A practical perspective on knowledge space theory: ALEKS and its data

Eric Cosyn, Hasan Uzun, Christopher Doble, Jeffrey Matayoshi
McGraw Hill/ALEKS Corporation, Irvine, CA, United States

Abstract

The ALEKS (Assessment and LEarning in Knowledge Spaces) educational software system is an instantiation of knowledge space theory (KST) that has been used by millions of students in mathematics, chemistry, statistics and accounting. The software employs a probabilistic assessment based on KST for placement into an appropriate course or curriculum, a learning mode in which students are guided through course material according to a knowledge structure, and regularly spaced re-assessments which are also based on KST. In each of these aspects, the interactions of the student with the system are guided by the theory and by insights learned from student data. We present several relationships between theory and data for the ALEKS system. We begin by surveying the ALEKS system and examining some practical aspects of implementing KST on a large scale. We then study the effectiveness of the ALEKS assessment using both standard statistical measures and ones adapted to the KST context. Finally, we examine the learning process in ALEKS via statistics for the learning mode and its associated knowledge structures.

Keywords: knowledge space theory, adaptive assessment, intelligent tutoring system, layers of a knowledge state

1. Introduction

Knowledge space theory (KST) is an approach to the assessment of knowledge that is based on a combinatoric and probabilistic model introduced by Doignon and Falmagne (1985). KST has inspired a large literature which includes hundreds of peer-reviewed journal articles and several books; a bibliographic database is maintained by Hockemeyer (2020) at the University of Graz in Austria. ALEKS, which stands for **A**ssessment and **LE**arning in **K**nowledge **S**paces, is a web-based intelligent tutoring system based on KST that is designed to assess and instruct students in mathematics, chemistry, statistics and accounting. ALEKS has been available commercially since the late 1990s, and currently, about four to five million students use it each year. During the two-decade existence of ALEKS, a great deal of data from student interactions with the system have been collected. These data offer insights into this large-scale implementation of KST; the goal of this paper is to examine the principal aspects of ALEKS in light of these data.

A number of studies have been done that scrutinize the relationship between the use of ALEKS and measures external to ALEKS, such as success in a course or performance on a standardized test. Among

these studies, one finds Hagerty et al. (2010); Mojarad et al. (2018) for college mathematics, Baxter and Thibodeau (2011) for accounting, Hickey et al. (2020) for college chemistry, Hu et al. (2007) for college statistics, Huang et al. (2016); Craig et al. (2013, 2011) for middle school mathematics, and Reddy and Harper (2013) for college course placement; see also Fang et al. (2019) for a meta-analysis. The present paper differs from these studies in that it focuses on measures internal to the ALEKS system, such as the system's success in predicting a student's performance on a particular ALEKS question or her readiness to learn new ALEKS content. Such a focus is meant to provide a pointed analysis of the workings of KST as implemented in ALEKS.

The paper is organized as follows. In the remainder of this section, we give background on both KST and ALEKS, including the relationship between KST and assessment and learning in ALEKS. In Section 2, we first examine the effectiveness of the ALEKS assessment using standard statistical measures. Later in the section, we again examine the ALEKS assessment, but this time using measures more firmly rooted in KST, namely, measures derived from the 'layers' of a knowledge state. In Section 3 we focus on the learning aspect of the ALEKS system, first surveying how efficiently students learn new material, and then analyzing the retention and forgetting of learned material, again featuring the notion of knowledge state layers in the latter case. Section 4 has a short recapitulation.

1.1. Background on knowledge space theory

At the core of KST is the concept of an *item*; for the context of ALEKS described in this paper, an item is a discrete concept or granular topic appearing as part of an academic course. For example, an item for an introductory high school algebra course might be *"Solving a compound linear inequality"*, and another might be *"Solving a word problem with two unknowns using a linear equation"*. An item is actually a collection of examples, called *instances*, each focused on the same, narrow topic, and designed to be equal in difficulty. One instance of the item *"Solving a compound linear inequality"* might be "Solve the compound inequality $4x + 5 \leq 21$ and $2x + 5 > 3$," while another instance might be "Solve the compound inequality $4x - 5 > -5$ or $3x + 2 \leq -13$." Each time the student is presented a given item, whether in an instructional setting or in an assessment, a different instance is chosen.

There are several hundred items that make up a typical academic course, and having the knowledge and skill to successfully complete all of the items means (according to KST) mastery of the course. A *knowledge state*, or simply *state*, is a particular set of items that some student could be capable of completing correctly. Denoting the set of all items by Q, a pair $(Q, \mathcal{K})$ is a *knowledge structure* if $\mathcal{K}$ is a family of subsets of Q containing all the knowledge states that are feasible, that is, that could characterize some student in the population. In other words, a student whose knowledge state is K can, in principle, correctly complete all the items in K and would fail to correctly complete any item not in K. (We say "in principle" because, for example, there is the possibility of careless errors, also called slips.) The family $\mathcal{K}$ always includes the empty

set (characterizing a student in the state of complete ignorance) and the set Q (characterizing a student mastering all of the items).

