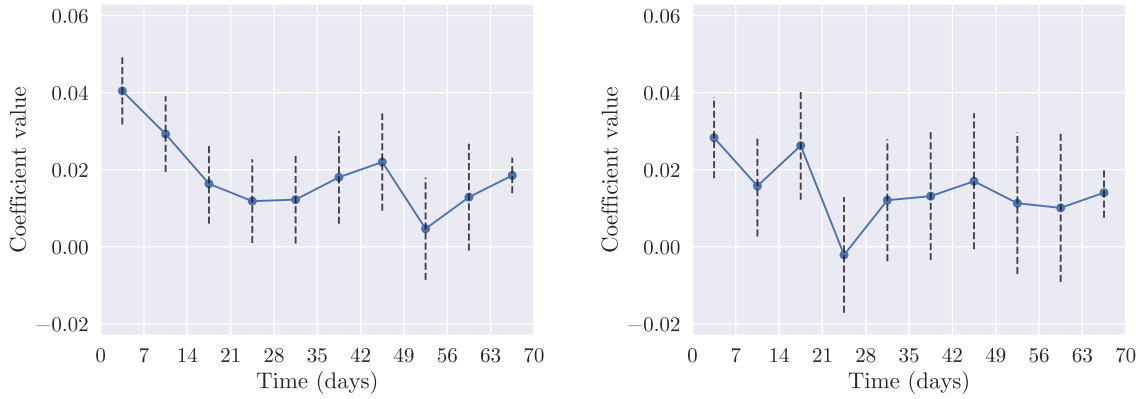
- Primary math (416,337): math courses from third grade through eighth grade
- High school math (221,136): high school math courses (e.g., Algebra 1, Geometry, etc.)
- College math (487,420): college math courses, starting at Basic Math and ending at Precalculus
- College chemistry (223,145): college chemistry courses (e.g., General Chemistry, Organic Chemistry, etc.)

For each category, we fit the regression model described in Section 5, with the results displayed in Figure 8. Interestingly, there do not appear to be any obvious patterns based on the subject matter or age of the students. For example, the primary math and college chemistry plots have similar trends—they both start around 0.04 and then decline as the time value increases—yet these student populations are very different in terms of both the age of the students and the subject matter of the courses they are taking.

10. Retention by Topic

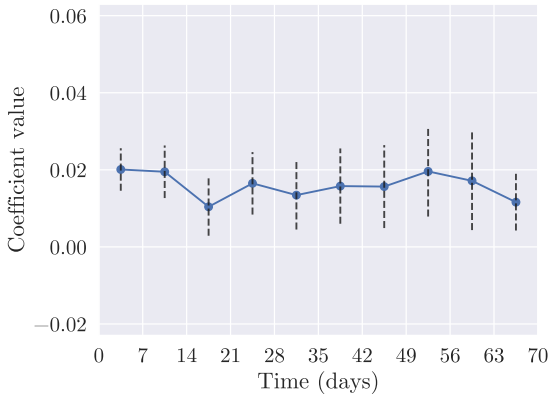
As the results in the previous section were somewhat inconclusive, we next take a slightly different approach and look at the retention differences between the mastery thresholds based on the specific topic. To start, for each topic we separate the data points depending on the mastery threshold used, and we then compute the retention rate for each group—that is, for each topic we have a retention rate when the low mastery threshold is used and a separate retention rate when the high mastery threshold is used. Then, for all of the items with at least 200 data points from each mastery threshold, we subtract the low mastery retention rate from the high mastery retention rate. The results for 1,376 topics with enough data points are shown in Figure 9a.

Notice there is a lot of noise in these values, as many topics have negative differences. This means that the retention rate with the higher mastery threshold is lower than with the lower mastery threshold, which seems implausible. That is, we have no reason to expect that more practice would result in a lower retention rate. Thus, to adjust for this, we next fit a multilevel model with the topics as the groups. For each topic, we include both a random intercept and a random slope, with the latter varying with the mastery threshold—this random slope will then give us an estimate of the retention difference for that topic depending on the mastery threshold. Multilevel models have an important property variously known as *shrinkage to the mean* (Snijders and Bosker, 2011) or *partial pooling* (Gelman et al., 2020). In our case, the estimates of the retention differences for the topics will be pulled towards the overall mean, with the size of this effect depending on the amount of data for the topic—generally, topics with more data will be less affected than those with little data. In our case, the expected benefit of this

