Forgetting curves and testing effect in an adaptive learning and assessment system

Jeffrey Matayoshi
McGraw-Hill Education/
ALEKS Corporation
Irvine, CA
jeffrey.matayoshi@aleks.com

Umberto Granziol
University of Padova
Padova, Italy
umberto.granziol@phd.unipd.it

Christopher Doble
McGraw-Hill Education/
ALEKS Corporation
Irvine, CA
christopher.doble@aleks.com

Hasan Uzun
McGraw-Hill Education/
ALEKS Corporation
Irvine, CA
hasan.uzun@aleks.com

Eric Cosyn
McGraw-Hill Education/
ALEKS Corporation
Irvine, CA
eric.cosyn@aleks.com

ABSTRACT

In the context of an adaptive learning and assessment system, ALEKS, we examine aspects of forgetting and aspects of a ‘testing effect’ (in which the act of simply being presented a problem in an assessment seems to assist in the learning process). Using a dataset consisting of over six million ALEKS assessments, we first look at the trend of student responses over the course of the assessment, finding little evidence for such a testing effect. We then refine our approach by looking at cases in which a question is repeated in an assessment; repeats are possible because some question is always chosen at random in an assessment for data-collection purposes. We find evidence of a testing effect for higher-performing students; for lower-performing students, we find a decreased willingness to attempt an answer the second time a problem is presented. Then, turning to forgetting, we find that the content representing the “high points” of a student’s learning sees a more precipitous drop in the student’s memory than does other content (perhaps because the “high point” skills and concepts may not have been practiced or developed much since the original learning event). Consequences and possible improvements for the ALEKS system, and also a brief comparison to recent work in the modeling of forgetting, are mentioned.

Keywords

Knowledge space theory, adaptive learning, forgetting curves, testing effect

1. INTRODUCTION

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 [13]. The artificial intelligence of ALEKS is a practical implementation of knowledge space theory (KST) [5, 7, 8], a mathematical theory that employs combinatorial structures to model the knowledge of learners in various academic fields of study including math [11, 15], chemistry [9, 18] and even dance education [19].

2. BACKGROUND

Memory and forgetting is an area that has seen significant research, pioneered by the late-nineteenth century work of Ebbinghaus with his ‘forgetting curves’ [2, 6]. Ebbinghaus posited that memory, as measured, say, by the ability to recall words presented in a list, decays exponentially with time; one such exponential model is given in Equations (7.1) and (7.2) in Section 7 below. A great deal of study has been done on the possible effects of various experimental conditions, such as whether the experiment probes explicit or implicit memory [12], the effect of the physical context in which the learning and recall take place [3, 17], and the extent to which the content is meaningful for the participant [10, 14], among many other experimental conditions. In the current paper, we will examine forgetting in the context of the adaptive learning and assessment system ALEKS, attempting to isolate the effect of aspects of the adaptivity on forgetting.

We will also look at a kind of ‘testing effect’ in which the act of simply being presented content in the adaptive assessment seems to assist in the learning process [1, 4]. We use the term ‘testing effect’ somewhat loosely here, as our use differs from that typically seen in the literature, since, for example, our situation does not include systematic feedback [16]. We use the term only to refer to a situation in which recall (or skill, or confidence) seems improved as content is encountered during an assessment.

In KST, an *item* is a problem that covers a discrete skill or concept. Each item is composed of many examples called *instances*; these instances are carefully chosen to be equal in difficulty and to cover the same content. A *knowledge state* in KST is a collection of items that, conceivably, a student at any one time could know how to do. In other words, roughly speaking, a set of items is a knowledge state if some student could know how to do all of the items in the set and not know how to do any of the items outside the set. For example, the empty set and full set are always considered knowledge states.

Another important concept from KST is the *inner fringe* of a knowledge state. An item is contained in the inner fringe of a knowledge state when the item can be removed from the state and the remaining set of items forms another knowledge state. Intuitively, the inner fringe items are the “high points” of a student’s knowledge, as they are not prerequisties required to master any of the other items in the knowledge state. This concept will be important for our work on forgetting in Sections 5 and 6.