Typically in KST, including in ALEKS, the number of states in the knowledge structure is much less than the number of subsets of Q. This is due in part to the inherent relatedness of items, as some items are prerequisites of other items. For example, it would be very unlikely, or even impossible, for a student to have mastered *"Solving a compound linear inequality"* without having mastered *"Solving a one-step linear equation."* For one particular ALEKS course (College Placement, described in this section below) comprised of 314 items, the number $|\mathcal{K}|$ of states in the knowledge structure is about 10^{23}, which is considerably smaller than $2^{314} \approx 10^{94}$, the number of subsets of a set with 314 elements. The construction of the knowledge structure will not be described here; see instead, e.g., Desmarais et al. (1995); Desmarais and Pu (2005); Koppen and Doignon (1990) and Chapters 15 and 16 in Falmagne and Doignon (2011).

Additional KST concepts that are important for the ALEKS system are the 'outer fringe' and the 'inner fringe' of a knowledge state. The *outer fringe* of a state K is the set of items q not in K such that $K \cup \{q\}$ is also a knowledge state. The outer fringe of a student's state may be thought of as the set of items the student is 'ready to learn.' The *inner fringe* of a state K is the set of items q in K such that $K \setminus \{q\}$ is also a knowledge state. That is, the inner fringe is made up of the items representing the 'high points' of the student's competence. If the knowledge structure is a *learning space* (see Definition A.1 in Appendix), as is the case in ALEKS, the knowledge state of a student is completely determined by the inner fringe and the outer fringe of the state (Theorem A.2). The fringes are put to use in the functioning of ALEKS, as described below.

The next section gives a brief overview of the ALEKS system. It is followed by a discussion of the implementation of KST in the ALEKS learning and assessment features. As in any other software platform, there are ongoing improvements made in ALEKS. Our aim is not to cover all the specifics of the current iteration of ALEKS, but rather to highlight certain processes and to describe the handling of some practical challenges that arise in such a large-scale implementation of KST.

1.2. A quick overview of the ALEKS system

A student begins her experience with ALEKS by enrolling in an ALEKS course. There are ALEKS mathematics courses ranging from third grade mathematics to college precalculus, and there are introductory college chemistry, statistics, and accounting courses. The full list of ALEKS courses and their sets of items are available at McGraw Hill (2020). When an instructor sets up an ALEKS course to be taken by a class of students, she has the option of selecting which items, among all the items available in the domain associated with the course, will make up the actual course content. Typically, an actual course consists of about 300 to 600 items.

Once enrolled, the student takes an *initial assessment* in ALEKS. The assessment aims to determine the student's knowledge state at the onset of the course. From the state resulting from the initial assessment, the student's outer fringe is determined and serves as the student's entry point in the ALEKS learning mode. The student chooses an item from her outer fringe and practices instances of the item. She is given feedback and access to explanations for those instances. If she performs well enough on those instances, ALEKS temporarily considers the item to be learned. The item is added to the student's state, the outer fringe is re-computed, and the student chooses another outer fringe item to practice.

Once the student has practiced a certain number of items, or once she has worked in the learning mode for a certain amount of time since her previous assessment, ALEKS gives her a *progress assessment* to confirm her learning. Based on this assessment, ALEKS updates the student's state. The student is again placed in the learning mode, and the process repeats—there is a cycle of learning and assessment until the student has worked through the course, at which time she may take a final assessment. The cycle of learning and assessment associated with the progress assessment acts as a form of *retrieval practice* (also known as *testing effect* or *test-enhanced learning* (Roediger III and Butler, 2011)). Numerous studies have shown that being forced to actively recall information helps to solidify that information in long-term memory (Bae et al., 2019; Rawson and Dunlosky, 2011; Roediger III and Butler, 2011; Roediger III and Karpicke, 2006a,b). Another benefit of the progress assessment is that, as it moves from one item to another outside the student's control, it enforces *interleaved practice* on items that were previously learned mostly through *massed practice* (see Dunlosky et al. (2013) for the benefit of interleaved practice over massed practice). These benefits make the interplay between assessment and learning a core feature of the ALEKS system.

Four ALEKS courses are highlighted in this paper. They have been selected as representative of the mathematics curriculum covered by ALEKS and for the abundance of data they provide. One of the courses is Sixth-Grade Math, which is either the last mathematics course taken in elementary school or the first mathematics course taken in middle school; as such, it is taken by students roughly ages 11-12. Another course is Algebra I, which is usually taught in ninth grade, and so is taken by students ages 14-15. A third course is College Algebra, which is a higher education course that is often a graduation requirement for students not in science or engineering majors, unless their high school credentials exempt them from it. We will call these three courses—Sixth-Grade Math, Algebra I, and College Algebra—*learning courses* because students taking them are not expected to have much knowledge of the course material upon starting the course, and the focus is on helping students learn new material. This is in contrast to the last course we highlight, College Placement, which is a placement test for incoming college students. It recommends the highest mathematics course that the student is likely to be successful in, from Basic Math (an arithmetic course) to first-year Calculus. If a student is not satisfied with her placement recommendation, she has the option to retake the test, but not before filling her gaps by practicing on a specific course also recommended

by the test. Another particularity of College Placement is that its item content is not customizable by the instructor.