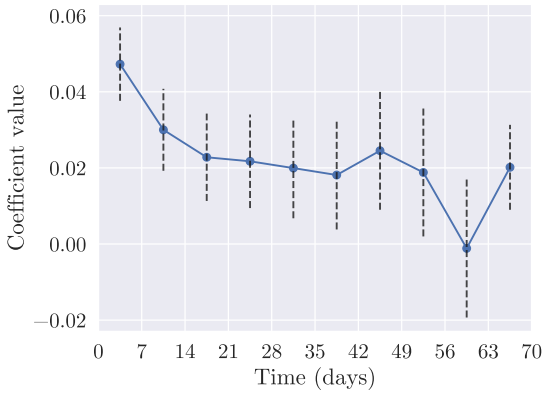


(a) Primary math

(b) High school math

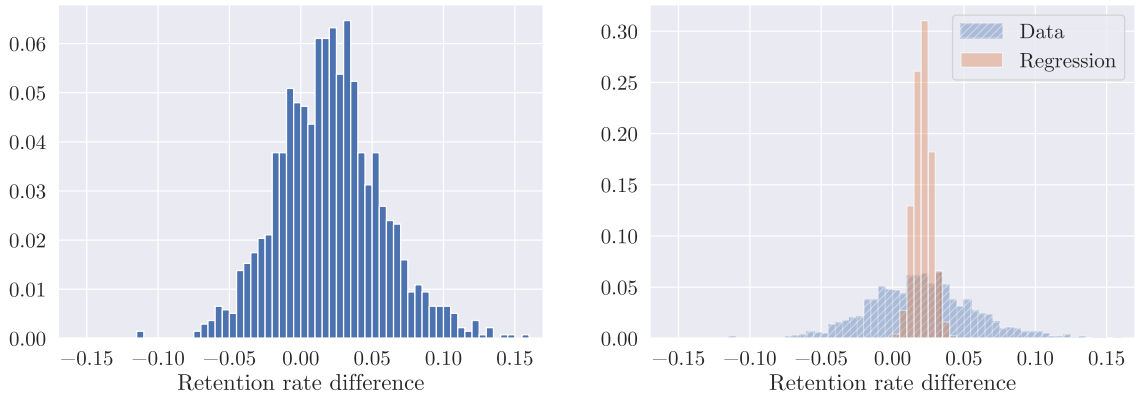


(c) College math



(d) College chemistry

Figure 8: Coefficient estimates of the retention rate differences for different course groupings.



(a) (b)

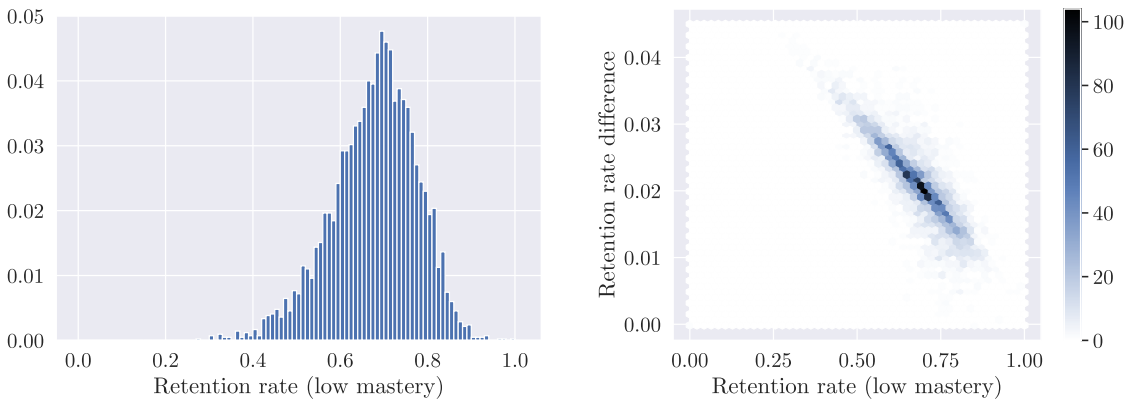
Figure 9: The relative frequency histogram in (a) shows the retention rate differences by topic. Each included topic has at least 200 data points from each mastery threshold. In (b), the estimated differences from the multilevel regression are also included.

effect is that it will adjust for the noisy values of the retention differences displayed in Figure 9a, hopefully leading to more reliable estimates of these differences.

The results from the multilevel model are shown in Figure 9b. The striped (blue) bars represent the original distribution from Figure 9a, while the solid (orange) bars show the distribution of the estimates from the multilevel model for the full set of 4,175 topics. The effect of the pooling is clear in the plot, as the distribution of the model estimates is much narrower and shifted to the right—in fact, there is only a single remaining negative estimate of -0.0002, with all the other estimates being positive.

