Does Practice Make Perfect? Analyzing the Relationship Between Higher Mastery and Forgetting in an Adaptive Learning System

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

ABSTRACT

As outlined by Benjamin Bloom, students working within a mastery learning framework must demonstrate mastery of the core prerequisite material before learning any subsequent material. Since many learning systems in use today adhere to these principles, an important component of such systems is the set of rules or algorithms that determine when a student has demonstrated mastery. A relevant issue when discussing mastery learning is its durability—in particular, we are interested in the relationship between different mastery learning thresholds and the forgetting of the learned material. As such, in this study we investigate this question using a large data set from the ALEKS adaptive learning system. Applying a quasi-experimental design, we find evidence that, while a higher mastery threshold is initially associated with a higher rate of knowledge retention, after several weeks this difference has largely disappeared.

Keywords

Mastery learning, forgetting, adaptive learning

1. INTRODUCTION

Many adaptive learning and intelligent tutoring systems in use today employ the principles of *mastery learning*. As outlined by Benjamin Bloom [9], in such a framework students must demonstrate mastery of the core prerequisite material before working on any subsequent material. Thus, an important component of any system implementing mastery learning is the set of rules or algorithms used to determine when a student has mastered a skill or problem type. Over the years, important families of models have been developed for this purpose, with perhaps the most noteworthy being Bayesian knowledge tracing (BKT) and its derivatives [6, 17, 43, 68], and the factors analysis family of models, with examples of the latter including Learning Factors Analysis (LFA) [13] and Performance Factors Analysis (PFA) [45]. Additionally, other simpler rules and heuristics, such as requiring students to correctly answer a certain number of questions in a row [32], are also utilized.[1]

As there is a balance between ensuring students have sufficiently mastered a problem type, while not subjecting them to more practice than necessary—variously referred to as "over practice" [14] or "overlearning" [51]—previous works have looked in detail at mastery learning thresholds and how to optimize them for various factors such as student learning efficiency [7, 14] and classification performance [22, 32]. 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 [46].

A related subject is that of knowledge retention and forgetting. In particular, the Ebbinghaus forgetting curve [4, 21] models the decay of knowledge over time, with numerous studies having looked at the conditions affecting these curves in settings as varied as laboratory experiments [26, 40, 42, 56], classrooms [2, 8, 25], and adaptive learning and intelligent tutoring systems [37, 38, 62, 65, 66]. Other works have shown that learning systems benefit greatly by accounting for forgetting [16, 35, 47, 63] and having personalized interventions and review schedules [34, 44, 55, 58, 67].

In this work, we are interested in the relationship between different mastery thresholds and the retention of knowledge. Additionally, we compare and contrast the frequencies at which problem types are successfully learned under these mastery thresholds. To perform these analyses, we take advantage of a "natural" experiment that occurs within the ALEKS adaptive learning system where, depending upon the outcome of an assessment given at the beginning of a course, problem types are assigned to two different mastery thresholds. By comparing the outcomes from these different thresholds, we hope to understand more about the relationship between higher mastery, extra practice, and forgetting.

2. BACKGROUND

In this section we give a brief background of the ALEKS system. Within the system, a *topic* is a problem type that covers a discrete unit of an academic course. Each topic contains many examples called *instances*, with these examples being chosen so that they cover the same content and are

[1]Interestingly, recent work has shown that some of these simpler models—including the one we consider in this study—can be viewed as special cases of BKT [19].

equal in difficulty. The topics in an ALEKS course contain many *prerequisite* relationships. That is, topic x is a prerequisite for topic y if x contains certain core concepts and material that must be learned before it's possible to learn the material in y.