Table 1 reports a few usage statistics for the four courses. The data are for students who took their initial assessment over the period from August 15 to September 30, 2018, for the three learning courses, and over the period from March to September, that same year, for College Placement. The data report the students' activity over the length of the course, which is a full academic year for Sixth-Grade Math and Algebra I, and a semester or an academic quarter for College Algebra. For College Placement, a retake of the test, when it occurs, happens typically within one or two weeks of the original test.

Table 1: Usage statistics for four ALEKS courses. The numbers are averages per student, with standard deviations in parentheses. These statistics cover several months in 2018-19 (as described in the text).

Course	*N*	**Number of items in actual course**	**Number of assessments**	**Number of items learned**	**Total time in hours**
Sixth-Grade Math	99,482	385.3 (68.1)	4.9 (3.6)	128.2 (110.8)	15.6 (14.8)
Algebra I	63,097	523.1 (172.0)	4.1 (3.6)	104.3 (112.9)	13.7 (16.9)
College Algebra	19,131	301.5 (123.7)	5.0 (4.0)	145.2 (113.6)	31.4 (27.1)
College Placement	548,391	314.0 (0.0)	1.3 (0.7)	-	2.1 (1.7)

1.3. Knowledge space theory as implemented in ALEKS

The above is only a brief overview of the ALEKS experience, with many details left out. What follows are discussions of some specifics of the ALEKS assesssment and learning mode, with a few practical challenges highlighted and their solutions outlined.

At the heart of any ALEKS course is the knowledge structure. As mentioned earlier, we will not cover here the construction of the knowledge structure, but we simply note that everywhere in the following, when we mention the knowledge structure (and by extension its knowledge states and operations on them), we actually mean the *projection* (Definition A.3 and Theorem A.4) of the knowledge structure on the subset of items chosen by the instructor for the course.

An ALEKS assessment can be viewed as a probabilistic search among all of the feasible states to uncover the student's latent state. Such an algorithm starts with some initial probability distribution on the states (Definition A.6). Then an item that roughly splits the distribution into two equal parts is selected. This item is such that the probabilities of the states that contain it add to about 0.5, and we say that the item has a likelihood close to 0.5. After each response from the student, the probabilities of the states are updated based on the response. For a correct response, the update results in an increase of probability for the states containing the item and a decrease for the other states. For an incorrect response, the update goes

in the other direction, with the probabilities of the states not containing the item being increased and the remainder decreased. For each subsequent question, the algorithm selects an item that has a likelihood close to 0.5, and the probabilities are updated based on the student's response. The assessment proceeds until a single knowledge state emerges with a very high probability[1]. Section 13.4 of Falmagne and Doignon (2011) provides a formal description of the principles outlined here.

Such an algorithm assumes that all the states in the knowledge structure can be listed and assigned a probability to begin the assessment. It also assumes that the probabilities of the states can be updated in a timely fashion, with the likelihoods of items computed from the states and then used in selecting the most informative question after each response. However, the sizes of the knowledge structures in ALEKS create several challenges to these assumptions. For most ALEKS knowledge structures, there are too many states to list even with modern computing power, and updating the probabilities of the states between questions is not feasible. To overcome this, the ALEKS system currently employs an assessment algorithm similar to the one outlined in Section 8.8 of Falmagne et al. (2013). In this algorithm, the set of items in the course is partitioned into several subsets, and the assessment is run in parallel, simultaneously on these subsets. Each subset gets its knowledge structure via projection from the full knowledge structure, and each of these knowledge substructures has a size that allows for the listing of its states, and for the probabilities of the states to be computed. Following the student's answer to the selected item, the probabilities of the states for the subset to which the item belongs are updated as described above. A key feature of the algorithm is its ability to carry the information gained from the student's answer to the other subsets (to which the item does not belong) and to update the probabilities for the states in these subsets.

One optimization used by the assessment is to perform its search on a subset of the knowledge structure that is relevant to that particular assessment. For example, initial assessments are taken by students who may have mastered the prerequisite material for the course (some of which is typically part of the domain of knowledge for that course) but not necessarily much more. The initial assessment thus performs the search on a smaller but very representative set of knowledge states that reflects this information. On the other hand, the focus of progress assessments is quite different. They are not designed for placement from scratch but mostly for testing the recent learning progress of the student. Consequently the progress assessment runs a local search, essentially in the *neighborhood* (Definition A.5) of the knowledge states recently crossed by the student. For both types of assessments, the probability distribution on the knowledge states at the onset of the assessment is not uniform but is instead informed by past assessment data from the course.

