Studying Retrieval Practice in an Intelligent Tutoring System

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

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

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

ABSTRACT

Retrieval practice (also known as testing effect or test-enhanced learning) is a well-studied and established technique for improving the retention of knowledge. Many previous works have confirmed the benefits of retrieval practice in laboratory experiments involving the memorization of words or facts. In this study, we build on these works and analyze retrieval practice in an intelligent tutoring system. Using a large data set composed of the actions of almost 4 million students studying math and chemistry, we look at the possible benefits of retrieval practice in the ALEKS adaptive learning and assessment system. We compare two different types of retrieval practice—one involving the assessment of learned material, and another involving the learning of closely related content that builds on the learned material—leveraging the scale of the available data to control for several confounding variables. Finally, we look at the timing of retrieval practice within the system and the possible effect it has on forgetting. The results indicate that a delay in retrieval practice is associated with better retention and that, while being assessed on learned material is beneficial, the learning of closely related content is associated with an even higher rate of retention.

Author Keywords

Retrieval practice; intelligent tutoring system; forgetting curves; knowledge space theory; marginal model; generalized estimating equations.

INTRODUCTION

Memory and forgetting is an active area of research that is associated with a significant amount of work, both within the education domain and, more generally, as part of the fields of psychology and cognitive science. Of particular interest for our current work is the famous Ebbinghaus forgetting curve [4, 13], a model that represents the decay of knowledge over time. Many studies have looked at the conditions affecting these curves in settings as varied as laboratory experiments [17, 27, 31, 46], classrooms [2, 7, 16], and adaptive learning and intelligent tutoring systems [49, 51, 52].

L@S '20, August 12–14, 2020, Virtual Event, USA.

ACM ISBN 978-1-4503-7951-9/20/08 ...$15.00.
http://dx.doi.org/10.1145/3386527.3405927

Previous work has shown that learning systems can benefit greatly by accounting for the decay of knowledge. For example, models of student learning have been improved by explicitly including aspects of forgetting [11, 37, 50]. Additionally, other studies have shown that personalized interventions and review schedules can improve students' long-term retention of knowledge [24, 35, 45, 48, 53].

Going further, an additional effect associated with forgetting is that of *retrieval practice* (also known as *testing effect* or *test-enhanced learning* [40]). The idea behind retrieval practice is that being forced to actively recall information can help with the long-term retention of that information [38, 40, 41, 43], with the benefits of this procedure having been confirmed in numerous studies [5, 21, 40, 41, 42]. Additionally, the importance of the timing of retrieval practice has also been investigated, with some studies indicating that having a delay between the initial learning and the retrieval practice can substantially improve long-term retention [20, 22, 36, 39].

As discussed in [1], the majority of studies on retrieval practice take place in laboratory settings. Only a relatively small number of studies have analyzed the effects of retrieval practice outside of these controlled environments (of note, however, is that two recent reviews found that retrieval practice does appear to be beneficial in classrooms [1, 30]). Furthermore, the previously mentioned studies on the benefits of delaying retrieval practice are, again, mainly from laboratory experiments, and they typically involve the memorization of words or facts. Thus, in this work our goal is to build on previous research and analyze the effects of retrieval practice within the environment of an intelligent tutoring system. In doing so, we look to evaluate the benefits of retrieval practice outside of a controlled setting and in relation to learning complex material that goes beyond the memorization of words or facts.

Specifically, we look at retrieval practice within the ALEKS system. ALEKS, which stands for "**A**ssessment and **LE**arning in **K**nowledge **S**paces", is a web-based, artificially intelligent, adaptive learning and assessment system [29]. In the absence of data from a completely controlled experiment, we instead leverage the scale inherent to adaptive systems to run our study. Using a large data set composed of the actions of close to 4 million students, we begin by looking at the possible benefits of retrieval practice on knowledge retention and forgetting in ALEKS. Next, we compare two different types of retrieval practice that occur within ALEKS, using the large amount of available data to control for several confounding variables. Lastly, we look in detail at the timing of retrieval practice within the system and the relationship it has with forgetting.

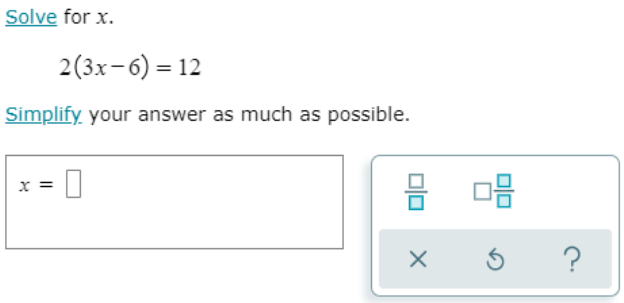

Figure 1. Screen capture of an ALEKS topic titled "Introduction to solving an equation with parentheses."

BACKGROUND

In this section we give a brief introduction to the ALEKS system and knowledge space theory (KST) [12, 14, 15], a mathematical framework that forms the foundation of ALEKS. KST uses combinatorial structures to model the knowledge of students in various academic fields. A *topic* in KST is a problem type that covers a discrete unit of an academic course.[1] Each topic is composed of many examples called *instances*, and these examples are carefully chosen to be equal in difficulty and to cover the same content. A *knowledge state* in KST is a collection of topics that, conceivably, a student at any one time could know how to do. Figure 1 contains a screen capture of the question and answer interface for an example math topic titled "Introduction to solving an equation with parentheses." As the title suggests, this topic introduces the technique of applying the distributive property to solve a linear equation. Note also that, rather than employing a multiple choice format, the question is open-ended, as the student is expected to enter the exact numerical solution (the majority of questions in ALEKS are similarly open-ended).

An important concept for our study of retrieval practice is that of prerequisite-postrequisite relationships between topics. Topic a is said to be a *prerequisite* for topic b if a must be learned before b can be learned; put another way, a contains necessary concepts and/or skills that must be learned before it's possible to completely master the material in b. In this relationship, b is then said to be a *postrequisite* of a. With regards to the ALEKS system, the prerequisite-postrequisite relationships between topics are carefully defined through a combination of human expertise and data, with two topics being labeled as a prerequisite-postrequisite pair only if there is strong evidence for this relationship. Regarding the topic in Figure 1, a typical postrequisite for this topic would require that a student solve a slightly more advanced equation; for example, one such postrequisite has the title "Solving a linear equation with several occurrences of the variable."

In ALEKS, the student is guided through a course via a cycle of learning and assessments. In an assessment, a student is presented a topic for which they can attempt to answer, or they can respond "I don't know" if they, presumably, have little knowledge of how to solve the problem. If the student attempts to answer the problem, the response is classified as either correct or incorrect. A course begins with an *initial assessment*, the goal of which is to accurately measure the starting knowledge of the student. The initial assessment classifies each of the topics in a course into one of the three following (mutually exclusive) categories.

- Topics that are most likely in the student's knowledge state (in-state)
- Topics that are most likely not in the student's knowledge state (out-of-state)
- The remaining topics (uncertain)

In the learning mode, at any one time the student can work on a subset of topics from the union of the out-of-state and uncertain categories, and this subset consists only of topics that the ALEKS systems believes the student is ready to learn. In the event that a student is unsure of the procedure for solving a problem, an explanation is available that contains a worked example of the current instance. A student is said to have "learned" or "mastered" a topic in the learning mode after a certain amount of success is demonstrated, where this success is computed based on the actions and responses of the student.