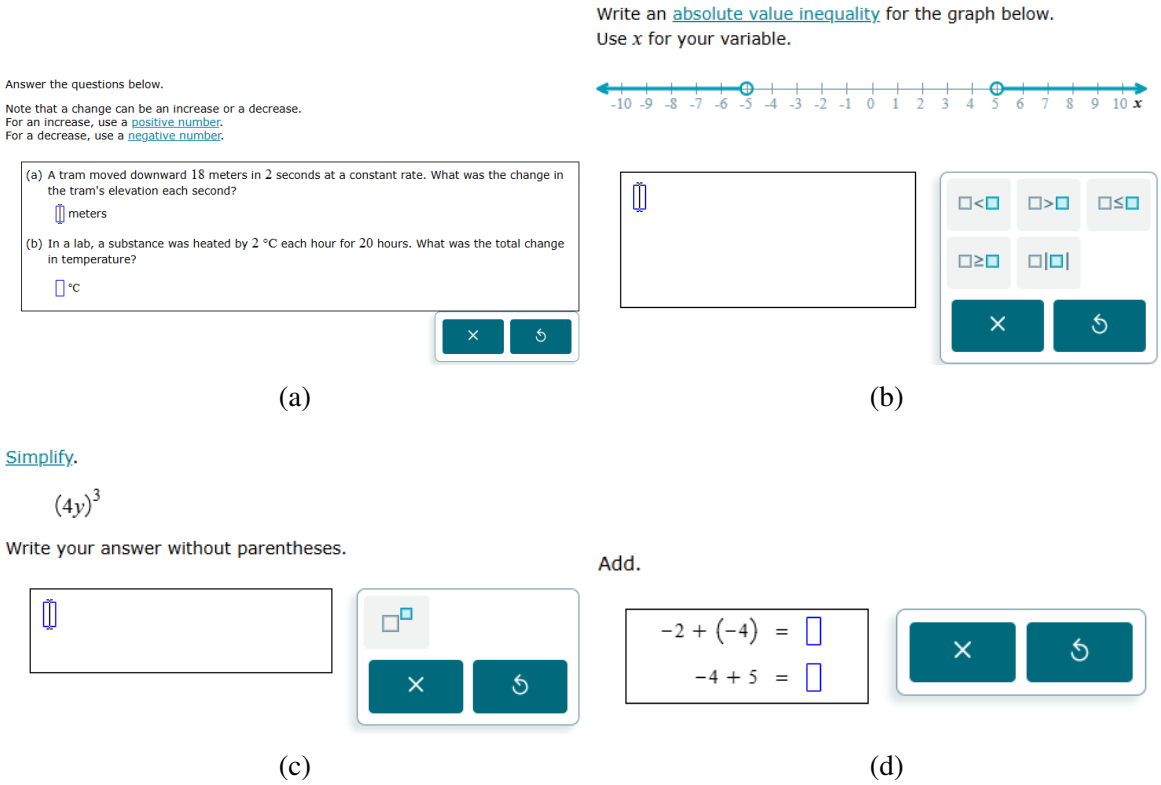
Based on the fact that the model estimates seem intuitively more reasonable, we next attempt to understand the results. The mean and median of the distribution from the multilevel model are both 0.021, with the 5th and 95th percentiles at 0.011 and 0.032, respectively. Thus, according to the multilevel model, there is relatively limited variation between the retention differences across the topics. This is somewhat surprising, as the overall topic retention rates vary quite widely based on the topic; this can be seen in Figure 10a, where we display the distribution of the topic retention rates for the low mastery threshold, as estimated by the multilevel model. Next, in Figure 10b we show a heatmap comparing the estimated retention rate gain from the high mastery threshold versus the estimated low mastery threshold retention rate for the topic. The Pearson correlation coefficient is -0.86, and the strong negative relationship is clear in the plot, where the items with high retention rates tend to improve less from the high mastery threshold. While it seems plausible that some of this is due to ceiling effects, with the high retention topics having less room to gain from the extra practice, it is also possible that there are other factors influencing this behavior.

To investigate this possibility further, we can compare and contrast example topics with extreme estimates for the retention rate difference. The first row of Figure 11—i.e., 11a and 11b—displays screen captures of topics with two of the highest estimated retention rate differences, while the second row—i.e., 11c and 11d—has screen captures of topics with two of the lowest differences. A seemingly significant distinction between the two rows is that topics



(a) (b)

Figure 10: The relative frequency histogram in (a) shows the multilevel model estimates of the topic-level retention rates for the low mastery threshold. Then, in (b), we show a heatmap of the estimated retention rate difference between the two mastery thresholds versus the estimated retention rate for the low mastery threshold. The Pearson correlation coefficient for the data in (b) is -0.86.



(a) (b)

(c) (d)

Figure 11: Screen captures of topics that have a relatively large estimated gain from the high mastery threshold—(a) and (b)—and topics that have very little estimated gain from the high mastery threshold—(c) and (d).

in the top row involve more complex material—this is supported by the fact that, in addition to having large estimated retention rate differences, they also have low overall retention rates. For example, properly understanding the topic in 11b requires a student to not only recognize that the direction of the inequality is determined by the placement of the line segment(s), but also to realize that open and closed endpoints represent different types of inequalities. Thus, it would seem reasonable that the extra practice afforded by the high mastery threshold could have a noticeable effect on the retention of the material. In comparison, the topics in the bottom row are "shallower" in nature and relatively more straightforward to understand, with high estimated retention rates; for example, the topic in 11c can be solved by memorizing a relatively simple formula. In any case, this discussion is fairly speculative, as we are looking at a small sample of topics; thus, understanding more about the relationship between the retention rate of a topic and the potential gain from the high mastery threshold is something we plan on studying further in future work.