[1]Note that, because of the probabilistic nature of the assessment procedure, this final state may contain an item to which the student gave an incorrect response. Such a response is regarded as due to a careless error. ALEKS items typically have open-ended answers, in the sense that the system avoids multiple choice formats and instead uses answer input tools that mimic what would be done with paper and pencil. As such, the lucky guess probability may be assumed to be very small.

Even the most efficient assessment algorithm needs to ask more questions than practical to converge to a single state, especially in the case of the initial assessment. For the College Placement course, the knowledge structure contains about $10^{23} \approx 2^{77}$ states. If all states were equally likely at the beginning of the assessment, an algorithm that could eliminate half of the remaining states after each question would still require 77 questions in the worst case to converge to a single state. It is not practical to ask a student this many questions in an assessment. Based on the feedback from students and instructors over the years, the number of questions in an initial assessment has been capped at 30. This number strikes a balance between the need to gather enough information about the student's knowledge state and the risk of overwhelming the student with too many questions.[2] All assessments also ask a randomly selected 'extra problem' used for testing purposes (see Section 2.1 below). So there are actually up to 29 questions in an initial assessment that are selected adaptively and used for uncovering the student's state.

In the ideal case, an assessment concludes once all the items can be categorized as either likely to have been mastered by the student ('in-state') or likely not to have been mastered by the student ('out-of-state'). Typically, an item falls in the former category if its likelihood is above 80% and the latter category if its likelihood is less than 20%. Most often, however, the assessment ends because it reached the maximum number of questions. While most items will then have a likelihood very close to either 1 or 0, the items not yet categorized as in-state or out-of-state will be categorized as 'uncertain'. The state assigned to the student at the end of the assessment does not include these uncertain items and so may underestimate the student's latent state. After the initial assessment, the learning of such uncertain items is fast-tracked (see below) and allows the new student to begin the learning process with easier material. As learning progresses, the initial underestimation quickly disappears.

Another distinction relevant to our discussion concerns the updating rule of the assessment. After an incorrect response, the probabilities of the states that do not contain the item are increased, while the probabilities of the remaining states are decreased. However, this update cannot be very aggressive given that there is a non-negligible chance for a careless error. This fact motivated the addition for each question of a button labeled "I DON'T KNOW" (or "I HAVEN'T LEARNED IT YET"), which the student can choose if she has no familiarity with the item. Selecting the button results in a substantial increase in the probabilities of the states not containing the item, thereby decreasing the total number of questions required to uncover the student's state. On the other hand, the open-ended nature of an ALEKS question makes lucky guesses very unlikely. So the update following a correct answer is also substantial[3].

[2]Regarding the latter concern, see Matayoshi et al. (2018) for evidence of a 'fatigue effect' experienced by students in ALEKS assessments.

[3]See Definition 13.4.4 in Falmagne and Doignon (2011) for a formal description of the updating rule. Remark 13.4.5 discusses a Bayesian interpretation of the update parameters of the rule that links them to the lucky guess and careless error rates of the items. In agreement with overall empirical estimates of these error rates, the update parameters in ALEKS are about 35 for a

We conclude this section with some specifics about what constitutes the learning of an item in the student's outer fringe. In the ALEKS learning mode, the student has access to a graphical list that presents all the items in the student's outer fringe and from which the student can select any item. Each item in the list comes with a 'learning progress bar' and a numerical value, or 'score', summarizing it. The initial score of an item is 0. When the student selects a new item to work on, an instance of that item with its explanation is presented. After reading the explanation, the student receives another instance for actual practice. Each time the student receives a new instance, she can either try to answer it or read the explanation for that instance. Afterward, the student gets another instance to practice, and so on. As the student practices the item, its score is updated according to the following rules.

- A correct answer increases the score by 1, but the second correct answer on a streak of two consecutive correct answers increases the score by 2 (for a total of 3).
- An incorrect answer decreases the score by 1, unless the score was already at 0.
- Reading an explanation does not change the score. However, reading the explanation when the new instance follows a correct answer breaks the streak of correct answers. So there is an implicit encouragement to try to answer without falling back on the explanation.

An item is considered to be (provisionally) learned, and is added to the student's knowledge state, when the student reaches a preset target score. Most often an item requires a target score of 5 to be learned. There are two cases when a target score of 3 is enough and the learning is fast-tracked. The first case is when the item is classified as uncertain following the initial assessment. The other case is when the item has been learned previously but then removed from the student's knowledge state following a progress assessment. Five consecutive incorrect answers is considered a failed learning attempt. In that case, the student is gently prompted to try another item.