While using the ALEKS software, the student is guided through a course via a cycle of learning and assessments. Each assessment (described below) updates the system’s assignment of a knowledge state to the student. Then, in the learning mode, the student is given problems to practice based on her knowledge state, with the system tracking the student’s performance and continually updating the student’s knowledge state. Subsequent assessments then modify the knowledge state as needed, and the process continues.

Each ALEKS assessment has about 15 to 29 questions, with each question comprising the presentation of some item to the student. The item is chosen in an adaptive way, that is, chosen based on the student’s previous responses during the assessment. (More specifically, the item is chosen to be maximally informative for the system’s evaluation of the student. The effect is that the assessment adapts to the level of the student, not necessarily becoming easier or harder for the student, as the assessment continues.) The student can elect to give an answer for the item, in which case her response is classified by the system as correct or incorrect, or she can choose to respond “I don’t know,” which she is encouraged to do if she has no idea how to approach the item. In addition, in each assessment, an *extra problem* is chosen uniformly at random from all of the items in the course and presented to the student as a question in the assessment. The student’s response to the extra problem does not affect the system’s evaluation of the student.

3. EXTRA PROBLEM BY RANK

For our first analysis, we will look at how responses (correct, incorrect, or “I don’t know”) to the extra problem evolve during the assessment. In other words, does the question rank of the extra problem have an effect on students’ responses? (By *question rank*, we mean the point in the assessment at which the question is asked, that is, the question number.) One hypothesis is that the extra problem success rate would increase throughout the assessment. (By *success rate*, we mean the proportion of the responses that are correct.) For example, it is possible that simply by working through repeated assessment questions, students experience a boost in performance; we will consider this phenomenon as a type of ‘testing effect’ [1, 4, 16]. One could imagine that this effect would be more pronounced after a long academic break, such as a summer or winter vacation, since the skills required for a particular course could suffer from a lack of recent use, and being assessed on these skills could help to sharpen them. As another example, there could be user interface issues for a student who is unfamiliar with the ALEKS system. Since the large majority of ALEKS problems require open-ended solutions, rather than multiple choice responses, it is possible that students would improve in performance as they became accustomed to the ALEKS interface. In both of these scenarios, the effect, if it existed, would seem to be more apparent earlier in a course, so we will look at data from ALEKS *initial assessments*, which are the assessments given at the start of an ALEKS course.

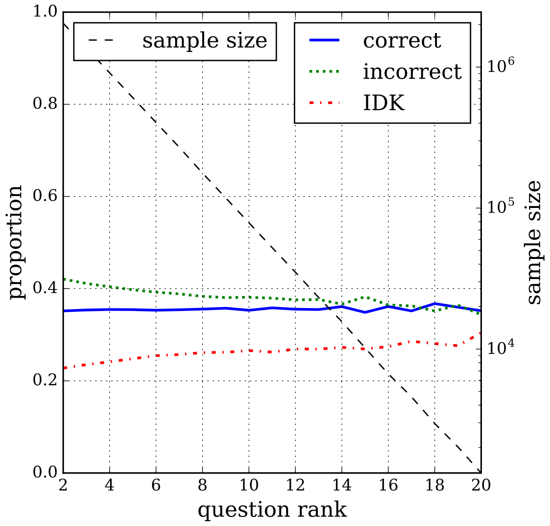

Figure 1: Proportions of responses to the extra problem by question rank for initial assessments. The types of responses are correct, incorrect, and “IDK” (“I don’t know”). Note that the sample size is shown on a logarithmic scale.

Note that both of the hypothesized effects in the previous paragraph would result in an increased extra problem success rate as the assessment progresses. However, one effect that would possibly lower the success rate, and that has been observed anecdotally, is that of assessment fatigue: as an assessment goes on, students may be more likely to respond incorrectly or not at all. This effect may be amplified by the open-ended answer interface used by ALEKS, which could make it more appealing for a student to respond “I don’t know” rather than make the effort to input a complete answer.

