Going for the Gold (Standard): Validating a Quasi-Experimental Study With a Randomized Experiment Comparing Mastery Learning Thresholds

Jeffrey Matayoshi
McGraw Hill ALEKS
jeffrey.matayoshi
@mheducation.com

Eric Cosyn
McGraw Hill ALEKS
eric.cosyn
@mheducation.com

Hasan Uzun
McGraw Hill ALEKS
hasan.uzun
@mheducation.com

Eyad Kurd-Misto
McGraw Hill ALEKS
eyad.kurd-misto
@mheducation.com

ABSTRACT

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 our 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 the amount of struggle students experience while learning.

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 [8], many modern adaptive learning and intelligent tutoring systems implement the principles of mastery learning. Typically in these systems, a set of rules or algorithms determines if a student has sufficiently grasped the core material in a topic, with perhaps the most noteworthy being Bayesian knowledge tracing (BKT) and its many derivatives [6, 11, 36, 58]. Another common set of models is the factor analysis family—examples of which include Learning Factors Analysis (LFA) [9] and Performance Factors Analysis (PFA) [38]—while simpler rules and heuristics, such as requiring students to correctly answer a certain number of questions in a row [24], 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 [4, 14] 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 [18, 34, 35, 46], classrooms [2, 7, 17], and adaptive learning and intelligent tutoring systems [29, 30, 52, 55, 56]. Furthermore, other studies have shown that accounting for forgetting [10, 27, 39, 53] and having personalized interventions and review schedules [26, 37, 45, 48, 57] can be beneficial for learning systems.

In the current study, we examine the relationship between different mastery thresholds and the long-term retention of the learned material. This is a continuation of the work in [28], 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 multiple objectives. First, given the inherent limitations of quasi-experimental studies, we want to see 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. Additionally, such a result would be of interest from a methodological standpoint, as it would demonstrate the utility of the techniques used in [28]. Finally, we would like to deepen our understanding by investigating how these relationships change based on the amount of struggle students experience while learning.

The outline of the paper is as follows. We start by giving a brief background of the ALEKS system in Section 2. Next, in Section 3 we summarize the results from [28], and we then follow with a description of our experimental setup in Section 4. After presenting our first analysis in Section 5, where we compare the experimental data with that from the original study in [28], in Section 6 we then look at how the findings change based on the amount of struggle experienced by students when learning a topic. Lastly, we finish with a discussion of these latest results and their potential implications for learning systems.

2. 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 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 [32, 33]. 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 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

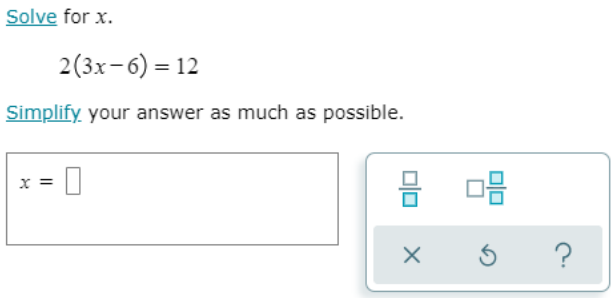


Figure 1: Screen capture of an ALEKS topic titled "Introduction to solving an equation with parentheses."

idea is that, as the system is not sure if the uncertain topics are actually known by the student, 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 (unless the score 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. Interestingly

[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.

enough, this mastery threshold, while arguably relatively straightforward, has been shown to have similarities with more sophisticated models, such as BKT [13]. 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.

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 [22, 54] and retrieval practice [5, 23, 40, 41, 42] have been shown to help with the retention of knowledge, the progress assessment plays a key role within the ALEKS system [27, 31]. In order to evaluate student knowledge retention, we can look to see how often students answer correctly to previously learned topics when they appear as an *extra 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.

3. PREVIOUS STUDY: ADJUSTING FOR SELECTION BIAS

A factor complicating our analysis is that there exists a selection bias with the assignment of the different mastery thresholds. That is, because of the way in which 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 [28] to look at the forgetting curves associated to 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

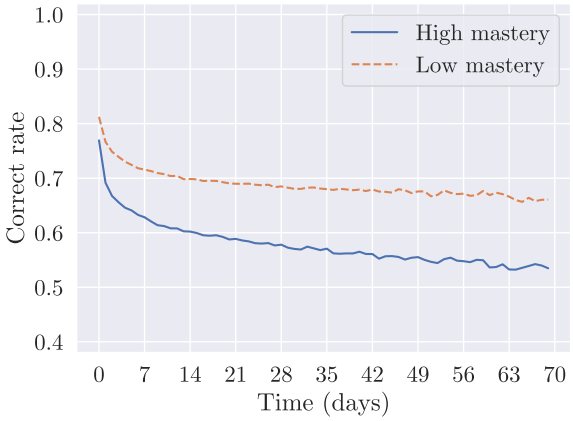


Figure 2: Forgetting curves comparing high mastery and low mastery topics based on the uncertain and unknown categories.