2. Evaluating the ALEKS assessment

In this section we give a detailed evaluation of the ALEKS initial assessment. We choose the initial assessment because it is the most challenging type of assessment for the system, as ALEKS has little to no information about the student to begin. Though the initial assessment for each ALEKS course runs via the algorithm described in Section 1.3 above, the student's experience in taking an initial assessment will vary based on the level of the student, the set of items in the particular course, the knowledge structure in place for those items, and other factors. With this in mind, we examine here the initial assessment for

correct answer, 5 for an incorrect answer, and 50 for "I don't know", with some variation from one course to another.

three courses that differ in a number of such factors. The three courses, which are introduced in Section 1.2, are Sixth-Grade Math (which covers mostly basic arithmetic and geometry), College Algebra (which covers exponentials and logarithms, matrices, and polynomial and rational functions), and College Placement (which covers areas such as proportions and percentages, integer arithmetic, linear and quadratic functions, polynomial and rational functions, exponentials and logarithms, and trigonometry). We note that College Placement has the most student use of any ALEKS course, involves the greatest diversity of student ability, and has an average initial assessment result nearest the "middle" of the course. For these reasons, the College Placement initial assessment is an especially useful and important one to study, and it is examined further in Section 2.3 below.

In the analyses that follow in Sections 2.2 and 2.3, data from over 3.1 million initial assessments gathered from 2012 until early 2020 were used. Detailed demographic information about the students, such as information regarding gender, age, geographic location, or college major, is not available. However, these ALEKS courses are taken predominantly by students in the United States, and there is roughly one student represented per initial assessment taken. (Some students in the College Placement sample took more than one initial assessment.) For the College Placement sample, each of the students was enrolled or soon-to-be enrolled at a college or university and took the assessment for the purpose of being placed in an appropriate mathematics course.

2.1. A key statistic: the extra problem in assessment

In each ALEKS assessment, one item is chosen uniformly at random from the set of items being used in the course, and the item is presented to the student as an assessment question. For the student, nothing distinguishes that *extra problem* from the other problems in the assessment. The answer to the extra problem is ignored by the adaptive assessment and does not affect its results. Rather, the extra problem and its answer are stored separately and later used to evaluate and improve the ALEKS system. Throughout this section and the next one, we will resort extensively to statistics built upon the extra problem to evaluate miscellaneous aspects of the ALEKS system.

2.2. Evaluating the assessment using standard measures

Section 2.2 treats an ALEKS assessment as a probabilistic classifier that attempts to predict, for each item, whether or not the item is contained in the student's knowledge state. Thus, for the most part, rather than looking at aspects of the assessment specific to KST, our intent is to evaluate the performance of the assessment using standard measures that can be applied to any model that makes probabilistic classifications. To perform this evaluation, we use the following procedure. At all times during an assessment, the ALEKS system has, for each item, an estimated probability that the item is contained in the student's state. To evaluate these estimates, we can compare the probabilities with the student responses to the extra problems;

in what follows, we encode these responses as either "1" (for a correct answer) or "0" (for an incorrect answer or an "I don't know" response). It is worth noting that each probability estimate from the ALEKS system corresponds to the likelihood that the student *knows* an item, rather than being an estimate of the probability that the student answers the item correctly. However, without access to the student's true knowledge state, we use the actual responses as a reasonable proxy to evaluate the probability estimates.

For the initial assessment for each of the three courses, we use the probability estimates returned at the end of the assessment to evaluate the following measures: the area under the receiver operating characteristic curve (AUROC), the point biserial correlation, and the accuracy. The AUROC is commonly used in evaluating probabilistic classifiers, and one of its strengths is that it is not sensitive to class imbalances (Fawcett, 2006). The point biserial correlation is a special case of the Pearson correlation coefficient in which one variable is dichotomous (i.e., the student response) and the other variable is continuous (i.e., the probability estimate from the ALEKS system). The AUROC and point biserial correlation use the exact probability scores. For the accuracy computation, any probability at or above 0.5 is assigned to the positive class (a correct answer) and anything below 0.5 is assigned to the negative class (an incorrect answer or an "I don't know" response); note that this assignment of classes is a standard procedure used to evaluate binary classifiers which does not necessarily correspond to the actual classifications by the ALEKS assessment. (We analyze the actual classifications of the assessment in more detail shortly.)[4]

The results are shown in Table 2, where *extra problem correct rate*, or simply *correct rate*, refers to the proportion of all responses to the extra problem that are graded as correct by the ALEKS system. We see that the strongest overall results are from the College Placement assessment. This fact is not unexpected, as College Placement is a specialized assessment that has the sole purpose of placing a student into the appropriate college-level mathematics course; as such, the items used in the course are chosen for being quality assessment questions. On the other hand, for the College Algebra and Sixth-Grade Math courses, which are learning courses, the emphasis is on fully covering the subject matter so that students can have a complete learning experience. Because of this, the items in these latter courses are designed to be used for both assessment *and* learning purposes; a natural consequence of this split focus is that the overall assessment performance of the items, while still acceptable, is not quite as strong as in the College Placement assessment.