To start, we will look at a dataset consisting of 6,132,681 initial assessments, grouping the responses to the extra problem by question rank. The results can be seen in Figure 1. The first thing to note is that the success rate (the proportion of correct responses) does not increase as the assessment goes on; its curve is essentially flat. Thus, whatever testing effect there may be is overwhelmed by other factors. In particular, the rate at which students answer “I don’t know” shows a steady rise as the question rank increases, and the incorrect rate shows a corresponding decrease; keeping in mind that the extra problem is a randomly chosen problem that is asked at a randomly chosen point in the assessment, we see evidence that students are experiencing some sort of fatigue. As students get further along in the assessment, they seem less willing to attempt a problem and more will-

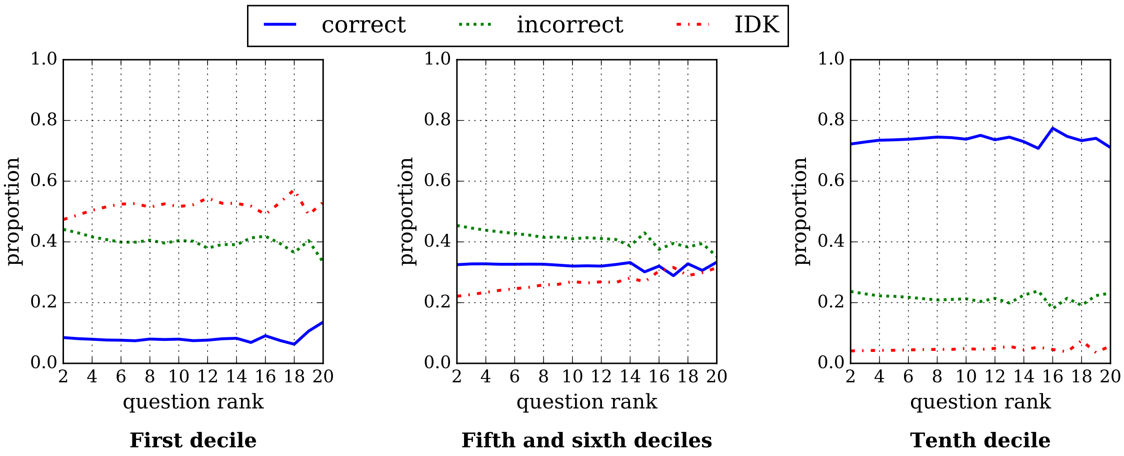

Figure 2: Proportions of responses to the extra problem by question rank for initial assessments, with percentage scores in (i) the first decile, (ii) the fifth or sixth decile, and (iii) the tenth decile.

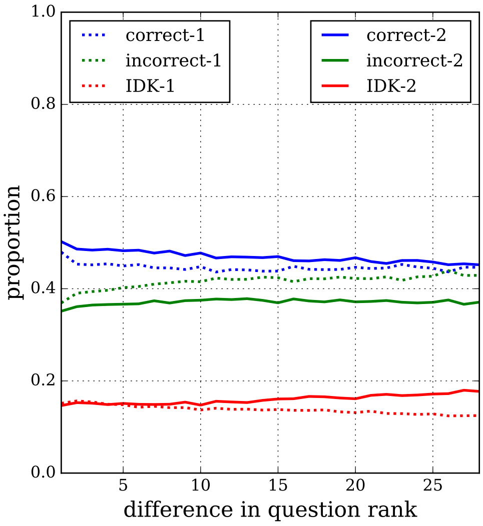

Figure 3: Proportions of responses for repeated items in initial assessments. The horizontal axis gives the difference in question rank between the two occurrences. The dotted curves (e.g., "correct-1") give the response proportions for the first occurrence, and the solid curves (e.g., "correct-2") give the response proportions for the second occurrence. The top set (pair) of lines represents the correct responses, the middle set represents the incorrect responses, and the bottom set represents the "I don't know" responses.

ing simply to respond "I don't know." What is striking is that, since the proportion of correct responses holds steady, it appears that many of these "I don't know" responses would have been incorrect responses earlier in the assessment; thus, one can alternatively interpret this as students being more "accurate" or "honest" in their self-assessment of the items they are capable of answering correctly.