Each subsequent *progress assessment* is given to a student after some time has been spent in the learning mode. Additionally, in each progress assessment an *extra problem* is chosen uniformly at random from all of the available topics in the course; this extra problem is then presented to the student as an assessment question. The response on the extra problem does not affect the results of the assessment, but the data gathered from these responses are instead used for validation and other statistics evaluating the ALEKS system. This extra problem is also important for the analyses in this current work, as it allows us to make accurate estimates of student knowledge retention. For the purposes of this work, we define *retention* as the act of answering a topic correctly when it appears as an extra problem at a point in time after the topic is studied in the learning mode. We then say that the *retention rate* is the correct answer rate on these extra problems. In what follows, any question appearing in a progress assessment that is not an extra problem is referred to as a *regular assessment problem*.

EXPERIMENTAL SETUP

Our starting data set is composed of the complete learning and assessment profiles of 3,945,684 students, with the students being drawn from courses across the entire spectrum of ALEKS products. The majority of these are math products, starting with third grade math and ending with college-level precalculus; however, there is also a sizable portion of college-level chemistry students in the data. The student actions took place over a time period starting at the beginning of 2016 and ending halfway through 2019.

As outlined in [32], many studies of retrieval practice use the following three-step experimental setup.

(1) Initial study of material

(2) Retrieval practice (or, a control condition as a comparison)

[1] It is standard practice in the KST literature to refer to a topic as an *item*. However, to avoid confusion with the usage of "item" in other disciplines, for this work we use "topic" exclusively.

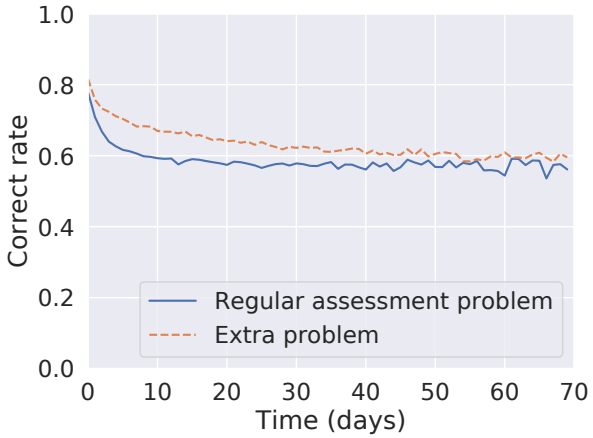

Figure 2. Forgetting curves comparing the regular assessment and extra problems from our assessment retrieval examples, based on the time since the topic was learned. The solid (blue) curve represents the correct answer rate to the regular assessment problems, while the dashed (orange) curve represents the correct answer rate to the extra problems.

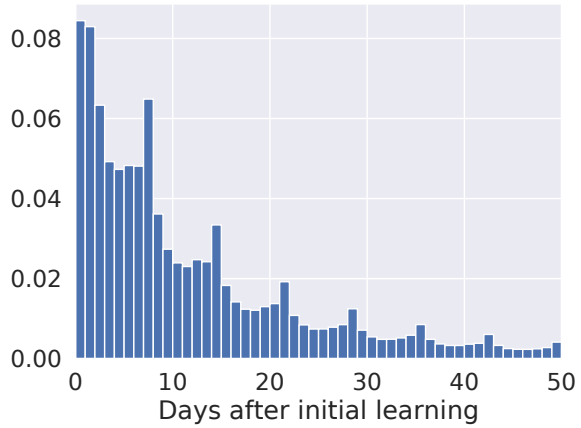

Figure 3. Relative frequency histogram showing the number of days between the initial learning of the topic and the assessment retrieval (i.e., when the topic appears as a regular assessment problem). Roughly 6% of the 851,266 data points have a value larger than 50 and don't appear in the histogram.

(3) Final test of retention

While we must re-emphasize that we won't be performing a true randomized experiment, we can instead leverage the large number of students in our data set to control for several different variables in an attempt to isolate, as much as possible, the effects of retrieval practice in ALEKS. In doing so, we compare two different types of retrieval practice. The first type we consider occurs when a topic is asked as a regular assessment problem on a progress assessment after the topic is mastered in the learning mode; we refer to this as *assessment retrieval*. In the specific context of this work, for all data points used to evaluate assessment retrieval, we require that the student has not mastered any postrequisites of the topic before it is asked on the progress assessment; by restricting our analysis in this way, we hope to reduce any possible bias that appears when students reinforce their knowledge by learning related content.

To evaluate the effect of assessment retrieval, we look at the response to the topic when it appears as an extra problem at a point in time after the retrieval occurred. As before, we want to control for any possible bias with regard to the placement of the topic in the student's knowledge state; thus, we again require that the student has not mastered any postrequisites before the topic appears as an extra problem. We can summarize our assessment retrieval data points as follows.

(1) A topic is studied and mastered in the learning mode

(2) Subsequent to (1), the topic appears as a regular assessment problem with no postrequisite topics being mastered

(3) Subsequent to (2), the topic appears as an extra problem with no postrequisite topics being mastered

One caveat with this approach is the following. If a student answers the regular assessment problem in step (2) incorrectly, in most cases they immediately return to the learning mode and review the problem; that is, they work on the topic in the learning mode until they again demonstrate a sufficient level of mastery. Thus, in these cases the retrieval practice includes more than the one regular assessment problem. While this complication means that we must be careful when drawing inferences about the effect of a single instance of retrieval practice, we can still draw some conclusions about the effectiveness of the overall process of assessment retrieval within ALEKS, as this extra review is an inherent and important part of the system.

The other type of retrieval we consider occurs when exactly one postrequisite of a topic is mastered after the original topic is learned; we refer to this as *learning retrieval*. To evaluate the effect of learning retrieval, we look at the response to the topic when it appears as an extra problem at a point in time after the postrequisite learning takes place, but before any additional postrequisite learning happens. Thus, by looking only at examples in which one postrequisite topic is learned, we want to isolate the effect of the learning of related material as much as possible. Importantly, to ensure that the learning retrieval and assessment retrieval effects are kept separate, from our learning retrieval data we exclude any examples where the topic appeared as a regular assessment problem before appearing as an extra problem. To summarize, our learning retrieval examples have the following characteristics.

(1) A topic is studied and mastered in the learning mode

(2) Subsequent to (1), a single postrequisite of the topic is mastered in the learning mode

(3) Subsequent to (2), the topic appears as an extra problem (a) with no additional postrequisite topics being mastered and (b) without ever having appeared as a regular assessment problem

Note that assessment retrieval is very similar to the standard form of retrieval practice that is frequently studied in the literature, as it involves a simple test of the learned knowledge. On the other hand, learning retrieval is fundamentally different in that it consists of being exposed to new material that

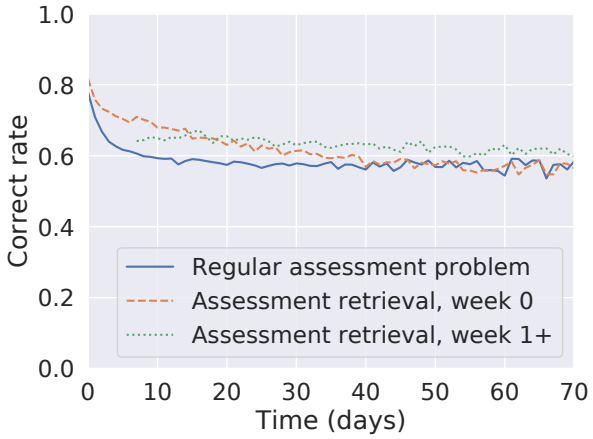

Figure 4. Assessment retrieval forgetting curves. As in Figure 2, the solid (blue) curve represents the correct answer rate to the regular assessment problems. The dashed (orange) curve represents the correct answer rate to the extra problems when the retrieval occurs within one week of learning, while the dotted (green) line represents the correct answer rate to the extra problems when the retrieval occurs more than one week after learning.