11. Discussion

We began this work with three specific objectives. First, we wanted to validate the results from our quasi-experimental study in Matayoshi et al. (2022), where we analyzed the relationship between different mastery learning thresholds and the retention of learned knowledge. We found that, overall, the results from the current study's randomized experiment agree with those from our earlier analysis. Specifically, both works found that, while there appear to be differences in retention rates between the two mastery thresholds, these differences are relatively small, and they tend to decrease as the time increases between the learning of the topic and its eventual appearance as an assessment question. For example, based on the results in Figures 4a and 4c, after several weeks the estimated difference in retention rates between the two mastery thresholds is less than 0.02. Given that the average retention rates after several weeks are around 0.6—this can be seen in Figure 3—the difference in retention rates is small in comparison.

Our second objective was to demonstrate the utility of the methodology we applied in our previous observational study. From this standpoint, we find it encouraging that the results of the RDD analysis in Matayoshi et al. (2022) align with the experimental results from the current work. In addition to giving us more confidence in the techniques we employed, more generally, we also hope that, in some small part, this encourages other researchers in the field to employ RDD techniques. As many learning systems use cutoffs and thresholds to make decisions, running an RDD analysis could be a promising alternative when a randomized experiment is not feasible.

The third objective of this study was to deepen our understanding of the relationship between the mastery thresholds and the long-term retention of knowledge; additionally, we wanted to look at the learning ratio—i.e., the proportion of topics worked on by students that are eventually mastered—for each mastery threshold. For the latter, we saw that, while students have a higher learning ratio with the low mastery threshold in comparison to the high mastery threshold, this difference, about 0.02, is small in comparison to the overall learning ratio of 0.80 (Table 2). Moreover, similar results hold whether the students show signs of struggle (i.e., they start with a wrong answer or a viewing of the explanation) or answer correctly on their first try.

Regarding knowledge retention, we observed that the difference in the mastery thresholds varied based on the amount of learning struggle exhibited by students. For example, when students start with a correct answer, the retention difference between the thresholds is consistently

less than 0.02. In comparison, when students start with a wrong answer, the estimated difference is over 0.04 at its maximum—however, it is notable that this difference declines as the time value increases. Furthermore, we also detected a strong negative relationship between the overall retention rate of a topic and the mastery threshold difference (Figure 10b)—that is, more difficult items with lower retention rates seem to benefit more from the extra practice afforded by the high mastery threshold.

Recall that the results of several of our analyses suggest that the difference in retention declines over time. While it is not a given that the same pattern would appear on other data sets and with other learning systems, at the very least, we believe that when studying retention in relation to mastery thresholds, it is important to look at various time intervals to get a complete view of the possible effects. That is, while a difference in retention may look quite large initially, the results could change after a longer period of time has passed. Furthermore, we can give further context to our results by drawing on studies from the field of cognitive psychology. In particular, Rohrer and Taylor (2006) mention studies by Reynolds and Glaser (1964) and Rohrer et al. (2005) in which the benefits of extra learning practice weaken over time. Similar results appear in the review from Driskell et al. (1992), where it was found that for cognitive tasks, the positive effects of overlearning were estimated to disappear somewhere around 38 days. Thus, the results of our current study are seemingly consistent with these previous findings.

Lastly, we finish with a discussion of the potential benefits of these findings for the ALEKS system. With the goal of allowing students to learn topics more efficiently, there are a few ways in which the current use of the two mastery thresholds in the system could be modified. To start, consider that the estimated differences in retention between the two mastery thresholds are relatively small. It seems reasonable that the high mastery threshold could be used less often than it currently is—this could easily be implemented by adjusting the initial assessment to define fewer topics in the unknown category. As another example, Figures 6 and 7 indicate that the extra practice from the high mastery threshold is less beneficial for students starting with a correct answer (i.e., the C* sequences), which could be an indication that these students, in many cases, are already familiar with the material in the topic. Based on these results, one practical change to the system is to make it slightly easier for students to master a topic with the high mastery threshold if they start with a correct answer. For example, if a student is learning with the high mastery threshold, they could be allowed to pass if they start with two consecutive correct answers instead of three, as the high mastery threshold now requires. Finally, the results in Section 10 suggest that topics with lower retention rates tend to benefit more from the high mastery threshold—this information could be used to selectively assign the high mastery threshold to certain topics where the benefit would be greatest. Hopefully, changes such as these would help to further improve the learning experience for students working in the ALEKS system.

References

Acharya, A., Blackwell, M., and Sen, M. 2016. Explaining causal findings without bias: Detecting and assessing direct effects. *The American Political Science Review 110,* 3, 512.

Agarwal, P. K., Bain, P. M., and Chamberlain, R. W. 2012. The value of applied research: Retrieval practice improves classroom learning and recommendations from a teacher, a principal, and a scientist. *Educational Psychology Review 24*, 437–448.