Next, in Table 3 we have partitioned the data points based on the actual classification made by the assessment (either in-state, out-of-state, or uncertain), for all three of our example ALEKS courses. For

[4]We also evaluated the performance of the assessment using the Matthews correlation coefficient (Matthews, 1975), which is recommended as a good overall measure for evaluating binary classifiers (Boughorbel et al., 2017; Chicco, 2017; Powers, 2011). However, as the results were very similar to those obtained from the use of the point biserial correlation, we have omitted the Matthews correlation from our analysis. This similarity is perhaps not too surprising, as the Matthews correlation is another special case of the Pearson correlation coefficient.

Table 2: Statistics for ALEKS initial assessments

Course	N	**Correct rate**	**AUROC**	**Point biserial**	**Accuracy**
Sixth-Grade Math	162,900	0.392	0.875	0.634	0.801
College Algebra	174,073	0.322	0.863	0.602	0.808
College Placement	2,775,432	0.509	0.889	0.671	0.814

Table 3: Statistics for ALEKS initial assessments, partitioned by item classification

Course	**In-state**		**Out-of-state**		**Uncertain**	
	Proportion	**Correct rate**	**Proportion**	**Correct rate**	**Proportion**	**Correct rate**
Sixth-Grade Math	0.333	0.794	0.497	0.110	0.170	0.428
College Algebra	0.283	0.740	0.610	0.110	0.107	0.427
College Placement	0.474	0.838	0.358	0.102	0.168	0.449

each classification category, Table 3 includes the proportion of items in the category, as well as the extra problem correct rate for those items. Again, we can see that College Placement has the best performance; for the in-state items, it has the highest correct rate (if an item is classified as in-state, it is desirable that the correct rate be high), while it then has the lowest correct rate for the out-of-state items (if an item is classified as out-of-state, one would expect it not to be answered correctly very often). Note that for the three courses, the correct rates for the uncertain items range between 0.42 and 0.45; assuming that there is some amount of careless error in these responses, it seems plausible that the proportion of these uncertain items actually known by students is somewhere around 0.5, which is the desired value. If, instead, the proportion of uncertain items actually known by students were very high or very low, this would be a sign that the classifications made by the ALEKS assessment are not well-calibrated.

For our next two analyses, we focus on College Placement which, for the reasons outlined in the introduction to Section 2, is a natural choice for an in-depth analysis. First, Figure 1 shows the evolution of the AUROC, point biserial, and accuracy values over the course of the assessment. Specifically, for the 2,688,472 College Placement assessments in the dataset that run for the full 29 questions, the values of the three measures are plotted at each point in the assessment (i.e., after each question). As shown, the values converge early in the assessment, with the changes being minimal after question 10. Essentially, this means that the assessment quickly obtains a fairly accurate picture of the student's knowledge, and then uses the remaining questions to fine-tune this picture.

We next look at the performance of the College Placement assessment at the level of the individual

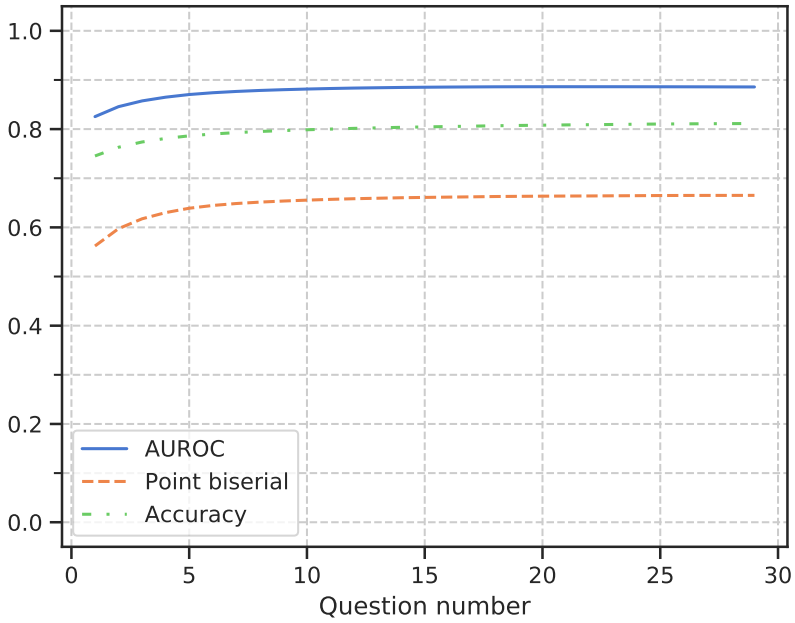
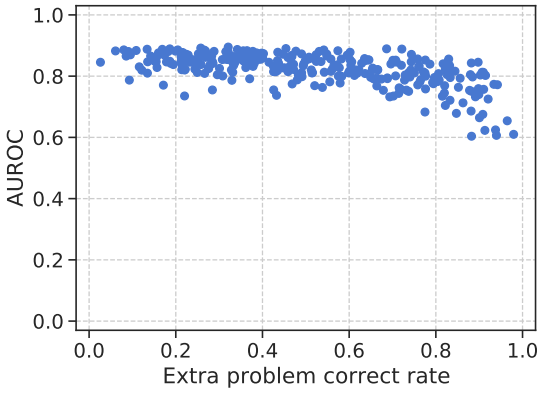


Figure 1: Classifier measure values by assessment question number.