At the start of an ALEKS course, the student's incoming knowledge is measured by an adaptive *initial assessment*. After each question of this assessment, a probability estimate is computed for each topic in the course, with this probability measuring how likely it is that the student knows the topic. At the 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 three categories.[2]

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

Next, in the ALEKS learning mode a student is presented a topic the system believes they are ready to learn, and the student can access additional topics they are ready to learn from a graphical list. In all cases, the topics being learned are from the uncertain and unknown categories. To determine mastery, a *high mastery* threshold is used for the unknown topics, while a *low mastery* threshold is used for the uncertain topics (we give precise definitions of these thresholds shortly). As the system is not sure if the uncertain topics are actually known by the student, relaxing the threshold allows the student to more quickly demonstrate mastery.

When learning a topic, a student can take three possible actions: submitting a correct answer, submitting a wrong answer, or accessing an explanation page with a worked solution to the current instance. We define the *learning sequence* to be the sequence of actions taken by the student while working on a particular topic. A student begins their learning sequence with a score of 0, whereupon they are presented an example instance with a worked explanation. Following this, the student receives another instance for actual practice. Each time the student receives a new instance, they can try to answer it, or they can access the explanation page. Note that a student is always given a new instance after a correct answer, viewing an explanation, or submitting two consecutive wrong answers. Depending on the student's action, the score is updated based on 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.
(2) An incorrect answer decreases the score by 1 (unless the score is already at 0).
(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.

[2]While beyond the scope of this study, the validity of both the probability estimates and the topic categorizations have been evaluated in works such as [18, 39].

For an unknown topic, the student must achieve a score of 5 before the topic is considered mastered—this is the aforementioned high mastery threshold. Alternatively, topics that are classified as uncertain only require a score of 3 to achieve mastery—this is the low mastery threshold. Finally, if a student gives five consecutive incorrect answers, this is considered to be a failed learning attempt—in such a case, the student is gently prompted to try another topic.

3. EXPERIMENTAL SETUP

Our study uses data from 13 different ALEKS mathematics products, ranging from elementary school mathematics to college-level algebra. From these products, we gathered data for a total of 2,235,061 students over a three-year period starting in January 2017. While we don't have access to detailed demographic information for our sample, we can say the majority of the K-12 students are from U.S. public schools, while the higher education products contain a mix of students from community colleges and four-year institutions, again mainly from the U.S.

To test the retention of the topics after they are mastered, we make use of the ALEKS *progress assessment*, an assessment given at regular intervals that is focused on the student's recent learning. The progress assessment plays a key role in the ALEKS system, as it enforces two learning strategies that have been shown to help with the retention of knowledge: spaced practice [30, 64] and retrieval practice [5, 31, 48, 49, 50]. To evaluate student knowledge retention, we define the *retention rate*—or, the *correct answer rate*—to be the proportion of the time that students answer topics correctly in the progress assessment after having mastered the topics in the ALEKS system.

Our analysis is complicated by a selection bias that exists with the assignment of the different mastery thresholds. 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, to compensate for this issue, we apply the elements of a regression discontinuity design (RDD) [59] to our analysis. RDD is a popular quasi-experimental design that is commonly used in fields such as econometrics [3] and political science [23]. The idea is that, given an experimental condition assigned by an arbitrary cutoff, it's plausible the data points close to this cutoff are similar, regardless of which side of the cutoff they ultimately fall on. In this study, we leverage the fact that a probability cutoff determines the assignment of the mastery threshold in the ALEKS system, with topics below the cutoff being assigned the high mastery threshold, while topics above the cutoff are assigned the low mastery threshold. By comparing topics with probabilities close to the cutoff, we hope to get accurate estimates of the differences between the two mastery thresholds.

In order to apply these ideas, we must account for the fact

that the ALEKS system also uses the information from the prerequisite relationships to assign the mastery threshold to topics. For example, suppose topic x is a prerequisite for topic y. Because of this relationship, if x is answered incorrectly during an ALEKS assessment, topic y will most likely be classified as unknown and thus given the high mastery threshold. Since this decision does not depend directly on the probability cutoff, we exclude these examples from all of our subsequent analyses, so that the probability cutoff is the sole determining factor in assigning the mastery threshold.