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 [28] we used a regression discontinuity design (RDD) [49], a popular method that frequently appears in fields such as political science [15] and econometrics [3], 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 5—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 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.

4. EXPERIMENTAL SETUP

Beginning in April 2023, across all ALEKS products we randomized the assignment of the different mastery thresholds to a small percentage of our users. Specifically, 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. 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.

To run our analysis, we extract extra problem data from April 2023 through May 2024. After processing the data to remove any extra problems that are not part of the experiment, we are left with slightly less than 1.4 million data points. Next, because we want to include the student's performance on the initial assessment as one of our control variables, we remove students for whom we do not have initial assessment data.[2] This leaves us with 1,003,696 data points from 548,028 unique students.

While our modeling procedure and analysis of the data closely follow the methodology used in [28], for completeness we next describe this methodology in detail. To compare the mastery thresholds, we apply a linear regression to estimate the average differences in retention between the mastery threshold groups; as our outcome variable is binary, this model is sometimes referred to as 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, [3] presents theoretical and empirical arguments along these lines. An additional criticism of the linear probability model is that estimating probability values near zero and one could be problematic, possibly leading to invalid probability estimates less than zero or greater than one. Nonetheless, based on previous works analyzing forgetting in the ALEKS system [12, 28, 29, 30, 31], we expect 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. Finally, as an additional check on this approach, we also fit logistic regression models and verify that the results are consistent with those from the linear regression models.

[2]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.

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

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) [19, 20, 25]. 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. In all cases, we use an *exchangeable structure*, which assumes that there is some common dependence between all the data in a group [19, 20, 47]. All of these models are fit using the GEE class in the `statsmodels` [44] Python library.[3]

Next, to facilitate comparisons with the work in [28], 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 topics classified as known) / (total 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 ($x_1 \times x_6$)

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).

[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 [19].

As discussed in [28], 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 [1, 43]. 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, we use 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* [21, 50]. 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 [1, 16, 21, 50, 51]. As with the results in [28], we do not see any substantial differences between the estimates using the sequential g-estimator and the estimates from our standard regression. As such, 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.

5. RESULTS

Using the full set of data from our randomized experiment—1,003,696 data points from 548,028 unique students—in Table 2 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, and vice versa regarding a failing sequence.

	High mastery (494,969)	Low mastery (508,727)
Correct answers	4.2 (3)	2.7 (2)
Wrong answers	2.3 (1)	1.9 (1)
Explanations	1.0 (0)	0.9 (0)
Total	7.5 (5)	5.5 (4)

Table 2: Comparison of learning sequence statistics for topics in the high mastery and low mastery groups. For each entry in the table, we show the average number of occurrences per sequence, with the corresponding median value in parentheses.

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 then we 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—however, it is difficult to tell for sure based only on the forgetting curves.

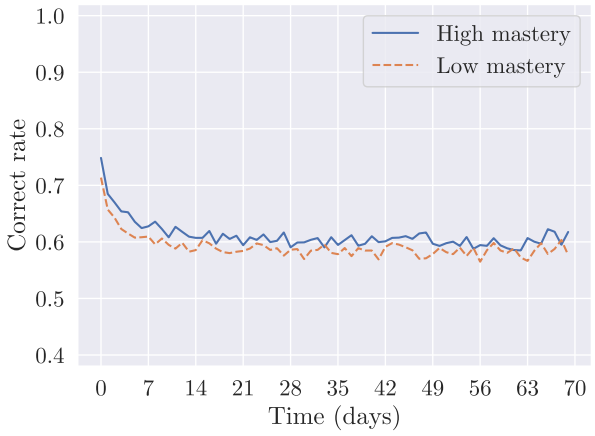


Figure 3: Forgetting curves comparing high mastery and low mastery topics from the randomized experiment.