To better understand these observed effects, we look more closely at the data based on the results of the initial assessment. We define the student's *initial assessment score* to be the percentage of the items in the course that are in the student's knowledge state according to the initial assessment, which gives a measure of the student's knowledge at the start of the course. Figure 2 shows the same results as in Figure 1, but this time separately for the three groups of students with initial assessment scores in (i) the first decile of all of the scores in the dataset, (ii) the fifth or sixth decile, and (iii) the tenth decile. From the plots in Figure 2, we can see that the (putative) fatigue effect is dependent on the group. The students in the middle group, with scores in the fifth and sixth deciles, seem to be most heavily affected, with a large increase in the "I don't know" rate as the assessment progresses. On the other hand, the students in the tenth decile show hardly any change over the course of the assessment, with the rates being mostly constant. Lastly, the students in the first decile are somewhere in the middle, with a sharp increase in the "I don't know" rate for the first few questions, and then a relatively flat curve thereafter.

4. REPEATED QUESTION

In the previous section, we saw that over the length of an assessment, the success rate was relatively flat. Thus, if there is any sort of boost from a testing effect, it is overwhelmed by other factors and is not apparent in our initial analysis. In the current section, we will take a more targeted approach and look at cases in which an item appears multiple times in an assessment. In particular, we will look at cases in which an item is first asked as an extra problem and then asked later in the same assessment as a "regular" question. (It is important to note that a different instance of the item is given each time, so that even though the type of problem being tested is the same, the particular example being presented is different.) Using a dataset composed of 644,462 initial assessments, each having some item repeated during the assessment, we can compare the success rates for the two occurrences of the repeated item. The results of this analysis are shown in Figure 3, where the horizontal axis gives the difference in question rank between the two occurrences. We can see that, overall, there is a gap between the success rates for the first and second occurrences, with the students being more successful on the second attempt. However, as with the analysis in Section 3, grouping the students by their initial assessment scores shows some pronounced differences. Figure 4 shows the results for students with initial assessment scores in the first decile; here,

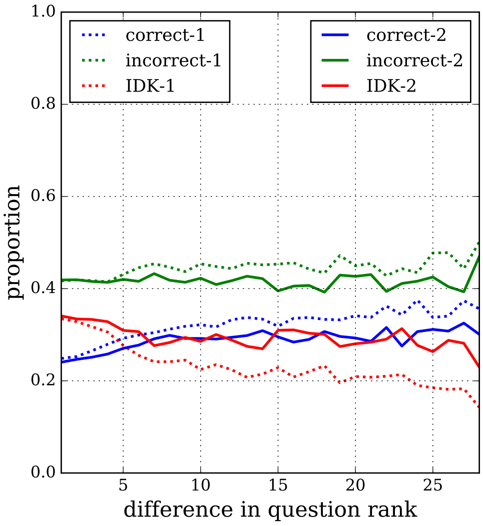

Figure 4: Proportions of responses for repeated items in initial assessments, for students with a percentage score in the first decile. The horizontal axis gives the difference in question rank between the two occurrences. Using the ordering at the leftmost edge of the horizontal axis, the top set (pair) of lines represents the incorrect responses, the middle set represents the "I don't know" responses, and the bottom set represents the correct responses.

in contrast to the overall trend, the students do worse on the second attempt. Interestingly, both the correct and incorrect rates *decrease* on the second attempt, with the "I don't know" rate showing a correspondingly large increase. Thus, it seems that the overall trend for students in this category is to be less confident, or at least less willing to attempt an answer, on their second attempt at a repeated item.

On the other hand, Figure 5 shows a much different trend for the students in the tenth decile. The "I don't know" rate is unchanged from the first attempt to the second, while a significant portion of the incorrect responses from the first attempt seemingly become correct responses in the second attempt. Thus, for students whose initial assessment scores are at the high end, it does appear that having multiple attempts at a problem gives a significant advantage.

As described in the previous section, the majority of students taking an initial assessment are returning from a break in schooling, often due to summer vacation. Thus, taking an ALEKS initial assessment may be one of the first chances in several months for a student to practice her math skills; in such a case, the simple act of working on an item may help the student recall some of the needed skills, or even to figure out new skills, which may then translate to greater success on a subsequent appearance of the item.