4. ANALYZING LEARNING RATIOS

Our first analysis attempts to quantify the differences between the mastery thresholds by comparing the *learning ratios*—that is, the proportions of topics worked on by students that are eventually mastered. As just discussed, we want to only look at data points which have the mastery threshold determined exclusively by the probabilities. Additionally, we restrict our analysis to learning data prior to a student's first progress assessment, as this assessment can alter the mastery threshold assigned to a topic, and we only select data points for which some work has taken place—for the latter, we minimally require that the student has at least looked at an example instance of the topic. Finally, out of the 2,154 total topics in our data set, we remove 7 that, due to technical issues, have systematic problems with their probability estimates. This leaves us with a total of 58,891,970 data points from 2,181,646 unique students.

Our next step is to select a reasonable *bandwidth* to conduct the RDD analysis, that is, a narrow interval around the probability cutoff to which the data points will be further restricted. While, all else equal, we want to have as much data as possible to work with, we also want our bandwidth to be narrow enough so that the included topics are expected to be similar in difficulty. We choose a bandwidth of 0.02 around the cutoff, which we believe works reasonably well at balancing these competing concerns. This leaves us with 1,949,102 data points from 984,138 unique students.

Table 1: Comparison of outcomes for the low mastery and high mastery groups.[3]

Mastery threshold	Learn	Fail	Inc.	No resp.
High (956,260)	0.847	0.057	0.062	0.034
Low (992,842)	0.865	0.052	0.050	0.033

For these data points, Table 1 shows the summary statistics after partitioning the learning outcomes into the following four categories.

- Learn: topic successfully mastered
- Fail: topic failed by submitting five consecutive incorrect answers
- Incomplete: at least one answer submitted, but 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

[3]Based on 10,000 cluster bootstrap samples—with the data from each student representing a single "cluster"—the 95% confidence intervals for the point estimates in Table 1 are all less than 0.002 in width.

Based on this partitioning, the learning ratio is simply the proportion of the outcomes in the Learn category. From Table 1 we can see that the topics in the high mastery group have a lower learning ratio in comparison to the low mastery topics—0.847 vs. 0.865, respectively. Furthermore, the high mastery group has larger proportions of incomplete and failed topics. Note that all of these results make intuitive sense—that is, all else being equal, for a given topic we expect it to be harder to learn under the high mastery threshold in comparison to the low mastery threshold.

One concern we have is that students may be actively seeking out one mastery threshold or the other, with perhaps the most prominent worry being that students would try to find the topics with the low mastery threshold, as this information is available to them. While our previous experience working with the ALEKS system has shown us that students mostly work on the specific topic the system presents to them, this is still worth investigating. As a start, we can look at the proportions in the No resp. column of Table 1—here, it's reassuring that these values are similar for the two different mastery thresholds.

Next, we can look at a *density* plot of the probabilities to see if there is an abrupt change as we move across the probability cutoff. Partitioning the interval $[-0.02, 0.02]$ into 100 bins of width 0.0004 each, in Figure 1 we plot the relative frequency (proportion) vs. the average distance from the threshold, based on the probabilities in each bin. While there's an increasing trend in the density as the x-values increase—which is a reflection of the distribution of the probabilities, rather than any particular student behavior—we are specifically interested in what happens around the probability cutoff, which is at 0 on the x-axis. As there doesn't appear to be clear evidence of a discontinuity—i.e., an abrupt increase or decrease—around the cutoff, we can use a *density test* to more precisely check for such a change [41]. Specifically, we apply the procedure outlined in [11], where the null hypothesis assumes there is no discontinuity in the density around the cutoff. Using the R implementation of this procedure in the `rddensity` package [12], the resulting p-value is 0.61—thus, the null hypothesis of no discontinuity is not rejected. Taking these results together, conservatively we can at least say there are no obvious signs of a bias from students electing to work on topics based on the mastery threshold.