Because of this, our next step is to apply the regression analysis described in Section 4. The resulting coefficients are displayed in Figure 4a and, for comparison, the original results from the regression discontinuity analysis in [28] 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, the overall differences are relatively small, with most of the estimated differences being less than 0.02. 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 study in [28], it is possible that some of these differences are due to these changes to the system.

We next run a type of matched analysis. Starting with the full set of 1,003,696 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 460,534 data points from a total of 140,686 unique students. Of these data points, 229,176 use the high mastery threshold, while 231,358 use the low mastery threshold.

The resulting coefficients estimates for the matched data are shown in Figure 4c, with the corresponding results from [28]

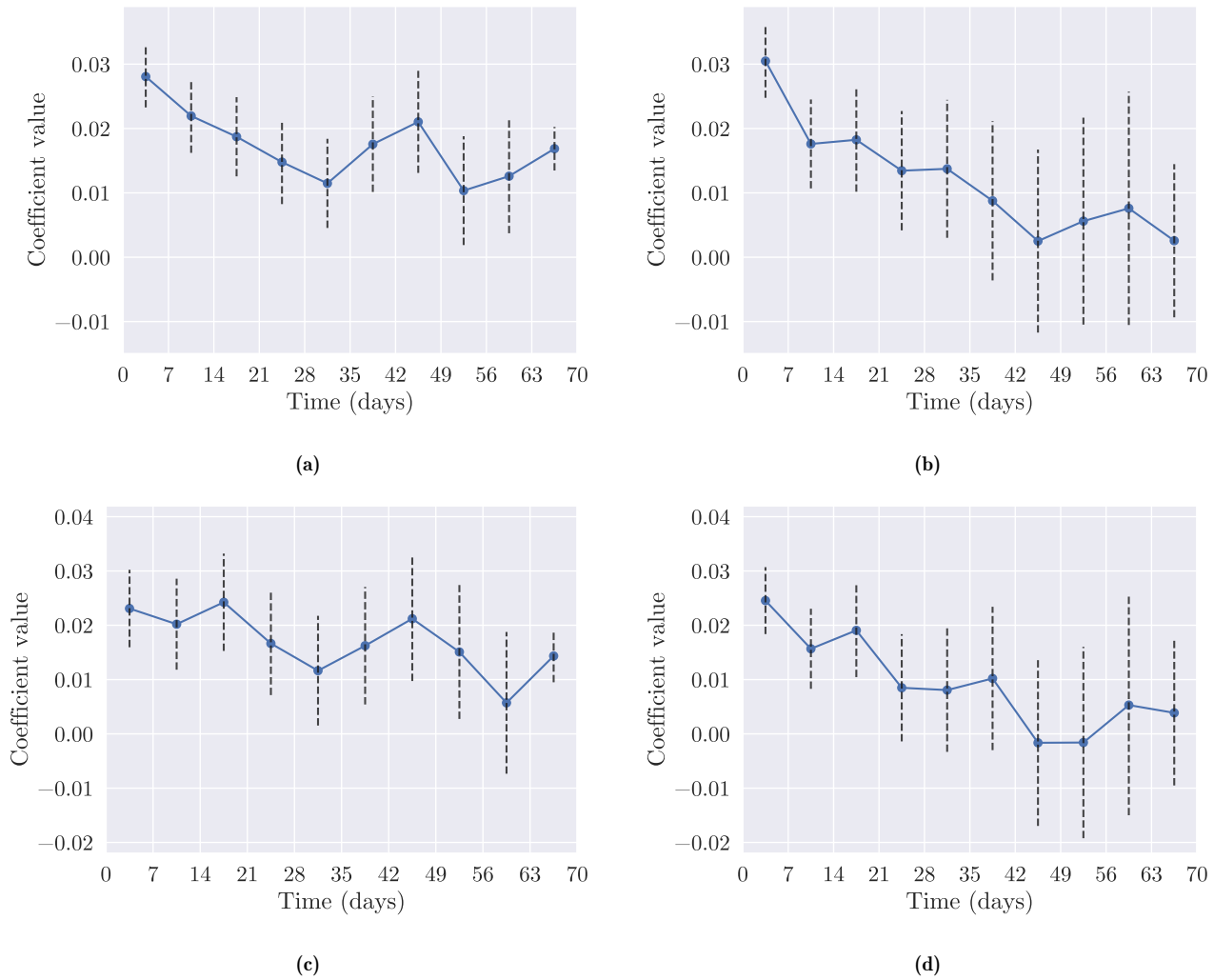


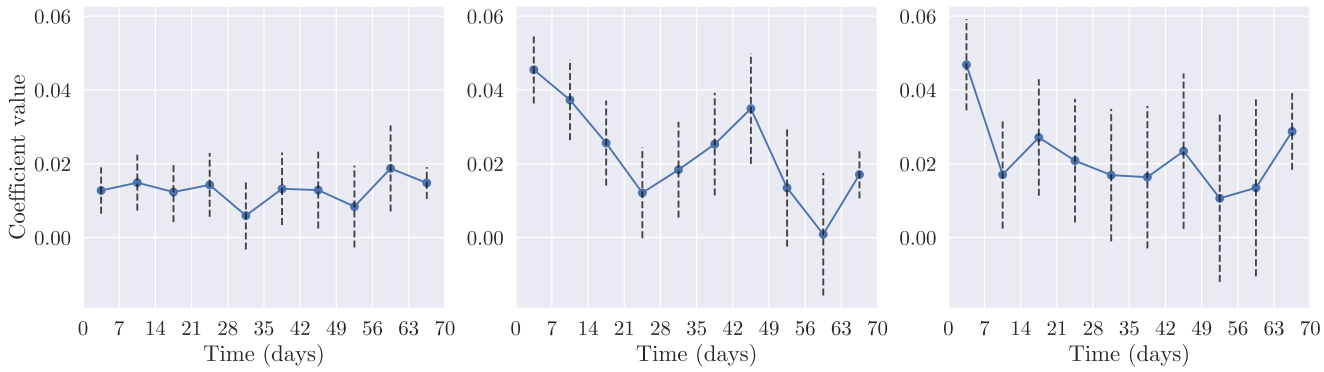
Figure 4: Coefficient estimates of the retention rate differences. The plot in (a) contains the estimates from the randomized experiment, while (b) contains the results from the regression discontinuity analysis in [28]. Then, (c) and (d) contain the corresponding results using the matched data; that is, (c) contains the estimates using the matched data from the randomized experiment, while (d) has the results from the regression discontinuity analysis and matched data in [28].

displayed in Figure 4d. As with the full dataset, 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.

6. 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 1,003,696 data points, partitioned by the mastery threshold and the first learning action, in Table 3 we show the average total number of learning actions per learning sequence, along with the average retention rate—in the latter case, this is 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 the 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,



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

Figure 5: Coefficient estimates based on the student's first action.

		C*	W*	E*
High mastery	N	279,808	144,621	70,540
	Actions	5.4 (3)	10.2 (8)	10.6 (8)
	Retention	0.67	0.57	0.54
Low mastery	N	283,346	151,785	73,596
	Actions	3.5 (2)	7.9 (6)	8.4 (6)
	Retention	0.65	0.55	0.52