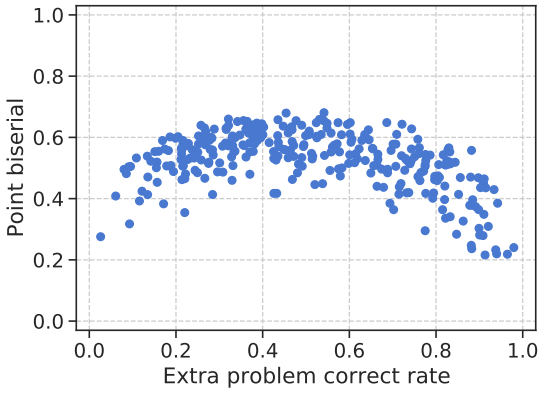
items. With its fixed set of 314 items from which the extra problems are chosen uniformly at random, the College Placement dataset has on average nearly 9000 data points per item.[5] For each item, the value of the measure (AUROC, point biserial correlation, or accuracy) may be compared to the correct rate for the item. Figures 2a–2c show the results for the AUROC, point biserial correlation, and accuracy, respectively. For the AUROC and point biserial correlation, the student's response to the extra problem is being compared directly to the probability returned by the ALEKS system, while for the accuracy, this probability is rounded to the nearest integer (either zero or one).

We begin by looking at the AUROC measure in Figure 2a. The items have a wide range of correct rates, so the AUROC is recommended since, as mentioned previously, it is not sensitive to class imbalances. This is supported by the fact that the trend of the AUROC values is mostly flat and does not have a significant dependence on the correct rate. The values seem to degrade only for the items with the highest correct rates; since, again, the AUROC is not sensitive to class imbalances, this appears to be strong evidence that the performance of the assessment degrades when making predictions on the easiest items.

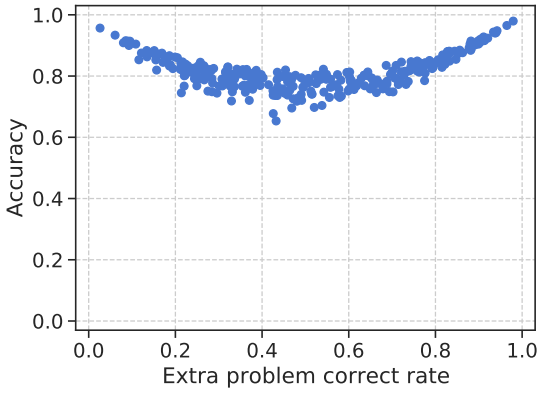
[5]This is in sharp contrast to ALEKS courses such as Sixth-Grade Math and College Algebra, for which the set of items appearing as extra problems is not fixed, as instructors are free to choose the items that appear in these courses. This results in a large proportion of items appearing rarely, or not at all, as extra problems, causing complications for an item-level analysis.



(a) AUROC



(b) Point biserial correlation



(c) Accuracy

Figure 2: AUROC, point biserial correlation, and accuracy versus extra problem correct rate.

Next, we turn to the point biserial correlation in Figure 2b. In contrast to the AUROC measure, the point biserial correlation has been shown to depend on the base rate of the dichotomous variable (Cohen, 1983; McGrath and Meyer, 2006). In our case, the base rate of the dichotomous variable corresponds to the extra problem correct rate. This dependence appears in the form of a pronounced curve in Figure 2b, where the point biserial correlations for the easiest and hardest items are lowest; the highest point biserial correlations then appear for the items in the middle range of difficulty. One final observation is that, while the point biserial correlations are low for the hardest items, the values for the easiest items appear to be even lower; this is consistent with the behavior of the AUROC measure, for which the lowest values also appeared for the easiest items.

Lastly, the plot for accuracy is displayed in Figure 2c. As the issues with accuracy values computed on imbalanced data are well known (Provost and Fawcett, 1997; Provost et al., 1998), unsurprisingly the accuracy plot also shows a pronounced curve, albeit in the opposite direction of the curve for the point biserial correlation. That is, the accuracy values are highest when the class imbalance is highest (i.e., for the easiest and hardest items) and lowest when the data are more balanced (i.e., for the middle difficulty items).

Based on the above analysis, there appears to be evidence that the performance of the ALEKS assessment degrades for the easiest items, that is, the items with the highest extra problem correct rates. A possible explanation for this is the following. In ALEKS, careless errors are much more common than lucky guesses; this is due to the fact that the majority of ALEKS items require an open-ended response, while relatively few require only multiple choice or true/false responses. Because of this, very difficult items have relatively little noise, since most students do not know how to solve them and are likely to give a wrong answer (again, since lucky guesses are rare). On the other hand, many students know the easy items but may make careless errors when solving them. When one considers this, along with the fact that discriminating students who know from those who do not know these easy items is difficult (simply because most students do in fact know these items), it is perhaps not surprising that the predictive performance of the assessment degrades for the easiest items.