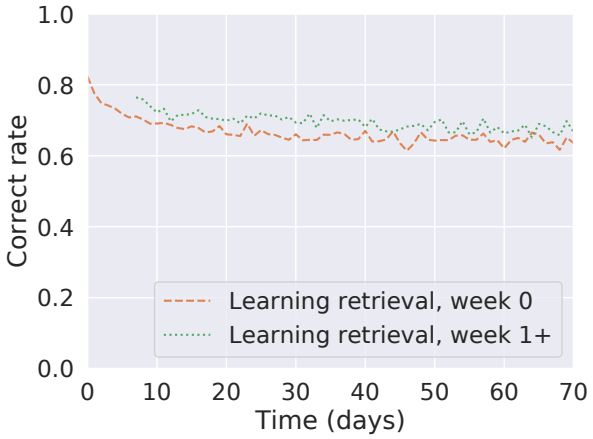

Figure 5. Learning retrieval forgetting curves. The dashed (orange) curve represents the correct answer rate to the extra problems when the retrieval occurs within one week of learning, while the dotted (green) line represents the correct answer rate to the extra problems when the retrieval occurs more than one week after learning.

builds on the learned knowledge. Due to these differences, in subsequent sections we spend a considerable amount of time analyzing and comparing the two types of retrieval.

EXPLORATORY ANALYSIS OF RETRIEVAL PRACTICE

For our initial analysis, we want to understand the effects of retrieval practice and whether it appears to benefit learning in ALEKS. We begin by looking at 851,266 examples that satisfy the requirements for assessment retrieval outlined in the previous section. For these data points, we compute two different forgetting curves that are indexed by the time since the topic was learned. The first curve is based on the correct answer rate to the regular assessment problems (i.e., the correct answer rate to the questions in (2) of the assessment retrieval definition given in the previous section). This curve can be viewed as the baseline forgetting curve, as it shows the correct rate of the learned topics without any retrieval practice. Next, after the retrieval practice has taken place, we can compare this curve to the forgetting curve for the extra problems (i.e., the correct answer rate to the questions in (3) of the assessment retrieval definition).

The results are shown in Figure 2. Similar to what has been shown in previous studies of ALEKS [25, 26], the correct rate for the regular assessment problems (shown by the solid line) decreases sharply within the first week or so, with the decline leveling off thereafter. On the other hand, we can see that, initially, the forgetting curve for the extra problems (dashed line) is above the curve for the regular assessment problems. However, the gap narrows as the time variable increases, and around 50 days or so the difference is minimal.

One possible confounding variable is the time at which the assessment retrieval takes place. Several previous works [20, 22, 36, 39] have shown that having a delay between the initial learning and the retrieval practice can be beneficial to long-term retention (one caveat is that these studies use smaller time scales than what we are considering in our analysis). To investigate this further, Figure 3 displays a relative frequency histogram of the number of days between the initial learning of the topic and the assessment retrieval. Of note is that roughly 42% of the acts of assessment retrieval occur within a week of the initial learning. Next, in Figure 4 we separate the effects associated with having the retrieval practice within one week, or at a later time. The solid (blue) line again represents the regular assessment problem curve from Figure 2. Then, the dashed (orange) curve shows the retention when the retrieval takes place less than seven days after learning, while the dotted (green) curve shows the retention when the retrieval takes place more than seven days after learning. From these curves, we can see that the delay in retrieval appears to be associated with better student knowledge retention, as the dotted curve is mostly flat and stays above the other curves more than two months after the initial learning. Interestingly, the curve representing early retrieval (i.e., retrieval within a week) has a higher retention rate initially, but this rate then declines rapidly and eventually converges with the baseline retention rate.

We next perform a similar analysis for learning retrieval. Our data set contains 298,659 examples that satisfy the requirements for learning retrieval outlined in the experimental setup section. In contrast to the assessment retrieval examples, we don't have a baseline forgetting curve as these examples specifically exclude the cases where the topic appears as a regular assessment problem. However, we can still look at the timing of the retrieval to see if this affects the overall retention. The results are shown in Figure 5 where, similar to the results in Figure 4, the later retrieval appears to be associated with better student knowledge retention. We explore the timing of retrieval practice in more detail in subsequent sections.

COMPARING THE BENEFITS OF LEARNING RETRIEVAL AND ASSESSMENT RETRIEVAL

In order to contrast the benefits of learning retrieval and assessment retrieval, we start by directly comparing the extra

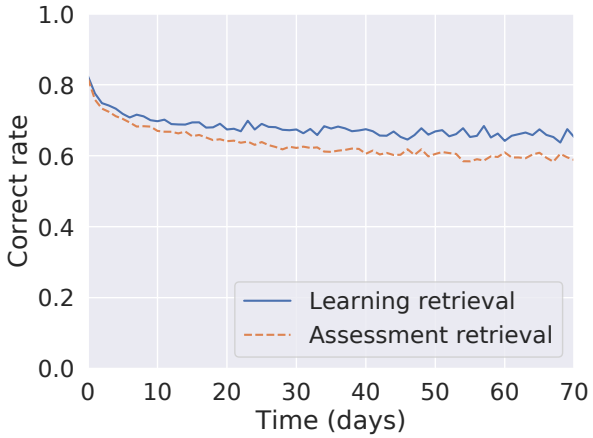

Figure 6. Forgetting curves comparing learning and assessment retrieval using all such examples in our data set (298,659 and 851,266 examples, respectively).

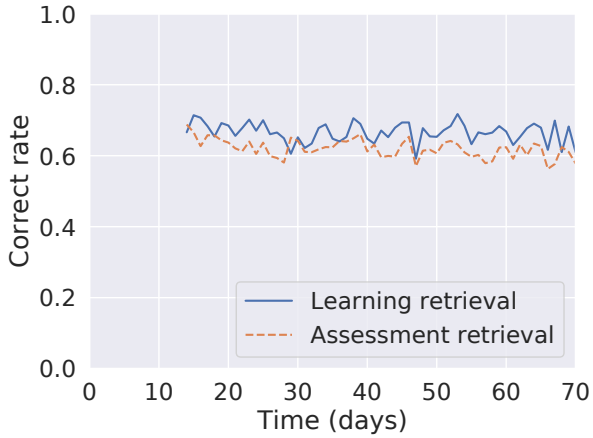

Figure 7. Forgetting curves comparing learning and assessment retrieval after equalizing at the student level. To focus on long-term retention we only look at data points for which the extra problem appeared at least 14 days after the initial learning of the topic.

problem forgetting curves in Figure 6. The learning retrieval curve contains 298,659 data points and the assessment retrieval curve has a sample size of 851,266; note that the learning retrieval curve combines the two extra problem curves from Figure 5, while the assessment retrieval curve is the same as the extra problem curve from Figure 2.