Angrist, J. D. and Pischke, J.-S. 2008. *Mostly Harmless Econometrics*. Princeton University Press,

Princeton.

AVERELL, L. AND HEATHCOTE, A. 2011. The form of the forgetting curve and the fate of memories. *Journal of Mathematical Psychology 55*, 25–35.

BAE, C. L., THERRIAULT, D. J., AND REDIFER, J. L. 2019. Investigating the testing effect: Retrieval as a characteristic of effective study strategies. *Learning and Instruction 60*, 206–214.

BAKER, R. S. J. D., CORBETT, A. T., AND ALEVEN, V. 2008. More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian Knowledge Tracing. In *Intelligent Tutoring Systems*, B. P. Woolf, E. Aïmeur, R. Nkambou, and S. Lajoie, Eds. Springer Berlin Heidelberg, Berlin, Heidelberg, 406–415.

BÄLTER, O., ZIMMARO, D., AND THILLE, C. 2018. Estimating the minimum number of opportunities needed for all students to achieve predicted mastery. *Smart Learning Environments 5,* 1, 1–19.

BARZAGAR NAZARI, K. AND EBERSBACH, M. 2019. Distributing mathematical practice of third and seventh graders: Applicability of the spacing effect in the classroom. *Applied Cognitive Psychology 33,* 2, 288–298.

BLOOM, B. S. 1968. Learning for mastery. *Evaluation Comment 1,* 2.

CEN, H., KOEDINGER, K., AND JUNKER, B. 2006. Learning Factors Analysis–a general method for cognitive model evaluation and improvement. In *International Conference on Intelligent Tutoring Systems*. Springer Berlin Heidelberg, Berlin, Heidelberg, 164–175.

CEN, H., KOEDINGER, K. R., AND JUNKER, B. 2007. Is over practice necessary? Improving learning efficiency with the cognitive tutor through educational data mining. *Frontiers in artificial intelligence and applications 158*, 511.

CHOFFIN, B., POPINEAU, F., BOURDA, Y., AND VIE, J.-J. 2019. DAS3H: Modeling student learning and forgetting for optimally scheduling distributed practice of skills. In *Proceedings of the 12th International Conference on Educational Data Mining*. International Eduational Data Mining Society, 29–38.

CORBETT, A. T. AND ANDERSON, J. R. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction 4,* 4, 253–278.

COSYN, E., UZUN, H., DOBLE, C., AND MATAYOSHI, J. 2021. A practical perspective on knowledge space theory: ALEKS and its data. *Journal of Mathematical Psychology 101*, 102512.

DOROUDI, S. 2020. Mastery learning heuristics and their hidden models. In *International Conference on Artificial Intelligence in Education*. Springer International Publishing, Cham, 86–91.

DRISKELL, J. E., WILLIS, R. P., AND COPPER, C. 1992. Effect of overlearning on retention. *Journal of Applied Psychology 77,* 5, 615.

EBBINGHAUS, H. 1885; translated by Henry A. Ruger and Clara E. Bussenius (1913). *Memory: A Contribution to Experimental Psychology*. Originally published by Teachers College, Columbia University, New York.

FANCSALI, S., NIXON, T., AND RITTER, S. 2013. Optimal and worst-case performance of mastery learning assessment with Bayesian knowledge tracing. In *Proceedings of the 6th International Conference on Educational Data Mining*. International Eduational Data Mining Society, 35–42.

GELMAN, A., HILL, J., AND VEHTARI, A. 2020. *Regression and Other Stories*. Cambridge University Press, Cambridge.

GOETGELUK, S., VANSTEELANDT, S., AND GOETGHEBEUR, E. 2008. Estimation of controlled direct effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70,* 5, 1049–1066.

GOOSSENS, N. A., CAMP, G., VERKOEIJEN, P. P., TABBERS, H. K., BOUWMEESTER, S., AND ZWAAN, R. A. 2016. Distributed practice and retrieval practice in primary school vocabulary learning: A multi-classroom study. *Applied Cognitive Psychology 30,* 5, 700–712.

HANLEY-DUNN, P. AND MCINTOSH, J. L. 1984. Meaningfulness and recall of names by young and old adults. *Journal of Gerontology 39*, 583–585.

HARDIN, J. W. AND HILBE, J. M. 2012. *Generalized Estimating Equations*. Chapman and Hall/CRC, New York.

HEAGERTY, P. J. AND ZEGER, S. L. 2000. Marginalized multilevel models and likelihood inference (with comments and a rejoinder by the authors). *Statistical Science 15,* 1, 1–26.

JOFFE, M. M. AND GREENE, T. 2009. Related causal frameworks for surrogate outcomes. *Biometrics 65,* 2, 530–538.