Table 3: Comparison of learning sequence statistics for topics in the high mastery and low mastery groups, partitioned by the first learning action. The values in the table include the sample size for each category; the average number of actions per sequence (with the median in parentheses); and the average retention rate, which is defined as the correct answer rate when the topic appears as an extra problem.

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 5, with each (sub)plot showing the estimated difference in retention for the specific subset of the learning sequences. Comparing the plots, 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. In comparison, the estimates for both the W* and E* sequences start off 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.

7. DISCUSSION

In this work, we set out to validate the results from our quasi-experimental study in [28], 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 extra problem.

From a methodological standpoint, we find it encouraging that the results of the RDD analysis in [28] aligned with the experimental results from the current work. In addition to giving us more confidence in the techniques we employed in [28], 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.

We next discuss 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. As such, 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, Figure 5 indicates that the extra practice from the high mastery threshold is less beneficial for students starting with a correct answer (i.e., the C* sequences). Based on this, there is a possible argument for using the low mastery threshold for all such sequences; specifically, any learning sequence starting with a correct answer could automatically use the low mastery threshold. Finally, as a follow-up to the current work, we plan on looking even further into the data from the randomized experiment, to

see how the results might change based on other factors—namely, the specific subject area the topic comes from, or the grade level of the student. Hopefully, such additional insights would assist in further optimizing the learning experience for students working in the ALEKS system.

8. REFERENCES

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

[2] P. K. Agarwal, P. M. Bain, and R. W. Chamberlain. 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, 2012.

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

[4] L. Averell and A. Heathcote. The form of the forgetting curve and the fate of memories. *Journal of Mathematical Psychology*, 55:25–35, 2011.