5. INNER FRINGE FORGETTING CURVE

In the next two sections we will examine forgetting as it applies to the ALEKS system. We will begin by looking at how the success rate of an inner fringe item changes as a function of the time since the item was first learned (with "learning" an item amounting to demonstrating a certain amount of success on the item in the learning mode). To do this, we will use data gathered from 286,345 ALEKS *progress assessments*, which are assessments given to a student after he has spent some time in the learning mode. The purpose of a progress assessment is to verify the student's recent learning. The progress assessments we examine here are limited to those for which the item presented as question 1 of the assessment is contained in the inner fringe of the student's knowledge state. Since the assessment is adaptive, we restrict our analysis to the first item presented to avoid any bias from the item-selection algorithm. We also look only at inner fringe items to reduce any bias that may come from the student working on items with related content: As mentioned, items in the inner fringe of a student's knowledge state are not required to master any of the other items in the knowledge state, so if an item is in the inner fringe, the student has not spent time learning new concepts that build on that specific item. For each of these progress assessments in which question 1 is an item appearing in the inner fringe of the student's knowledge state, we compute the number of days from the time the student learned the item to the time the item appeared in the progress assessment.

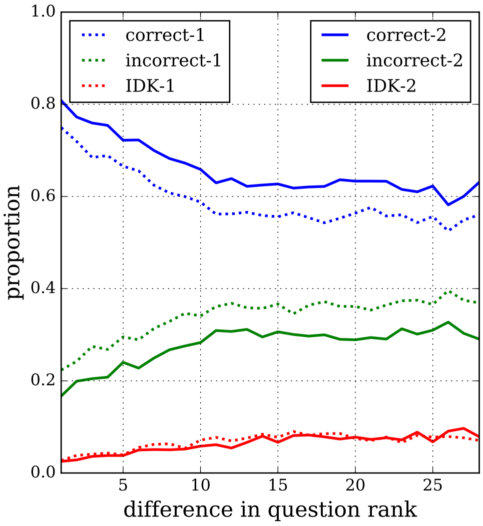

Figure 5: Proportions of responses for repeated items in initial assessments, for students with a percentage score in the tenth decile. The horizontal axis gives the difference in question rank between the two occurrences. The top set (pair) of lines represents the correct responses, the middle set the incorrect responses, and the bottom set the "I don't know" responses.

The results are in Figure 6. In this figure, the solid curve (the one near the top of the figure) can be considered a forgetting curve [2, 6]. As shown, there is a clear decrease in the success rate as the number of days since the item was learned increases, while the rates of incorrect and "I don't know" responses both increase. The changes are greatest over the initial few days and then flatten out somewhere between one and two weeks. As an aside, we can also see in Figure 6 the weekly cycle of student use, which causes the sample size to peak every seven days.

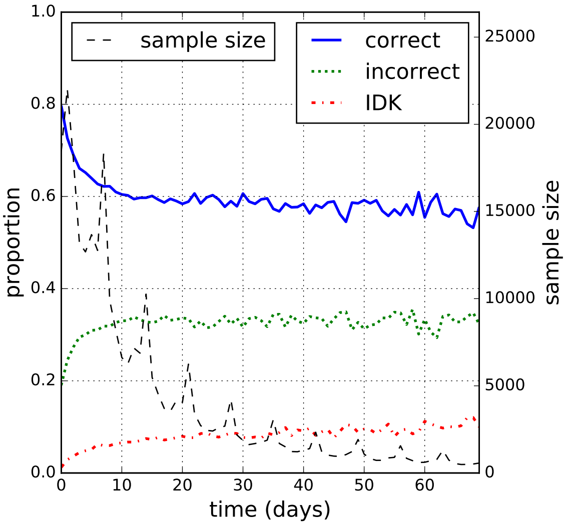

Figure 6: Proportions of responses as a function of the time (in days) since the item appearing as question 1 in a progress assessment was learned.