5. FORGETTING AND RETENTION

In this next section, we attempt to estimate the differences in retention and forgetting between the topics that have been learned with the two mastery thresholds. While doing so, there are two important factors to consider. First, the students who learn with the high mastery threshold get more practice, as they tend to answer more questions in comparison to those who learn with the low mastery threshold. Second, as we saw previously the high mastery threshold is associated with a lower learning ratio in comparison to the low mastery threshold. This indicates there is a selection bias when we look at the students who learn a topic with the high threshold and compare them to students who learn with the low threshold—that is, the students who pass the high threshold tend to have a slightly better grasp of the material, or are perhaps slightly stronger students. Note that we'd expect both of these factors to benefit the knowledge

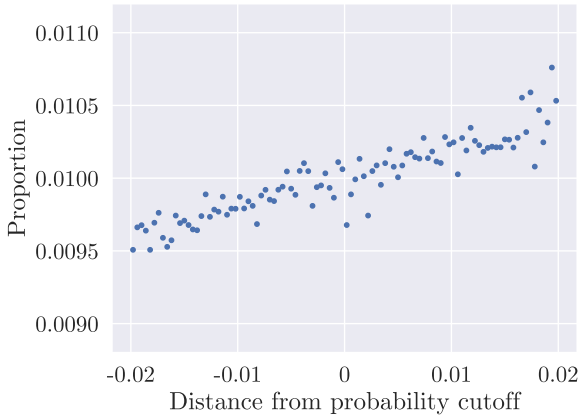

Figure 1: Density plot of the probability values, based on the distance from the probability cutoff. (More precisely, the x-axis reports the difference *topic probability − cutoff probability*.)

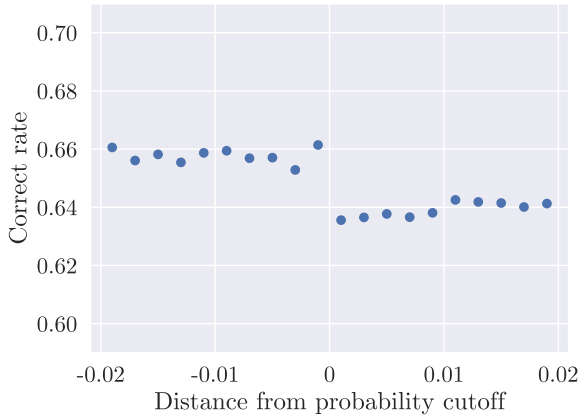

Figure 2: Retention (correct) rates based on the distance from the probability cutoff.

retention of the students who learn with the high threshold.

While it would be of interest to isolate these factors as much as possible, we won't be able to separate them in the analysis that follows, as they are directly connected to the specific mastery threshold used. On the one hand, from a purely scientific perspective this is unfortunate, as it would be of interest to, say, precisely compare the associations between the different amounts of practice. On the other hand, from a more practical viewpoint we can still analyze the overall differences between the mastery thresholds, which is arguably of more value for designing and improving adaptive learning and intelligent tutoring systems.

Table 2: Average number of actions per learning sequence—numbers in parentheses show the relative proportion of each action, based on the average total number of actions in the bottom row.

	High mastery (278,126)	Low mastery (302,191)
Correct answers	3.6 (0.66)	2.3 (0.61)
Wrong answers	1.3 (0.23)	1.0 (0.27)
Explanations	0.6 (0.10)	0.5 (0.12)
Total	5.4	3.8

Starting with the data set summarized in Table 1, we extract the subset of data points that (a) are successfully learned and (b) appear as questions in the student's first progress assessment—this leaves us with 580,317 data points from 436,735 unique students. In Table 2 we show the learning sequence statistics partitioned by the mastery threshold. The high mastery threshold topics have about 1.6 more learning events, with about 1.5 of these being either correct or wrong answers. Next, in Figure 2 we show a plot of the retention rates based on the distance from the probability cutoff, with the data points being divided into equal-width bins of 0.005, starting at -0.02 and ending at 0.02. For each bin the y-value represents the average correct answer rate when a topic appears in a student's first progress assessment, while the x-value is the average distance from the probability cut-