[5] C. L. Bae, D. J. Therriault, and J. L. Redifer. Investigating the testing effect: Retrieval as a characteristic of effective study strategies. *Learning and Instruction*, 60:206–214, 2019.

[6] R. S. J. d. Baker, A. T. Corbett, and V. Aleven. More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian Knowledge Tracing. In *Intelligent Tutoring Systems*, pages 406–415. Springer Berlin Heidelberg, 2008.

[7] K. Barzagar Nazari and M. Ebersbach. Distributing mathematical practice of third and seventh graders: Applicability of the spacing effect in the classroom. *Applied Cognitive Psychology*, 33(2):288–298, 2019.

[8] B. S. Bloom. Learning for mastery. *Evaluation Comment*, 1(2), 1968.

[9] H. Cen, K. Koedinger, and B. Junker. Learning Factors Analysis–a general method for cognitive model evaluation and improvement. In *International Conference on Intelligent Tutoring Systems*, pages 164–175. Springer, 2006.

[10] B. Choffin, F. Popineau, Y. Bourda, and J.-J. Vie. DAS3H: Modeling student learning and forgetting for optimally scheduling distributed practice of skills. In *Proceedings of the 12th International Conference on Educational Data Mining*, pages 29–38, 2019.

[11] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278, 1994.

[12] E. Cosyn, H. Uzun, C. Doble, and J. Matayoshi. A practical perspective on knowledge space theory: ALEKS and its data. *Journal of Mathematical Psychology*, 101:102512, 2021.

[13] S. Doroudi. Mastery learning heuristics and their hidden models. In *Artificial Intelligence in Education*, pages 86–91. Springer International Publishing, 2020.

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

[15] A. Gelman, J. Hill, and A. Vehtari. *Regression and Other Stories*. Cambridge University Press, 2020.

[16] S. Goetgeluk, S. Vansteelandt, and E. Goetghebeur. Estimation of controlled direct effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):1049–1066, 2008.

[17] N. A. Goossens, G. Camp, P. P. Verkoeijen, H. K. Tabbers, S. Bouwmeester, and R. A. Zwaan. Distributed practice and retrieval practice in primary school vocabulary learning: A multi-classroom study. *Applied Cognitive Psychology*, 30(5):700–712, 2016.

[18] P. Hanley-Dunn and J. L. McIntosh. Meaningfulness and recall of names by young and old adults. *Journal of Gerontology*, 39:583–585, 1984.

[19] J. W. Hardin and J. M. Hilbe. *Generalized Estimating Equations*. Chapman and Hall/CRC, 2012.

[20] P. J. Heagerty and S. L. Zeger. Marginalized multilevel models and likelihood inference (with comments and a rejoinder by the authors). *Statistical Science*, 15(1):1–26, 2000.

[21] M. M. Joffe and T. Greene. Related causal frameworks for surrogate outcomes. *Biometrics*, 65(2):530–538, 2009.

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

[23] J. D. Karpicke and H. L. Roediger. The critical importance of retrieval for learning. *Science*, 319(5865):966–968, 2008.

[24] K. Kelly, Y. Wang, T. Thompson, and N. Heffernan. Defining mastery: Knowledge tracing versus n-consecutive correct responses. In *Proceedings of the 8th International Conference on Educational Data Mining*, pages 630–631, 2015.

[25] K.-Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

[26] R. V. Lindsey, J. D. Shroyer, H. Pashler, and M. C. Mozer. Improving students long-term knowledge retention through personalized review. *Psychological science*, 25(3):639–647, 2014.

[27] J. Matayoshi, E. Cosyn, and H. Uzun. Evaluating the impact of research-based updates to an adaptive learning system. In *International Conference on Artificial Intelligence in Education*, pages 451–456. Springer, 2021.

[28] J. Matayoshi, E. Cosyn, and H. Uzun. 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*. ERIC, 2022.

[29] J. Matayoshi, U. Granziol, C. Doble, H. Uzun, and E. Cosyn. Forgetting curves and testing effect in an adaptive learning and assessment system. In *Proceedings of the 11th International Conference on Educational Data Mining*, pages 607–612, 2018.

[30] J. Matayoshi, H. Uzun, and E. Cosyn. Deep (un)learning: Using neural networks to model retention and forgetting in an adaptive learning system. In *Artificial Intelligence in Education-20th International Conference, AIED 2019*, pages 258–269,

2019.