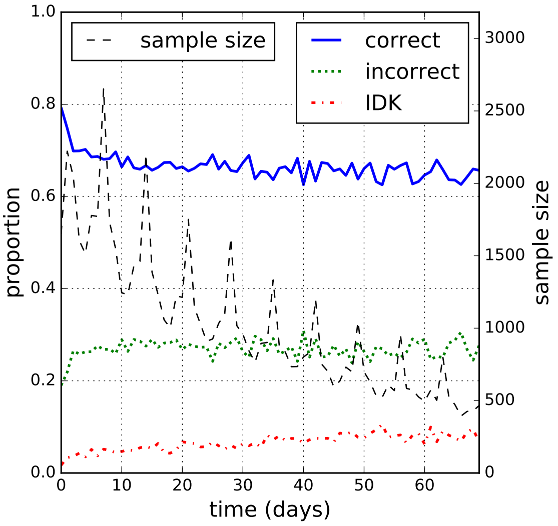

Figure 7: Proportions of responses as a function of the time (in days) since the extra problem appearing in a progress assessment was learned.

6. EXTRA PROBLEM OVER TIME

For the next part of our analysis, we again use data generated by ALEKS progress assessments. Rather than looking at the first item presented, however, we instead focus on the extra problem. Restricting our analysis to extra problems that have been previously learned by the student, we can again look at the response rates as a function of the time since the item was first learned. Using data from 72,045 progress assessments that fit the criteria, we show the results in Figure 7. (Furthermore, for ease of comparison, we display the information from Figures 6 and 7 in Figure 8.)

While there is a drop in the success rate over the first few days, in comparison to Figure 6 this drop is less pronounced, and it levels off within a shorter amount of time. The reason for this is most likely that we are no longer looking only at items in the inner fringe of the student's knowledge state. Recall that, if an item is in the student's inner fringe, then the student has not (at least in theory) mastered any subsequent material that requires complete mastery of that item. However, this no longer holds for a randomly chosen item from the student's knowledge state; for example, the student may have mastered one or more subsequent items that require complete mastery of the extra problem, which may have the effect of reinforcing the learning of the concepts in the extra problem. Thus, the flatter nature of the extra problem forgetting curve can be viewed as a consequence of the adaptive nature of the ALEKS system, which serves to reinforce the original learning.

On the other hand, the success rate on the extra problem does exhibit a noticeable decline over the first several days after the item is learned. It is during this period that more targeted review and/or practice may be beneficial.

7. DISCUSSION AND FUTURE WORK

In the above analyses, we observed the following: (1) Students, especially those near the middle of the range in content knowledge, tend to replace incorrect responses with ones of "I don't know" as the assessment progresses; (2) Students at the upper end in content knowledge tend to improve on an item the second time the item is asked in an assessment, while students on the lower end tend to do worse the second time, or at least to become less confident; (3) Items that give the "high points" of a student's learning see a more precipitous drop in the student's memory than do other items (perhaps because the skills and concepts in these "high point" items may not have been practiced or developed much since the original learning event). A possible improvement to the ALEKS learning and assessment software based on these observations may be to introduce pointed feedback during an assessment to provide encouragement or guidance to students who are at risk of fatiguing or declining in confidence. Another may be to have a dedicated review period for "high point" items, perhaps given in conjunction with a progress assessment itself, to help with immediate forgetting.

In addition to suggesting improvements to the ALEKS system, our analyses may both inform and be informed by the extensive literature on memory. Take, for example, the particular forgetting curve analysis in [2], in which the authors examine models of forgetting given by

$$R(t) = a + (1-a) \times b \times P(t), \quad 0 < a, b < 1, \tag{7.1}$$

for different functions $P(t)$. Here, $R(t)$ gives the probability of retention at time t, and a and b are parameters. One such function $P(t)$ examined in [2] is

$$P(t) = (1+t)^{-\beta}, \tag{7.2}$$

in which $\beta > 0$ is a parameter. Fitting $R(t)$ (with this form of $P(t)$) to the success rates for question 1 and extra problem data gives the smooth curves shown in Figure 8. The fit is strong, with the $R(t)$ curves closely following the trend of the data. (The increasing jaggedness of the correct curves in Figure 8 stems from the decreasing sample sizes, as shown in Figures 6 and 7.) For reference, we report that for this fit, the parameters a, b and β are estimated to be 0.55, 0.56