KANG, S. H. 2016. Spaced repetition promotes efficient and effective learning: Policy implications for instruction. *Policy Insights from the Behavioral and Brain Sciences 3,* 1, 12–19.

KARPICKE, J. D. AND ROEDIGER, H. L. 2008. The critical importance of retrieval for learning. *Science 319,* 5865, 966–968.

KELLY, K., WANG, Y., THOMPSON, T., AND HEFFERNAN, N. 2015. Defining mastery: Knowledge tracing versus n-consecutive correct responses. In *Proceedings of the 8th International Conference on Educational Data Mining*. International Eduational Data Mining Society, 630–631.

LIANG, K.-Y. AND ZEGER, S. L. 1986. Longitudinal data analysis using generalized linear models. *Biometrika 73,* 1, 13–22.

LINDSEY, R. V., SHROYER, J. D., PASHLER, H., AND MOZER, M. C. 2014. Improving students long-term knowledge retention through personalized review. *Psychological Science 25,* 3, 639–647.

MATAYOSHI, J., COSYN, E., AND UZUN, H. 2021. Evaluating the impact of research-based updates to an adaptive learning system. In *International Conference on Artificial Intelligence in Education*. Springer International Publishing, Cham, 451–456.

MATAYOSHI, J., COSYN, E., AND UZUN, H. 2022. Does practice make perfect? Analyzing the relationship between higher mastery and forgetting in an adaptive learning system. In *Proceedings of the 15th International Conference on Educational Data Mining*, A. Mitrovic and N. Bosch, Eds. International Educational Data Mining Society, 316–324.

MATAYOSHI, J., COSYN, E., UZUN, H., AND KURD-MISTO, E. 2024. Going for the gold (standard): Validating a quasi-experimental study with a randomized experiment comparing mastery learning thresholds. In *Workshop on Causal Inference in Educational Data Mining, EDM 2024*.

MATAYOSHI, J., GRANZIOL, U., DOBLE, C., UZUN, H., AND COSYN, E. 2018. Forgetting curves and testing effect in an adaptive learning and assessment system. In *Proceedings of the 11th International Conference on Educational Data Mining*. International Eduational Data Mining Society, 607–612.

MATAYOSHI, J., UZUN, H., AND COSYN, E. 2019. Deep (un)learning: Using neural networks to model retention and forgetting in an adaptive learning system. In *International Conference on Artificial Intelligence in Education*. Springer International Publishing, Cham, 258–269.

MATAYOSHI, J., UZUN, H., AND COSYN, E. 2020. Studying retrieval practice in an intelligent tutoring system. In *Proceedings of the Seventh ACM Conference on Learning @ Scale*. Association for Computing Machinery, New York, NY, USA, 51–62.

MATAYOSHI, J., UZUN, H., AND COSYN, E. 2022. Using a randomized experiment to compare the performance of two adaptive assessment engines. In *Proceedings of the 15th International Conference on Educational Data Mining*. International Educational Data Mining Society, 821–827.

MATAYOSHI, J. S. AND COSYN, E. E. 2024. Neural network-based assessment engine for the determination of a knowledge state. US Patent App. 18/536,844.

MCBRIDE, D. M. AND DOSHER, B. A. 1997. A comparison of forgetting in an implicit and explicit memory task. *Journal of Experimental Psychology: General 126*, 371–392.

PAIVIO, A. AND SMYTHE, P. C. 1971. Word imagery, frequency, and meaningfulness in short-term memory. *Psychonomic Science 22*, 333–335.

PARDOS, Z. A. AND HEFFERNAN, N. T. 2011. KT-IDEM: Introducing item difficulty to the knowledge tracing model. In *Proceedings of the 19th International Conference on User Modeling, Adaption, and Personalization*. UMAP'11. Springer Berlin Heidelberg, Berlin, Heidelberg, 243–254.

PAVLIK, P. I. AND ANDERSON, J. R. 2008. Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied 14,* 2, 101.

PAVLIK, P. I., CEN, H., AND KOEDINGER, K. R. 2009. Performance Factors Analysis–a new alternative to knowledge tracing. In *Artificial Intelligence in Education-14th International Conference, AIED 2009*. IOS Press, NLD, 531–538.

PELÁNEK, R. AND ŘIHÁK, J. 2017. Experimental analysis of mastery learning criteria. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*. Association for Computing Machinery, New York, NY, USA, 156–163.

QIU, Y., QI, Y., LU, H., PARDOS, Z. A., AND HEFFERNAN, N. T. 2011. Does time matter? Modeling the effect of time with Bayesian knowledge tracing. In *Proceedings of the 4th International Conference on Educational Data Mining*. International Eduational Data Mining Society, 139–148.

REYNOLDS, J. H. AND GLASER, R. 1964. Effects of repetition and spaced review upon retention of a complex learning task. *Journal of Educational Psychology 55,* 5, 297.

ROEDIGER III, H. L. AND BUTLER, A. C. 2011. The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences 15*, 20–27.