The curves start with roughly the same values, but the assessment retrieval curve then shows a steeper decline and stabilizes at around the 0.6 mark; for comparison, the learning retrieval curve ends with a value of about 0.64. While the difference in the figures is clear, another possible confounding variable that is not controlled for is the individual student. That is, it is possible that certain types of students may be more prone to one or the other type of retrieval, which may bias our comparison. For example, if a student is very strong and learns many problems, she may be more likely to appear under the learning retrieval category; on the other hand, a student who learns a topic and then does very little else is more likely to appear under the assessment retrieval category. So, to control for this possible bias, we use the following matching procedure to randomize the "treatment" (i.e., the type of retrieval) at the student level.

- Find all examples of assessment and learning retrieval in which the extra problem appears at least two weeks after the initial learning
- From this reduced data set, find all students who have examples of both types of retrievals
- Choose half of the students at random; for these students, choose at random one assessment retrieval example each
- For the other half of the students, choose at random one learning retrieval example each

After implementing this procedure, we now have a data set in which each of 67,080 students appear exactly once; thus, we've removed the dependence in our data at the student level. Additionally, by dividing the students randomly between the two retrieval categories we are attempting to remove any sampling bias from the type of retrieval. Finally, by removing examples where the time from the initial learning to the time the topic appeared as an extra problem is less than two weeks, we can focus on how the retrieval practice affects long-term retention; note that, as shown in Figure 6, after two weeks the curves have mostly flattened out.

The forgetting curves based on this new data set are shown in Figure 7. The curves are close, with the learning retrieval curve appearing to be slightly higher. To get a more precise measure of the differences between these two categories, our next step is to fit a logistic regression with the student's answer to the extra problem (correct or incorrect) as the dependent (response) variable, while the following are our independent (predictor) variables.

- x1: Time in days between the initial learning and appearance as an extra problem
- x2: Uncertain status of the topic (one if the topic is classified as uncertain; zero if it is classified as out-of-state)
- x3: Number of correct answers when learning topic
- x4: Number of incorrect answers when learning topic
- x5: Number of explanations viewed when learning topic
- x6: Time in days between the initial learning and the retrieval
- x7: Type of retrieval (one if learning retrieval; zero if assessment retrieval)

The variable x7 is our main focus, as its coefficient gives an indication of the relative difference between the retention rates of each type of retrieval. The other variables are introduced to control for possible confounding effects. The overall forgetting aspect is represented by x1; as mentioned previously, by looking only at examples for which x1 is greater than two weeks, we're attempting to minimize this effect to some degree.

Variable	*Mean*	*Median*	*Standard Deviation*
x1	69.71	55	50.56
x2	0.38	0	0.48
x3	3.17	3	1.28
x4	1.31	1	2.03
x5	0.86	0	2.13
x6	15.96	7	23.87
x7	0.5	0.5	0.5

Table 1. Descriptive statistics for independent (predictor) variables.

Then, x2 differentiates between topics that are classified as uncertain by the initial assessment and those classified as out-of-state; as some uncertain topics may be known by students at the time of the initial assessment (the assessment simply didn't have enough information to make this determination), we typically expect the average correct rate for uncertain topics to be higher than for out-of-state topics. Next, variables x3-x5 are used to control for the amount of difficulty the student experiences when learning the topic, while x6 then represents the timing of the retrieval. Table 1 contains descriptive statistics for each of these variables.

However, even with these additional predictors, another confounding variable that we have not completely controlled for is the topic that is being studied, as the forgetting curves are aggregated over all the topics in our data set. Given that students are guided through a course based on their current knowledge, the topics that appear are not independently sampled; that is, the data points pertaining to the same topic (most likely) share some underlying dependence or correlation. As evidence of this, the topic variable was by far the most important feature for the model of retention that was built in [26], and thus it seems likely that the effects of retrieval practice may differ somewhat across the topics. To control for this, we use a multilevel design where the data points associated with each topic are considered a "group" or "cluster." We then build a marginal (or, population-average) model using a generalized estimating equation (GEE) [18, 19, 23]. All our models are fit using the GEE class in the `statsmodels` [44] Python library.

GEE models were developed specifically to handle correlated data, and they are commonly used in epidemiological studies and studies containing repeated measurements. When using a GEE model, we must specify the type of correlation structure for the data within each group. An advantage of GEE models is that, even if this structure is misspecified, the parameter estimates are statistically consistent, and only the efficiency of these estimates is compromised [18, 23]. Two common choices are an exchangeable correlation structure and an independence correlation structure. The exchangeable structure assumes that there is some common dependence between all the data in a group, while the independence structure assumes that there is no dependence within each group [18, 19, 47].

We use the Quasi-AIC (QIC) score [33] to help choose between these two different correlation structures. Since the estimating equations used in GEE models are not necessarily likelihood based, the QIC score is an alternative to the Akaike

Variable	*Coefficient*	*SE*	*z*	*95% CI*
const	0.423	0.044	9.511	[0.336, 0.510]
x1	-0.003	0.000	-12.576	[-0.003, -0.002]
x2	0.525	0.025	20.945	[0.476, 0.574]
x3	0.065	0.012	5.278	[0.041, 0.089]
x4	-0.097	0.007	-14.281	[-0.111, -0.084]
x5	-0.092	0.006	-14.259	[-0.105, -0.079]
x6	0.002	0.000	5.460	[0.001, 0.003]
x7	0.157	0.020	7.666	[0.117, 0.197]

Table 2. Results from fitting a GEE logistic regression model with an exchangeable correlation structure to the data from Figure 7. The indicator variable representing the type of retrieval (x7) is statistically significantly different from zero (the p-value is approximately 2×10^{-14}).

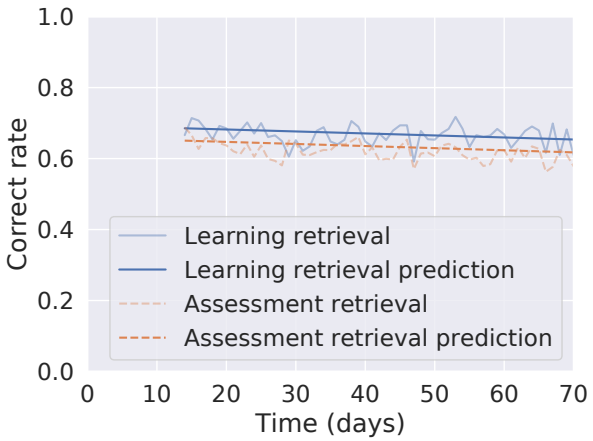

Figure 8. Forgetting curves and predictions comparing learning and assessment retrieval. At all times the learning retrieval prediction is greater than the assessment retrieval prediction by at least 0.035.

Information Criterion (AIC) [3] that can be used to compare the fits of different GEE models [18, 33]. Based on a comparison of the QIC scores for the two correlation structures, along with our prior belief that the data points pertaining to the same topic are correlated, we use the exchangeable correlation structure as our preferred model. The results from this model are reported in Table 2.