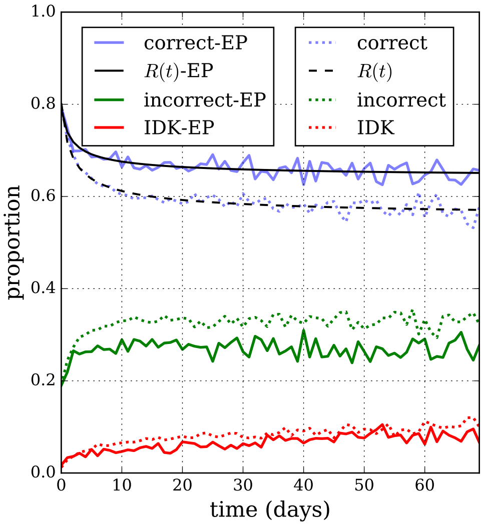

Figure 8: A direct comparison of Figures 6 and 7. The solid curves (e.g., "correct-EP") are from Figure 7, giving the proportions of responses as a function of the time (in days) since the extra problem was learned. The dotted curves are from Figure 6, giving the proportions of responses as a function of the time (in days) since the item appearing as question 1 was learned. Also shown are the curves obtained from fitting $R(t)$ given by Equation (7.1) (with $P(t)$ as in (7.2)) to the data.

and 0.59, respectively, for the question 1 curve; for the extra problem curve, these parameters are estimated to be 0.64, 0.42 and 0.58, respectively.

It is a natural next step to implement such a model to improve students' experiences using ALEKS by improving, for example, the scheduling of progress assessments, the item-selection algorithm, and the timing and content of review periods for newly learned items.

Further, it is feasible that the very large data sets examined in this paper may contribute to the discussion of competing mathematical models of forgetting. For example, the authors in [2] also examine forgetting functions of the form

$$R(t) = a + (1 - a) \times be^{-\alpha t} \tag{7.3}$$

and of the form

$$R(t) = 0.116 + (1 - 0.116) \times b \times (1 + \gamma t)^{-\beta}, \tag{7.4}$$

comparing the various special cases of (7.1) given by (7.2)-(7.4). Our data would likely contribute to this and similar discussions.

8. REFERENCES

[1] Agarwal, P., Bain, P., and Chamberlain, R. 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.

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

[3] Baddeley, A., Eysenck, M. W., and Anderson, M. C. *Memory.* Psychology Press, New York, 2009.

[4] Carrier, M., and Pashler, H. The influence of retrieval on retention. *Memory and Cognition 20* (1992), 632–642.

[5] Doignon, J.-P., and Falmagne, J.-C. Spaces for the assessment of knowledge. *International Journal of Man-Machine Studies 23* (1985), 175–196.

[6] Ebbinghaus, H. *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).

[7] Falmagne, J.-C., Albert, D., Doble, C., Eppstein, D., and Hu, X., Eds. *Knowledge Spaces: Applications in Education.* Springer-Verlag, Heidelberg, 2013.

[8] Falmagne, J.-C., and Doignon, J.-P. *Learning Spaces.* Springer-Verlag, Heidelberg, 2011.

[9] Grayce, C. A commercial implementation of knowledge space theory in college general chemistry. In *Knowledge Spaces: Applications in Education*, J.-C. Falmagne, D. Albert, C. Doble, D. Eppstein, and X. Hu, Eds. Springer-Verlag, 2013, ch. 5, pp. 93–114.

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

[11] Huang, X., Craig, S., Xie, J., Graesser, A., and Hu, X. Intelligent tutoring systems work as a math gap reducer in 6th grade after-school program. *Learning and Individual Differences 47* (2016), 258–265.

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

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

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

[15] Reddy, A., and Harper, M. Mathematics placement at the University of Illinois. *PRIMUS 23* (2013), 683–702.

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

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

[18] Taagepera, M., and Arasasingham, R. Using knowledge space theory to assess student understanding of chemistry. In *Knowledge Spaces: Applications in Education*, J.-C. Falmagne, D. Albert, C. Doble, D. Eppstein, and X. Hu, Eds. Springer-Verlag, 2013, ch. 6, pp. 115–128.

[19] Yang, Y., Leung, H., Yue, L., and Deng, L. Automatic dance lesson generation. *IEEE Transactions on Learning Technologies 5* (2012), 191–198.