[31] J. Matayoshi, H. Uzun, and E. Cosyn. Studying retrieval practice in an intelligent tutoring system. In *Proceedings of the Seventh ACM Conference on Learning @ Scale*, pages 51–62, 2020.

[32] J. Matayoshi, H. Uzun, and E. Cosyn. Using a randomized experiment to compare the performance of two adaptive assessment engines. In *Proceedings of the 15th International Conference on Educational Data Mining*. ERIC, 2022.

[33] J. S. Matayoshi and E. E. Cosyn. Neural network-based assessment engine for the determination of a knowledge state, 2024. US Patent App. 18/536,844.

[34] D. M. McBride and B. A. Dosher. A comparison of forgetting in an implicit and explicit memory task. *Journal of Experimental Psychology: General*, 126:371–392, 1997.

[35] A. Paivio and P. C. Smythe. Word imagery, frequency, and meaningfulness in short-term memory. *Psychonomic Science*, 22:333–335, 1971.

[36] Z. A. Pardos and N. T. Heffernan. 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, pages 243–254. Springer-Verlag, 2011.

[37] P. I. Pavlik and J. R. Anderson. Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, 14(2):101, 2008.

[38] P. I. Pavlik, H. Cen, and K. R. Koedinger. Performance Factors Analysis–a new alternative to knowledge tracing. In *Artificial Intelligence in Education-14th International Conference, AIED 2009*, 2009.

[39] Y. Qiu, Y. Qi, H. Lu, Z. A. Pardos, and N. T. Heffernan. Does time matter? modeling the effect of time with Bayesian knowledge tracing. In *Proceedings of the 4th International Conference on Educational Data Mining*, pages 139–148, 2011.

[40] H. L. Roediger III and A. C. Butler. The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, 15:20–27, 2011.

[41] H. L. Roediger III and J. D. Karpicke. The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1(3):181–210, 2006.

[42] H. L. Roediger III and J. D. Karpicke. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3):249–255, 2006.

[43] P. R. Rosenbaum. 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, 1984.

[44] S. Seabold and J. Perktold. Statsmodels: Econometric and statistical modeling with Python. In *9th Python in Science Conference*, 2010.

[45] B. Settles and B. Meeder. 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)*, pages 1848–1858, 2016.

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

[47] C. Szmaragd, P. Clarke, and F. Steele. Subject specific and population average models for binary longitudinal data: a tutorial. *Longitudinal and Life Course Studies*, 4(2):147–165, 2013.

[48] B. Tabibian, U. Upadhyay, A. De, A. Zarezade, B. Schölkopf, and M. Gomez-Rodriguez. Enhancing human learning via spaced repetition optimization. *Proceedings of the National Academy of Sciences*, 116(10):3988–3993, 2019.

[49] D. L. Thistlethwaite and D. T. Campbell. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6):309, 1960.

[50] S. Vansteelandt. Estimating direct effects in cohort and case-control studies. *Epidemiology*, pages 851–860, 2009.

[51] S. Vansteelandt, S. Goetgeluk, S. Lutz, I. Waldman, H. Lyon, E. E. Schadt, S. T. Weiss, and C. Lange. 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, 2009.

[52] Y. Wang and J. E. Beck. Incorporating factors influencing knowledge retention into a student model. In *Proceedings of the 5th International Conference on Educational Data Mining*, 2012.

[53] Y. Wang and N. T. Heffernan. Towards modeling forgetting and relearning in ITS: Preliminary analysis of ARRS data. In *Proceedings of the 4th International Conference on Educational Data Mining*, pages 351–352, 2011.

[54] Y. Weinstein, C. R. Madan, and M. A. Sumeracki. Teaching the science of learning. *Cognitive Research: Principles and Implications*, 3(1):2, 2018.

[55] X. Xiong and J. E. Beck. A study of exploring different schedules of spacing and retrieval interval on mathematics skills in ITS environment. In *International Conference on Intelligent Tutoring Systems*, pages 504–509. Springer, 2014.

[56] X. Xiong, S. Li, and J. E. Beck. Will you get it right next week: Predict delayed performance in enhanced ITS mastery cycle. In *The Twenty-Sixth International FLAIRS Conference*, 2013.

[57] X. Xiong, Y. Wang, and J. B. Beck. Improving students' long-term retention performance: A study on personalized retention schedules. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, pages 325–329. ACM, 2015.

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