As seen in Table 2, the coefficient of the retrieval variable (x7) is significantly different from zero, with the 95% confidence interval ranging from 0.117 to 0.197. To get a sense of how much the variable affects the model predictions, we use the mean values from Table 1 for variables x1 to x6, and then compare the results for x7=0 and x7=1. For x7=0 the predicted probability is 0.618, while it is 0.654 when x7=1. To visualize these results, in Figure 8 we've reproduced the forgetting curves from Figure 7, but we've also added the predictions from the logistic regression based on the time since the original learning; in these predictions, we've again used the average values for x2 to x6, while now varying both x1 and x7.

While the above model gives some indication that learning retrieval is associated with better retention, our exploratory analysis in the previous section indicates that the timing of the retrieval has a large influence on the overall effectiveness of

Variable	*Description*
x6	Retrieval within 7 days of learning
x7	Retrieval between 7 and 14 days after learning
x8	Retrieval between 14 and 21 days after learning
⋮	
x14	Retrieval between 56 and 63 days after learning
x15	Retrieval more than 63 days after learning

Table 3. Indicator variables representing the number of weeks after learning that the retrieval practice occurs.

retrieval practice. Though we've included a variable for this in our regression, our next model attempts to more completely take this timing into account. To that end, we replace x6 and x7 (the time of the retrieval and the type of retrieval, respectively) with a more comprehensive set of variables. We first add 10 indicator (dummy) variables representing the number of weeks after learning that the retrieval takes place. These indicator variables are described in detail in Table 3.

Next, to test for a difference between the two types of retrieval, we add additional indicator variables that model the interaction effects between the timing of the retrieval and the type of retrieval. Specifically, these variables start with the same values that are described in Table 3, with the difference being that we multiply each of these values with a categorical variable that represents the type of retrieval (one if learning retrieval; zero if assessment retrieval). Thus, we simply repeat the values of x6 through x15 if we have an instance of learning retrieval and, otherwise, the values of x16 through x25 are all zero.

For example, suppose we have an instance of learning retrieval that occurs 19 days after the initial learning. Then, according to Table 3, the value of x8 would be one (as the retrieval occurs between 14 and 21 days after learning). Furthermore, since this is a learning retrieval example, the corresponding interaction term, x18, would also have a value of one. On the other hand, if this were an example of assessment retrieval, x8 would still have a value of one, but then all of the interaction terms (including x18) would have a value of zero. Since the values of x16 through x25 are (possibly) non-zero only for learning retrieval examples, by examining the coefficients of these predictors we can compare the two types of retrieval, while controlling for the timing of the retrieval more completely.

The coefficients of the interaction terms (x16 through x25) are shown in Figure 9. While only the first three values (representing weeks 0, 1 and 2) are statistically significantly different from zero, it's clear that the overall trend is for the coefficients to be positive. Thus, since these variables represent the interaction between the occurrence of learning retrieval and the timing of the retrieval, this is evidence that learning retrieval is associated with better retention than assessment retrieval is, even when taking into account the timing of the retrieval.

Combining the various results from this section, there is strong evidence that, within the ALEKS system, learning retrieval is associated with better retention than assessment retrieval. This evidence appears even after controlling for several confounding variables, through a combination of randomization and and the use of a multilevel model. As mentioned previously, one complication is that in some of the assessment retrieval examples, a student actually reviews the topic before it appears as an extra problem. However, as this extra retrieval practice should only be beneficial to retention (or, at the very least, it should not adversely affect retention), and in light of the fact that assessment retrieval is associated with lower retention in ALEKS, it's not much of a stretch to conclude that this same association would be present if the reviewing component of assessment retrieval were completely removed.

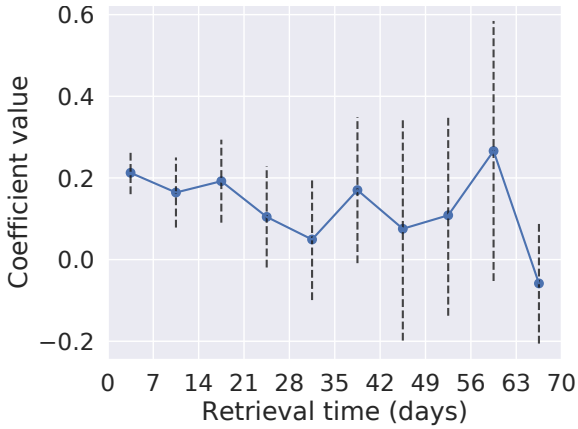

Figure 9. Coefficients of interaction terms (x16 through x25) for retrieval time and learning retrieval variables. Vertical lines represent the 95% confidence intervals. While only the first three terms are statistically significantly different from zero, overall the coefficients show a positive trend, signifying that learning retrieval seems to be associated with better retention in comparison to assessment retrieval.

A possible explanation for the superior results associated with learning retrieval is that, within ALEKS, learning a topic is an involved process requiring a student to solve multiple instances of the topic; furthermore, this process applies and consolidates the knowledge from the prerequisite topic. In comparison, being assessed on a topic only requires answering one instance of the topic, regardless of whether or not the submitted answer is correct. Additionally, several concepts and results from the educational psychology literature may be involved here, and we return to this analysis in the discussion section.

ANALYZING THE TIMING OF RETRIEVAL PRACTICE

In this section we take a deeper look at the effects of the timing of the retrieval practice. Starting from our data set consisting of 851,266 assessment retrieval examples, we extract two new data sets to analyze. For the first of these new data sets, we choose the examples such that we have a fixed interval for the time between the initial learning and the appearance of the extra problem. Specifically, we find all the students who have at least one data point for which (a) the retrieval occurred less than 56 days (8 weeks) after the initial learning and (b) the extra problem appeared at least 60 days after the initial learning, but no more than 90 days. The motivation for (b) is that, to separate as much as possible the effects of the timing of the retrieval from the overall forgetting of knowledge that occurs, we want the appearance of the extra problem to be in

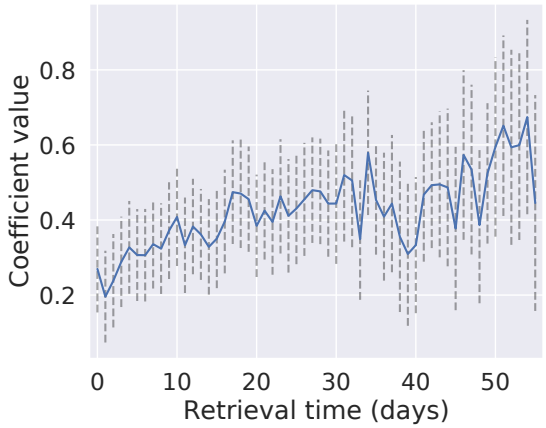

Figure 10. GEE logistic regression coefficients for indicator variables representing the number of days between the learning of the topic and the assessment retrieval, using the fixed interval data set. Vertical lines represent the 95% confidence intervals. All the extra problems appeared between 60 and 90 days after the initial learning took place.