ROEDIGER III, H. L. AND KARPICKE, J. D. 2006a. The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science 1,* 3, 181–210.

ROEDIGER III, H. L. AND KARPICKE, J. D. 2006b. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science 17,* 3, 249–255.

ROHRER, D. AND TAYLOR, K. 2006. The effects of overlearning and distributed practice on the retention of mathematics knowledge. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition 20,* 9, 1209–1224.

ROHRER, D., TAYLOR, K., PASHLER, H., WIXTED, J. T., AND CEPEDA, N. J. 2005. The effect of overlearning on long-term retention. *Applied Cognitive Psychology 19,* 3, 361–374.

ROSENBAUM, P. R. 1984. The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society: Series A (General) 147,* 5, 656–666.

SEABOLD, S. AND PERKTOLD, J. 2010. Statsmodels: Econometric and statistical modeling with Python. In *9th Python in Science Conference*. 92–96.

SETTLES, B. AND MEEDER, B. 2016. A trainable spaced repetition model for language learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 1848–1858.

SMITH, S. M. 1979. Remembering in and out of context. *Journal of Experimental Psychology: Human Learning and Memory 4*, 460–471.

SNIJDERS, T. AND BOSKER, R. 2011. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. SAGE Publications, London.

SZMARAGD, C., CLARKE, P., AND STEELE, F. 2013. Subject specific and population average models for binary longitudinal data: a tutorial. *Longitudinal and Life Course Studies 4,* 2, 147–165.

TABIBIAN, B., UPADHYAY, U., DE, A., ZAREZADE, A., SCHÖLKOPF, B., AND GOMEZ-RODRIGUEZ, M. 2019. Enhancing human learning via spaced repetition optimization. *Proceedings of the National Academy of Sciences 116,* 10, 3988–3993.

THISTLETHWAITE, D. L. AND CAMPBELL, D. T. 1960. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology 51,* 6, 309.

VANSTEELANDT, S. 2009. Estimating direct effects in cohort and case–control studies. *Epidemiology 20,* 6, 851–860.

VANSTEELANDT, S., GOETGELUK, S., LUTZ, S., WALDMAN, I., LYON, H., SCHADT, E. E., WEISS, S. T., AND LANGE, C. 2009. On the adjustment for covariates in genetic association analysis: A novel, simple principle to infer direct causal effects. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society 33,* 5, 394–405.

WANG, Y. AND BECK, J. E. 2012. Using student modeling to estimate student knowledge retention. In *Proceedings of the 5th International Conference on Educational Data Mining*. International Eduational Data Mining Society, 200–203.

WANG, Y. AND HEFFERNAN, N. T. 2011. Towards modeling forgetting and relearning in ITS: Preliminary analysis of ARRS data. In *Proceedings of the 4th International Conference on Educational Data Mining*. International Eduational Data Mining Society, 351–352.

WEINSTEIN, Y., MADAN, C. R., AND SUMERACKI, M. A. 2018. Teaching the science of learning. *Cognitive Research: Principles and Implications 3,* 1, 2.

XIONG, X. AND BECK, J. E. 2014. A study of exploring different schedules of spacing and retrieval interval on mathematics skills in ITS environment. In *International Conference on Intelligent Tutoring Systems*. Springer International Publishing, Cham, 504–509.

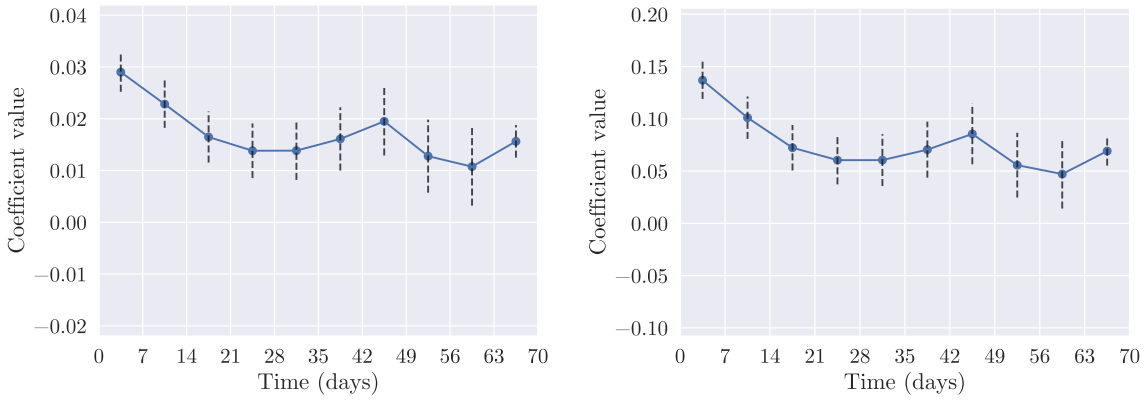
XIONG, X., LI, S., AND BECK, J. E. 2013. Will you get it right next week: Predict delayed performance in enhanced ITS mastery cycle. In *The Twenty-Sixth International FLAIRS Conference*. Association for the Advancement of Artificial Intelligence, 533–537.