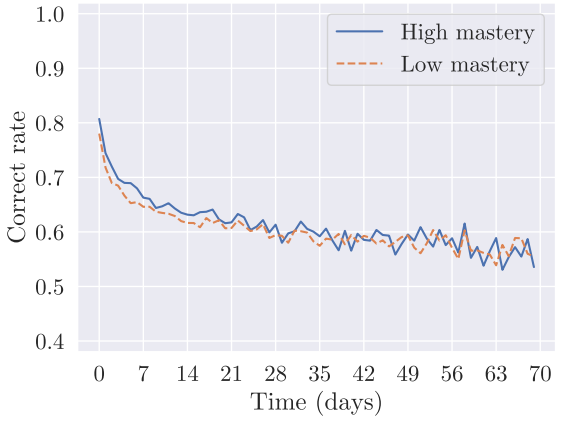

Figure 3: Mastery threshold forgetting curves.

off. Note there is a clear drop in retention as we move across the probability cutoff to the low mastery threshold topics. However, as we are studying the retention of knowledge, one factor we haven't taken into account is time. In particular, we next look at the forgetting curves for these data to see how these relationships might change at different time scales.

To generate these curves, for each data point we compute the time in days between the learning of the topic and its appearance on the progress assessment. Then, we group the data points into bins of width one day, compute the retention rate within each bin, and plot the results in Figure 3. The solid (blue) line shows the curve for the high mastery threshold topics, while the dashed (orange) line shows the curve for the low mastery threshold topics. For time values less than two weeks, the retention rate for the high mastery threshold group is higher—however, for larger time values it's not quite as clear how much of a difference, if any, exists between the retention curves.

We next use a linear regression model to more precisely estimate the differences in retention between the two mastery threshold groups—as our outcome variable is binary, this

model is known as a linear probability model. While the use of a generalized linear model—such as logistic regression—is typically recommended with binary outcome variables, we prefer to use a linear regression here to make it easier for us to interpret the coefficients. Although the use of a linear model with a binary outcome variable could theoretically lead to biased estimates, it's been argued that this bias is typically low [3]. Additionally, a criticism of the linear probability model is that it could give invalid probability estimates less than zero or greater than one. However, based on previous works analyzing forgetting in the ALEKS system [18, 36, 37, 38], we expect the probability estimates of a correct answer to be bounded well away from zero and one.

As some students appear multiple times within our data, we treat data points associated to the same student as a "group" or "cluster"—this leaves us with 436,735 clusters, one for each unique student. To handle these clusters appropriately, in each of our analyses we fit a marginal model using the generalized estimating equation (GEE) class in the `statsmodels` [54] Python library. GEE models are commonly applied in epidemiological studies and analyses containing repeated measurements [27, 28, 33, 57], making them well-suited for our study.

Our regression models include the following predictors.

- 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 event in learning sequence (correct, incorrect, or explanation)
- x_6: Categorical variable encoding time (in weeks) since topic was learned (see Table 3)
- x_7: Interaction between mastery and time ($x_1 \times x_6$)

The variables x_1 and x_7 are our main focus, as we want to estimate the average difference in retention between the two mastery groups—additionally, we want to see if these differences vary across the categories of the time variable. The remaining predictors are control variables, as they can help adjust for factors such as variation in starting knowledge (x_3), general differences between students using the various ALEKS products (x_4), and the amount of initial struggle experienced by the students while learning the topics (x_5). Finally, it's generally considered good practice to include the assignment variable, represented here by the probability estimate x_2, in the regression as well [23].