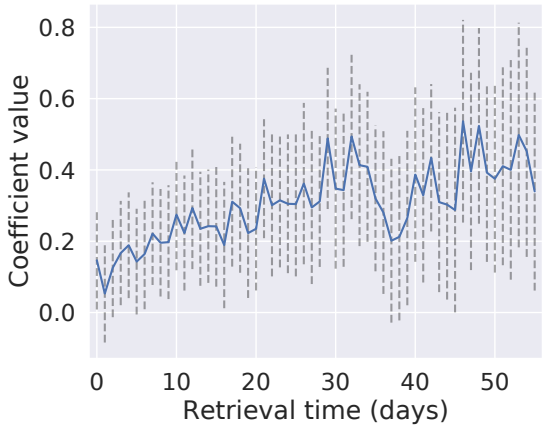

Figure 11. GEE logistic regression coefficients for indicator variables representing the number of days between the learning of the topic and the assessment retrieval, using the moving interval data set. Vertical lines represent the 95% confidence intervals. In contrast to the model in Figure 10, all the extra problems appeared between 60 and 90 days after the assessment retrieval took place. So, for example, if the retrieval happened 12 days after the learning of the topic, the extra problem appeared somewhere between 72 and 102 days after the initial learning.

a fixed time window after learning. Then, for each student we randomly choose one of these data points, ensuring that we remove the dependence at the student level.

Our second data set uses a moving interval for the appearance of the extra problem, and this interval varies based on the timing of the retrieval. To accomplish this, we find all students with at least one data point for which (a) the retrieval occurred less than 56 days (8 weeks) after the initial learning and (b) the extra problem appeared at least 60 days after the retrieval took place, but no more than 90 days after the retrieval. Thus, while the 60 to 90 day window in the previous data set was based on the initial learning, now it is based on the timing of the retrieval. So, for example, if the retrieval happened 12 days after the learning of the topic, the extra problem appears somewhere between 72 and 102 days after the initial learning.

As it's not obvious that one of the two procedures is superior, by analyzing both data sets we hope to obtain stronger evidence for any effects associated with the timing of the retrieval. To that end, we again fit multilevel GEE logistic regressions using the topics to form our groups. We use variables x1 through x5, introduced in the previous section, as well as 56 indicator variables, one for each of the possible days after learning in which the retrieval takes place. For the fixed interval data set, the values of the indicator variable coefficients are shown in Figure 10. We can see that the coefficients are smallest within the first few days of learning, and then steadily increase until about day 30. While the subsequent values are noisier (this is exemplified by the larger error bars), there again seems to be an upward trend starting at around day 40.

The results for the moving interval data set are shown in Figure 11, where the overall trend of the coefficient values is similar to that shown in Figure 10; the values are smallest within the first few days, and then steadily increase until roughly day 30, as before. However, one difference from Figure 10 is that there is seemingly less of an increase after day 40. (In this analysis, note that we're simply comparing the trends of the coefficient values in the two figures, and we're not trying to make any comparisons or statements about the relative sizes of the coefficients; the latter would be problematic as these are separate regression models fit on different data sets).

For the moving interval data set the appearance of the extra problem is always at least 60 days after the retrieval practice; thus, the overall time between the initial learning and the extra problem is longer, on average, than in the fixed interval data. While we've controlled for the learning of any postrequisite material in ALEKS when choosing our assessment retrieval data, it's likely a student is learning outside of ALEKS as well, and all things being equal we would expect a student who is further along in the course to know more. As this effect doesn't exist with the results from Figure 10, by taking into account the results in both figures we are attempting to compensate for these issues; it is therefore noteworthy that the results are similar with both of these groupings.

The lowest benefit of retrieval at day zero is seemingly consistent with prior research (albeit, on different time scales) such as [22], where it was shown that delaying the first act of retrieval practice led to better long-term retention. Additionally, in works such as [20] and [36], it was shown that having longer spacing between repeated retrieval attempts resulted in a large improvement in long-term retention. The reason given for these effects is that delaying the retrieval practice makes it more difficult, and it is thought that some difficulty in retrieval is beneficial for long-term retention [28, 41]. This idea is sometimes described as *desirable difficulties* [8, 9].

Next, in Figures 12 and 13 we perform the same analysis for our learning retrieval data. Here, the results are much different. While the values within the first few days appear to be lowest, the rest of the values are relatively flat, and the overall increasing trend is very mild. Thus, this seems to indicate that learning retrieval is much less sensitive to the actual timing of

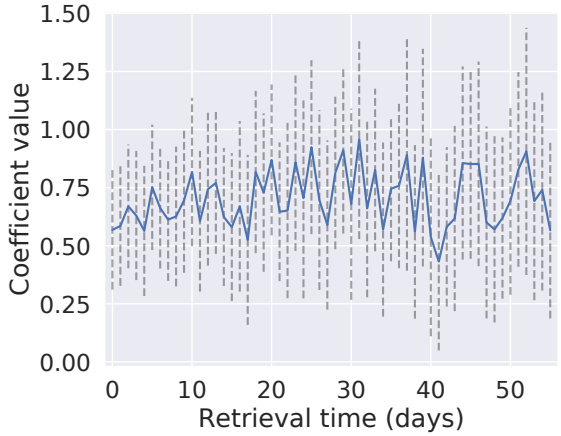

Figure 12. GEE logistic regression coefficients for indicator variables representing the number of days between the learning of the topic and the learning retrieval, using the fixed interval data. Vertical lines represent the 95% confidence intervals. All the extra problems appeared 60 to 90 days after the initial learning took place.

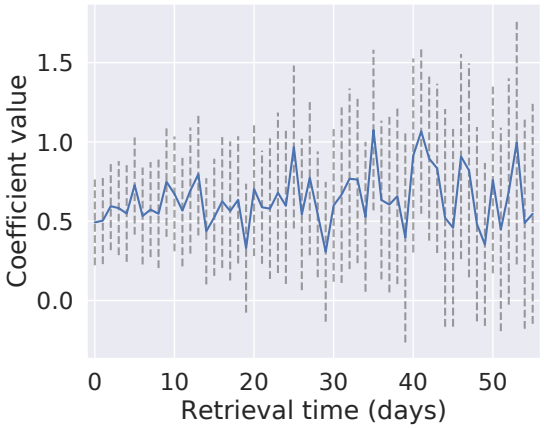

Figure 13. GEE logistic regression coefficients for indicator variables representing the number of days between the learning of the topic and the learning retrieval, using the moving interval data. Vertical lines represent the 95% confidence intervals. As opposed to the model in Figure 12, all the extra problems appeared between 60 and 90 days after the learning retrieval took place. So, for example, if the retrieval happened 12 days after the learning of the topic, the extra problem appeared somewhere between 72 and 102 days after the initial learning.

the retrieval. Relating this to the idea of desirable difficulties, one explanation for this lack of sensitivity is that learning a postrequisite is always somewhat difficult and is affected less by forgetting, regardless of whether it happens at day 0 or day 100. That is, as the postrequisite topic is more advanced than the original topic, it presents a more formidable challenge than simply answering the original topic in an assessment; as such, the difficulty of this challenge is always present, regardless of the timing of the retrieval. Additionally, the more involved nature of learning retrieval, in which a student works on multiple instances of a topic, may also play a role. Specifically, it seems possible that this extra practice could compensate for any forgetting that has occurred; thus, by the time the student masters the postrequisite topic, any lingering effects from forgetting have disappeared.