XIONG, X., WANG, Y., AND BECK, J. B. 2015. Improving students' long-term retention performance: A study on personalized retention schedules. In *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*. Association for Computing Machinery, New York, NY, USA, 325–329.

YUDELSON, M. 2016. Individualizing Bayesian knowledge tracing. Are skill parameters more important than student parameters? In *Proceedings of the 9th International Conference on Educational Data Mining*. International Eduational Data Mining Society, 556–561.

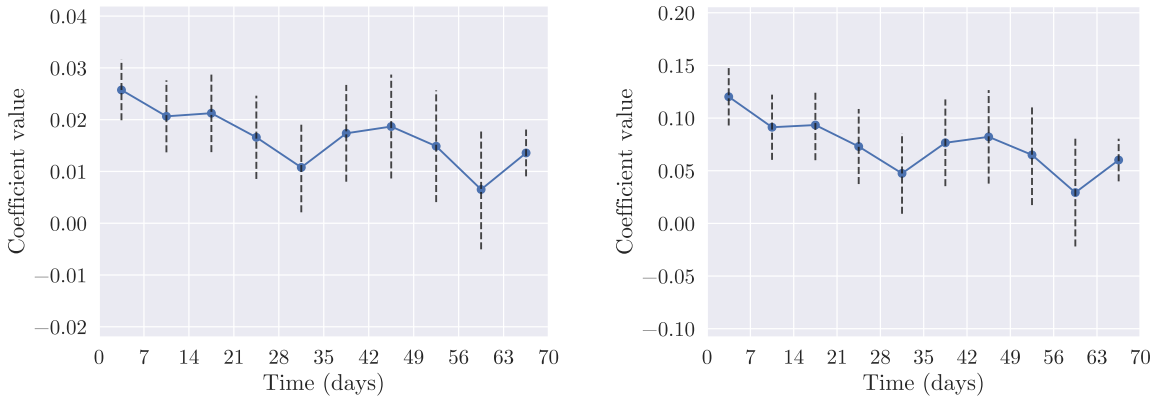
12. APPENDIX

As discussed in Section 5, we chose linear regression models to estimate the differences in retention, rather than logistic regression models, to simplify the interpretation of the coefficient estimates. To check our methodology, the main results from Section 7 are reproduced here and compared to the corresponding results from applying logistic regression models. These comparisons are reported in Figures 12–14. In all cases, the linear regression and logistic regression plots are almost identical.



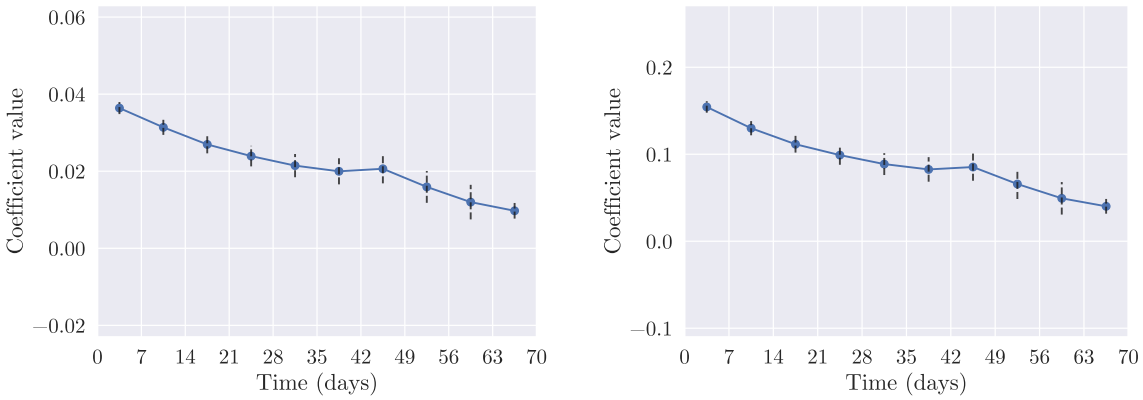
(a) (b)

Figure 12: Comparison of linear regression and logistic regression models using the full data set. The plot in (a) shows the linear regression results from Figure 4a, while the plot in (b) has the results from a logistic regression on the same data. Note the difference in the y-axis scales.



(a) (b)

Figure 13: Comparison of linear regression and logistic regression models using the matched data set. The plot in (a) shows the linear regression results from Figure 4c, while the plot in (b) has the results from a logistic regression on the same data. Note the difference in the y-axis scales.



(a) (b)

Figure 14: Comparison of linear regression and logistic regression models using all questions from progress assessments—i.e., not just the randomly chosen extra problem—that are part of the experiment. The plot in (a) shows the linear regression results from Figure 5, while the plot in (b) has the results from a logistic regression on the same data. Note the difference in the y-axis scales.