Regarding the time since the topic was learned, a complication with this variable is that it's technically a *post-treatment* variable—that is, it's measured after the "treatment" occurs, where in our case the treatment corresponds to the successful learning of the topic with the high mastery threshold. If a causal link is suspected between the post-treatment variable and the treatment, including the post-treatment variable in the regression could bias the estimate of the coefficient for the treatment variable [1, 53]. While we don't have a compelling reason to think there is a causal link between the mastery threshold and the time variable, we use the following procedure in an attempt to address the possibility of post-treatment bias. First, we run our regression analysis including the categorical variable for time. Next, we re-run our analysis using the two-step regression procedure known as the *sequential g-estimator* [29, 60]. Using 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, 24, 29, 60, 61]. Comparing the results using the sequential g-estimator to our first regression, we do not see any substantial differences—for example, in the first model we fit, the estimates of the coefficients of interest differ by less than 0.0013 in absolute value. Thus, to simplify the exposition, in what follows we describe and report the results from the models fit without using the sequential g-estimator.

Table 3: 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

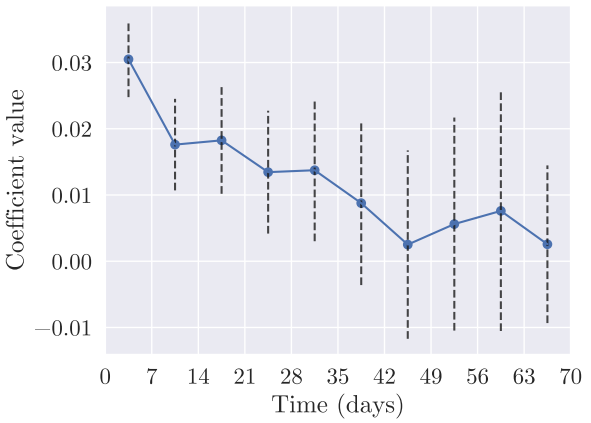

Figure 4: Coefficient estimates of the retention rate differences.

Figure 4 shows the results from fitting a model with variables x_1 through x_7. For the given time category, each (blue) dot represents the estimated average difference in retention between the two mastery thresholds, with the dashed lines showing the 95% confidence interval for each estimate. For example, the first dot represents the data points with a retention time of less than seven days, where the high mastery group has an estimated average retention rate that's higher by about 0.03, with a 95% confidence interval of (0.025, 0.036). The general trend suggests that larger time values are associated with smaller retention differences between the two groups. These results appear to be consistent with the plots shown in Figure 3, where the gap between the two forgetting curves is smaller for the larger time values.

Next, we take a different approach and use a type of matched

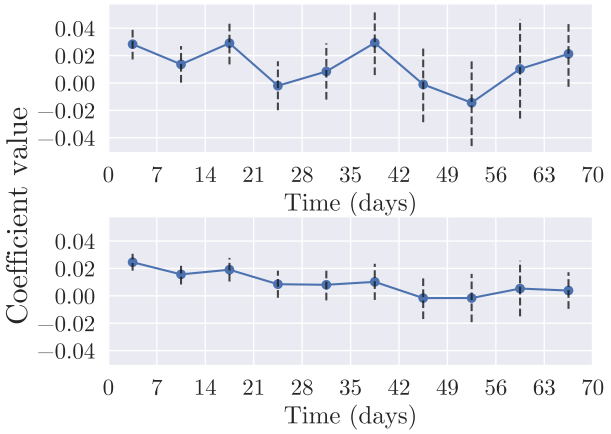

Figure 5: Matched data coefficient estimates, using bandwidths of 0.02 (top) and 0.04 (bottom).

design. Starting with the 580,317 data points from our previous analysis, we restrict the data to the 59,689 students who have learned topics using both mastery thresholds. This leaves us with 72,279 high mastery topics and 73,851 low mastery topics. The results from fitting a model using the variables x_1 through x_7 are shown in the top plot of Figure 5. Similar to the results from the unmatched data in Figure 4, the estimated average difference is about 0.03 for week 1, while then dropping below 0.02 for week 2. While the point estimates appear to have a negative trend—that is, the higher time values are associated with smaller estimated retention rate differences between the two mastery thresholds, on average—the trend is not nearly as pronounced as the one in Figure 4, and the estimates tend to be noisier.