DISCUSSION

In this work we studied in detail the effects associated with retrieval practice in the ALEKS adaptive learning system. After controlling for several confounding variables, we found evidence that, within the system, the students in the learning retrieval category had better retention rates in comparison to the students in the assessment retrieval category. That is, the students who learned a postrequisite topic retained the knowledge of the original topic better than the students who were assessed on the original topic. Additionally, we also looked at how the timing of this retrieval affects retention. While the timing had a more dramatic influence on assessment retrieval, in both types of retrieval the retention was lowest when the retrieval practice happened within the first few days of learning.

Our experimental design attempted to control for several confounding variables; this included variables related to both the topics and the individual students. As an alternative, instead of the randomized matching procedure we applied to our data, another possibility would have been to use a within subjects design. However, using this design in combination with the topic clusters results in multiple levels that are not nested, complicating the building of a marginal model. Additionally, we believe that the randomized matching procedure is less biased and more closely approximates a fully randomized experiment. While the method we employed has the obvious drawback of removing data from our analysis, the large amount of students in the full data set compensated for this and gave us the flexibility we needed to complete our analyses. Thus, for all of these reasons, we decided to use this particular experimental design to focus on the overall benefits of retrieval practice within ALEKS.

Now, as this is still fundamentally a quasi-experimental study, it is possible that some confounding variables were not properly controlled for. Thus, further research is necessary before any strong inferences and claims can be made about the causal effects of learning and assessment retrieval in intelligent tutoring systems. With that caveat in mind, however, we can analyze the results from this study within the context of the existing body of work on retrieval practice. Regarding the advantage of learning retrieval, admittedly, this is not the result that the authors expected to see when this work began. However, with the wisdom gained from the benefit of hindsight, there are several factors that may explain and contribute to this effect. To start, from an implementation standpoint, mastering a topic in the learning mode is a more intensive activity in comparison to being assessed on a topic. Mastering a topic requires a student to work on multiple instances of the topic, and it has been shown that repeated acts of retrieval are more beneficial than a single act [21, 36, 40]. Another important aspect is obtaining feedback during retrieval practice, which also helps with the long-term retention of knowledge [6, 10, 34]. When the retrieval practice happens during the assessment, the student does not receive immediate feedback on the correctness of their answer; on the other hand, while learning the postrequisite topic, the student gets feedback each time an

answer is submitted and, additionally, has access to an explanation of the problem that contains a worked out example.

Furthermore, based on the idea that difficult retrieval practice is more beneficial than easy practice (i.e., desirable difficulties), learning retrieval would appear to have another advantage. As the postrequisite topic contains more advanced material than the original topic does, it's likely that learning the postrequisite is more difficult in comparison to being tested on the original topic; this extra difficulty would then, presumably, be more beneficial to retention. The desirable difficulties concept also coincides with the observation that both types of retrieval practice are associated with better retention when the retrieval is delayed by at least a few days. This last result aligns with previous work in this area [20, 22, 36] showing that a delay in retrieval introduces extra difficulties, which then benefits retention.

Additionally, our analysis provides evidence that the effectiveness of assessment retrieval continues to increase as the retrieval is delayed further. Based on the plots in Figures 10 and 11, the best results are obtained when the assessment retrieval takes place at least 21 or so days after learning. A possible explanation for this observation again comes from the desirable difficulties framework. That is, being tested on a topic immediately after learning it is much less challenging than being tested on the topic 21 days later; this is clearly seen from the regular assessment problem forgetting curve in Figure 2, where the correct rate drops by roughly 20 percentage points from day 0 to day 21. Thus, the increased difficulty associated with the later retrieval times could explain, at least in part, the greater benefit on retention after a few weeks. On the other hand, learning retrieval is much less sensitive to the timing of the retrieval (Figures 12 and 13), and there is evidence that the full benefits of learning retrieval are gained after a short delay of only a few days. A possible explanation is that the difficulty of learning a postrequisite is not as dependent on the time since the original topic was learned; that is, forgetting has less of an effect on the difficulty of the retrieval, as learning retrieval requires the mastery of completely new material (and not just the recalling of already learned material). The result is that the difficulty of the task stays relatively constant over time, as then does the effect on retention. Furthermore, the repeated practice that occurs with learning retrieval is likely a factor as well, as it may help compensate for any forgetting that has occurred.

Putting this all together, the suggested effects of retrieval practice that we observed are consistent with the previous body of research in this area. Importantly, however, while other studies observing such effects have concentrated more on pure memorization and are mostly performed in controlled laboratory settings, our setup is much different. The material that is learned and tested in our data consists of complex math and chemistry problems, rather than words or simple facts. Additionally, all of the learning takes place within an intelligent tutoring system in an uncontrolled manner. While the obvious drawback is that this introduces extra complexities when attempting to isolate these effects, the upside is that we get to observe retrieval practice "in the wild", so to speak; that is, we observe these effects in a messy, real-world learning environment. Thus, it is both interesting and informative to see many of the same general effects that have been observed in laboratory studies of forgetting and retrieval practice play out in the environment of an intelligent tutoring system.

If these findings hold up under further scrutiny, there are a couple of ways in which they can be used to benefit students working in adaptive learning and intelligent tutoring systems. Given the hypothesized effectiveness of learning retrieval, one could argue that allowing students to progress faster through these systems and learn more material is a viable strategy, as such learning can act as an ongoing form of retrieval practice. That is, relying less on assessing and confirming already learned material, and focusing more on the learning of new concepts, may help student learning. Additionally, the observation that the effectiveness of retrieval practice increases after a long delay is important, as in the case of assessment retrieval it's an argument for decreasing the frequency and duration of these types of assessments.

Confirming the benefits of learning retrieval would also lend some validity to the prerequisite-postrequisite pairs used in the ALEKS system. Given that the analysis of learning retrieval relies on strong prerequisite-postrequisite relationships between topics, another direction for future research would be to analyze these specific relationships in more detail. For example, do certain types of postrequisite topics have characteristics that work better for learning retrieval? On the other hand, if a postrequisite does not seem to be helpful for learning retrieval, further investigation into the validity of the prerequisite-postrequisite pair may be warranted.

As mentioned previously, learning complex material in subjects such as math and chemistry is, in some sense, very different from the type of retrieval practice typically encountered in the literature. While we have positioned this study within the context of such previous work, much of what we observed may in fact be due to the benefits of practicing procedural knowledge, as a reviewer suggested, rather than the more straightforward act of simply retrieving information. Based on this idea, it would be of interest to see if the benefits of retrieval practice are influenced by the specific focus and content of the topics. Exploring these relationships further would improve our understanding of the potential benefits associated with the various types of retrieval practice.

REFERENCES

[1] Olusola O. Adesope, Dominic A. Trevisan, and Narayankripa Sundararajan. 2017. Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research* 87, 3 (2017), 659–701.

[2] Pooja K. Agarwal, Patrice M. Bain, and Roger W. Chamberlain. 2012. The value of applied research: Retrieval practice improves classroom learning and recommendations from a teacher, a principal, and a scientist. *Educational Psychology Review* 24 (2012), 437–448.

[3] Hirotugu Akaike. 1974. A new look at the statistical model identification. *IEEE Trans. Automat. Control* 19, 6 (1974), 716–723.

[4] Lee Averell and Andrew Heathcote. 2011. The form of the forgetting curve and the fate of memories. *Journal of Mathematical Psychology* 55 (2011), 25–35.