In an attempt to get cleaner estimates of the average retention differences between the two mastery groups, we enlarge our data set by using a wider bandwidth of 0.04—this gives a total of 1,163,706 data points. Restricting our analysis to the 168,339 students who have examples of topics learned with both mastery thresholds, we have a new data set consisting of 476,105 data points. The results from the regression fit on this new data set are shown in the bottom plot of Figure 5, where it's instructive to see that there is a clearer pattern than in the top plot—that is, using the enlarged data set, there appears to be a fairly strong negative association between the time values and the estimated differences in retention, similar to the results shown in Figure 4.

6. DISCUSSION

In this work we presented a detailed comparison of two mastery learning thresholds that are used in the ALEKS system. Attempting to adjust for selection effects and other possible confounding variables, we utilized elements of a regression discontinuity design by leveraging the fact that the assignment of the mastery threshold is determined by a probability cutoff value. Focusing on topics with probabilities close to this cutoff, we looked at the learning outcomes for the two different mastery threshold groups, with the results suggesting that, while differences do exist, they are not particularly large. For example, the average learning ratio difference between the two groups was less than 0.02. Additionally, we used regression models to estimate the average difference in knowledge retention between the mastery threshold groups. The overall retention rates were more similar than we might have expected a priori and, furthermore, we saw evidence that the difference in retention rates between the two groups was negatively associated with time—that is, longer time gaps between the initial learning and the test of retention had smaller average differences in retention.

While performing our analyses, we investigated, and attempted to adjust for, several potential sources of bias and confounding. Nonetheless, being an observational study it's possible that other sources of bias exist. Thus, in what follows we interpret our results within the larger literature on learning and retention, and also discuss potential implications for the ALEKS system, all while keeping this caveat in mind. To start, given that the difference in retention between the mastery threshold groups was smaller for larger time values, this might suggest that any possible gains from the high mastery threshold are not persistent. Notably, prior research on learning has shown that massed practice (i.e., grouping learning into a single session) and overlearning (i.e., continuing to practice after a skill has been mastered), while possibly beneficial in the short-term, do not lead to learning that is especially durable or long lasting [30, 51]. However, as most experiments studying these learning strategies involve simple verbal tasks in a laboratory setting [10, 15, 51], we found it informative to see similar results for students learning mathematics in an adaptive learning system.

Furthermore, while only a limited number of experiments investigate these issues for learning mathematics, two particular studies seem relevant and informative for our current work. First, the results in [51] indicated that the gains from massed practice of mathematics problems did not appear to be as durable as those from using distributed practice—specifically, while the benefits from these techniques appeared similar after a week, with a longer gap of four weeks distributed practice was superior. Second, in [52] two massed practice conditions for learning mathematics problems—with these conditions being somewhat analogous to our high mastery and low mastery conditions—were compared, with no clear difference in performance observed between the conditions. Thus, the outcomes of these two studies are seemingly consistent with the results from our current work.

If the results of this study prove to be valid, a possible adjustment to the ALEKS system is to reduce the usage of the higher mastery threshold for topics close to the probability cutoff. The benefit of this approach is that it would allow students to more efficiently learn additional topics. Taking a slightly different view, it's well-documented that distributed practice is more effective as an overall learning strategy in comparison to massed practice [10, 15, 20, 30, 64]. Thus, rather than removing the high mastery threshold completely, perhaps the extra practice enforced by the high mastery threshold could be distributed over multiple learning sessions. Additionally, as previous work found evidence that retrieval practice within the ALEKS system is associated with better retention [38], it would be of interest to find the most effective way of combining the principles of both retrieval and distributed practice in the system. This is a line of research we are currently exploring in more detail.