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

[6] Robert L. Bangert-Drowns, Chen-Lin C. Kulik, James A. Kulik, and MaryTeresa Morgan. 1991. The instructional effect of feedback in test-like events. *Review of Educational Research* 61, 2 (1991), 213–238.

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

[8] Robert A. Bjork. 1994. Memory and metamemory considerations in the training of human beings. In *Metacognition: Knowing about knowing*, Janet Metcalfe, Arthur P. Shimamura, et al. (Eds.). MIT press.

[9] Robert A. Bjork. 1999. Assessing our own competence: Heuristics and illusions. In *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application*, Daniel Gopher and Asher Koriat (Eds.). MIT Press.

[10] Andrew C. Butler, Jeffrey D. Karpicke, and Henry L. Roediger III. 2008. Correcting a metacognitive error: feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34, 4 (2008), 918.

[11] Benoît Choffin, Fabrice Popineau, Yolaine Bourda, and Jill-Jênn Vie. 2019. DAS3H: Modeling Student Learning and Forgetting for Optimally Scheduling Distributed Practice of Skills. In *Proceedings of the 12th International Conference on Educational Data Mining*. 29–38.

[12] Jean-Paul Doignon and Jean-Claude Falmagne. 1985. Spaces for the assessment of knowledge. *International Journal of Man-Machine Studies* 23 (1985), 175–196.

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

[14] Jean-Claude Falmagne, Dietrich Albert, Christopher Doble, David Eppstein, and Xiangen Hu (Eds.). 2013. *Knowledge Spaces: Applications in Education*. Springer-Verlag, Heidelberg.

[15] Jean-Claude Falmagne and Jean-Paul Doignon. 2011. *Learning Spaces*. Springer-Verlag, Heidelberg.

[16] Nicole A.M.C. Goossens, Gino Camp, Peter P.J.L. Verkoeijen, Huib K. Tabbers, Samantha Bouwmeester, and Rolf A. Zwaan. 2016. Distributed Practice and Retrieval Practice in Primary School Vocabulary Learning: A Multi-classroom Study. *Applied Cognitive Psychology* 30, 5 (2016), 700–712.

[17] Patricia Hanley-Dunn and John L. McIntosh. 1984. Meaningfulness and recall of names by young and old adults. *Journal of Gerontology* 39 (1984), 583–585. Issue 5.

[18] James W. Hardin and Joseph M. Hilbe. 2012. *Generalized Estimating Equations*. Chapman and Hall/CRC.

[19] Patrick J. Heagerty and Scott L. Zeger. 2000. Marginalized multilevel models and likelihood inference (with comments and a rejoinder by the authors). *Statist. Sci.* 15, 1 (2000), 1–26.

[20] Jeffrey D. Karpicke and Althea Bauernschmidt. 2011. Spaced retrieval: absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 37, 5 (2011), 1250.

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

[22] Jeffrey D. Karpicke and Henry L. Roediger III. 2007. Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33, 4 (2007), 704.

[23] Kung-Yee Liang and Scott L. Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73, 1 (1986), 13–22.

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

[25] Jeffrey Matayoshi, Umberto Granziol, Christopher Doble, Hasan Uzun, and Eric Cosyn. 2018. Forgetting Curves and Testing Effect in an Adaptive Learning and Assessment System. In *Proceedings of the 11th International Conference on Educational Data Mining*. 607–612.

[26] Jeffrey Matayoshi, Hasan Uzun, and Eric Cosyn. 2019. 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*. 258–269.

[27] Dawn M. McBride and Barbara Anne Dosher. 1997. A comparison of forgetting in an implicit and explicit memory task. *Journal of Experimental Psychology: General* 126 (1997), 371–392. Issue 4.

[28] Mark A. McDaniel and Gilles O. Einstein. 2005. Material Appropriate Difficulty: A Framework for Determining When Difficulty Is Desirable for Improving Learning. In *Decade of Behavior. Experimental Cognitive Psychology and its Applications*, A. F. Healy (Ed.). American Psychological Association, 73–85.

[29] McGraw-Hill Education/ALEKS Corporation. 2019. What is ALEKS? `https://www.aleks.com/about_aleks`. (2019).

[30] Bruna Fernanda Tolentino Moreira, Tatiana Salazar Silva Pinto, Daniela Siqueira Veloso Starling, and Antônio Jaeger. 2019. Retrieval practice in classroom settings: A review of applied research. *Frontiers in Education* 4 (2019), 5.

[31] Allan Paivio and Padric C. Smythe. 1971. Word imagery, frequency, and meaningfulness in short-term memory. *Psychonomic Science* 22 (1971), 333–335. Issue 6.

[32] Steven C. Pan and Timothy C. Rickard. 2018. Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin* 144, 7 (2018), 710.

[33] Wei Pan. 2001. Akaike's information criterion in generalized estimating equations. *Biometrics* 57, 1 (2001), 120–125.

[34] Harold Pashler, Nicholas J. Cepeda, John T. Wixted, and Doug Rohrer. 2005. When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31, 1 (2005), 3.

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

[36] Mary A. Pyc and Katherine A. Rawson. 2009. Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language* 60, 4 (2009), 437–447.

[37] Yumeng Qiu, Yingmei Qi, Hanyuan Lu, Zachary A. Pardos, and Neil T. Heffernan. 2011. Does Time Matter? Modeling the Effect of Time with Bayesian Knowledge Tracing.. In *Proceedings of the 4th International Conference on Educational Data Mining*. 139–148.

[38] Katherine A. Rawson and John Dunlosky. 2011. Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General* 140, 3 (2011), 283.

[39] Katherine A. Rawson, Kalif E. Vaughn, and Shana K. Carpenter. 2015. Does the benefit of testing depend on lag, and if so, why? Evaluating the elaborative retrieval hypothesis. *Memory & Cognition* 43, 4 (2015), 619–633.

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

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

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

[43] Henry L. Roediger III, Adam L. Putnam, and Megan A. Smith. 2011. Ten benefits of testing and their applications to educational practice. In *Psychology of Learning and Motivation*. Vol. 55. Elsevier, 1–36.

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

[45] Burr Settles and Brendan Meeder. 2016. A trainable spaced repetition model for language learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1848–1858.

[46] Steven M. Smith. 1979. Remembering in and out of context. *Journal of Experimental Psychology: Human Learning and Memory* 4 (1979), 460–471. Issue 5.

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

[48] Behzad Tabibian, Utkarsh Upadhyay, Abir De, Ali Zarezade, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. 2019. Enhancing human learning via spaced repetition optimization. *Proceedings of the National Academy of Sciences* 116, 10 (2019), 3988–3993.

[49] Yutao Wang and Joseph E. Beck. 2012. Incorporating Factors Influencing Knowledge Retention into a Student Model. In *Proceedings of the 5th International Conference on Educational Data Mining*.

[50] Yutao Wang and Neil T. Heffernan. 2011. Towards Modeling Forgetting and Relearning in ITS: Preliminary Analysis of ARRS Data. In *Proceedings of the 4th International Conference on Educational Data Mining*. 351–352.

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

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

[53] Xiaolu Xiong, Yan Wang, and Joseph Barbosa Beck. 2015. Improving Students' Long-term Retention Performance: A Study on Personalized Retention Schedules. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*. ACM, 325–329.