7. 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):512–529, 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] O. Bälter, D. Zimmaro, and C. Thille. Estimating the minimum number of opportunities needed for all students to achieve predicted mastery. *Smart Learning Environments*, 5(1):1–19, 2018.

[8] 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.

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

[10] S. K. Carpenter, N. J. Cepeda, D. Rohrer, S. H. Kang, and H. Pashler. Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review*, 24(3):369–378, 2012.

[11] M. D. Cattaneo, M. Jansson, and X. Ma. Simple local polynomial density estimators. *Journal of the American Statistical Association*, 115(531):1449–1455, 2020.

[12] M. D. Cattaneo, M. Jansson, and X. Ma. *rddensity: Manipulation Testing Based on Density Discontinuity*, 2021. R package version 2.2.

[13] 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.

[14] H. Cen, K. R. Koedinger, and B. Junker. Is over practice necessary?-Improving learning efficiency with the cognitive tutor through educational data mining. In *International Conference on Artificial Intelligence in Education*. IOS Press, 2007.

[15] N. J. Cepeda, H. Pashler, E. Vul, J. T. Wixted, and D. Rohrer. Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological bulletin*, 132(3):354–380, 2006.

[16] 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.

[17] 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.

[18] 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.

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

[20] J. Dunlosky, K. A. Rawson, E. J. Marsh, M. J. Nathan, and D. T. Willingham. Improving students learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public interest*, 14(1):4–58, 2013.

[21] 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).

[22] S. Fancsali, T. Nixon, and S. Ritter. Optimal and worst-case performance of mastery learning assessment with Bayesian knowledge tracing. In *Proceedings of the Sixth International Conference on Educational Data Mining*, pages 35–42, 2013.

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

[24] 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.

[25] 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.

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

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

[28] 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.

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

[30] 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.

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

[32] K. Kelly, Y. Wang, T. Thompson, and N. Heffernan. Defining mastery: Knowledge tracing versus n-consecutive correct responses. pages 630–631, 2015.

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

[34] 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.

[35] 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.

[36] 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.

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

[38] 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.

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

[40] 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.

[41] J. McCrary. Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of econometrics*, 142(2):698–714, 2008.

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

[43] 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.

[44] P. I. Pavlik and J. R. Anderson. Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, 14(2):101–117, 2008.

[45] P. I. Pavlik, H. Cen, and K. R. Koedinger. Performance Factors Analysis–a new alternative to knowledge tracing. In *International Conference on Artificial Intelligence in Education*, pages 531–538, 2009.

[46] R. Pelánek and J. Řihák. Experimental analysis of mastery learning criteria. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 156–163, 2017.

[47] 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 Fourth International Conference on Educational Data Mining*, pages 139–148, 2011.

[48] 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.

[49] 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.

[50] 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.

[51] D. Rohrer and K. Taylor. 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, 2006.

[52] D. Rohrer and K. Taylor. The shuffling of mathematics problems improves learning. *Instructional Science*, 35(6):481–498, 2007.

[53] 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.

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

[55] 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.

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

[57] 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.

[58] 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.

[59] D. L. Thistlethwaite and D. T. Campbell. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology*, 51(6):309–317, 1960.

[60] S. Vansteelandt. Estimating direct effects in cohort and case-control studies. *Epidemiology*, 20:851–860, 2009.

[61] 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.

[62] Y. Wang and J. E. Beck. Incorporating factors influencing knowledge retention into a student model. In *Proceedings of the Fifth International Conference on Educational Data Mining*, pages 201–203, 2012.

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

[64] Y. Weinstein, C. R. Madan, and M. A. Sumeracki. Teaching the science of learning. *Cognitive Research: Principles and Implications*, 3(1):1–17, 2018.

[65] 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.

[66] 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.

[67] 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.

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