2.3. Evaluating the assessment using knowledge state layers

We now look at how the extra problem rate is affected by the position of the item relative to the knowledge state. To do this, we employ the concept of the *layers* of a knowledge state. Following Doble et al. (2019), we define the layers of a knowledge state via the *surmise relation*.

Definition 2.1. Let $(Q, \mathcal{K})$ be a knowledge structure. For any item $q \in Q$, let $\mathcal{K}_q$ denote the family $\{K \in \mathcal{K} \mid q \in K\}$. The *surmise relation* $\precsim$ is a relation on Q defined by

$$q \precsim r \iff \mathcal{K}_q \supseteq \mathcal{K}_r. \tag{2.1}$$

We can now define the layers of a knowledge state.

Definition 2.2. Let $(Q, \mathcal{K})$ be a learning space. As such, it is known that the surmise relation is a partial order (that is, reflexive, antisymmetric, and transitive). We define the *outer layer* S^{ol} of a subset S of Q as

$$S^{ol} = \{q \notin S \mid \forall r \notin S, r \precsim q \Rightarrow r = q\}.$$

The outer layer of S is thus the set of the minimal items with respect to the restriction of $\precsim$ to $Q \setminus S$. Similarly, we define the *inner layer* S^{il} of a subset S of Q as

$$S^{il} = \{q \in S \mid \forall r \in S, q \precsim r \Rightarrow r = q\}.$$

The inner layer of S is thus the set of the maximal items with respect to the restriction of $\precsim$ to S. We define recursively the n^{th} *outer layer* K^{ol_n} of a state K as

$$\begin{aligned} K^{ol_1} &= K^{ol}, \\ K^{ol_n} &= (K \cup \bigcup_{i=1}^{n-1} K^{ol_i})^{ol} \text{ if } n \geq 2. \end{aligned}$$

Similarly, we define the n^{th} *inner layer* K^{il_n} of K as

$$\begin{aligned} K^{il_1} &= K^{il}, \\ K^{il_n} &= (K \setminus \bigcup_{i=1}^{n-1} K^{il_i})^{il} \text{ if } n \geq 2. \end{aligned}$$

It follows immediately from the definition of outer layer that, for any state K and any item $q \notin K$, there is a natural number n such that $q \in K^{ol_n}$. Moreover, the outer layers of K are order-preserving with respect to the surmise relation: for all $q, r \notin K$, with $q \in K^{ol_i}$ and $r \in K^{ol_j}$,

$$q \precsim r \;\Rightarrow\; i \leq j.$$

Similarly, the inner layers of K are order-reversing with respect to the surmise relation: for all $q, r \in K$, with $q \in K^{il_i}$ and $r \in K^{il_j}$,

$$q \precsim r \;\Rightarrow\; i \geq j.$$

The notion of layer exposed here has similarity with but is distinct from the notion of n^{th}-fringe introduced in Hockemeyer (1997). Additionally, in Appendix A of Cardinal et al. (2013) the "levels" of a partially ordered set are discussed, and this concept is equivalent to the outer layers of the empty set specifically when the knowledge space is defined by a partial order. The exact relation between layers and fringes of a knowledge state is examined in Doble et al. (2019). The layers can be thought of as a partition of the items in Q based on their difficulty and complexity in relation to a knowledge state K. Empirically, this can be seen in Figure 3a, which shows the correct rates for the College Placement extra problems as a function of the knowledge state layer. For this analysis, we exclude any extra problems that were classified in the uncertain

category at the end of the assessment; this gives a clearer picture of the different behavior of the items depending on whether or not they are in the student's knowledge state.[6] The correct rate essentially starts at zero for the outermost layers, and then begins to increase around outer layer 8, eventually arriving at a value of roughly 0.27 for outer layer 1. There is then a large jump from outer layer 1 to inner layer 1, with inner layer 1 having a correct rate of roughly 0.7. The correct rate then continues to increase throughout the rest of the inner layers.

Figure 3b shows the corresponding rates for the "I don't know" responses (that is, the proportion of all responses to the extra problem for which the student does not attempt to answer but chooses "I don't know"). Note that these rates behave somewhat differently from the correct rates. To start, the "I don't know" rates for the outer layers show larger differences compared to the correct rates, with the overall change being slightly more than 0.4; by comparison, the overall change in the correct rate for the outer layers is less than 0.3. Then, moving to the inner layers, we see markedly different behavior, as the "I don't know" rates are almost constant, showing very little change from the inner fringe onward. This is an interesting contrast, as it signals that students appear to recognize when they do in fact know how to solve an item. On the other hand, understanding the behavior in the outer layers is complicated by the fact that College Placement is a placement exam, and in such a context students may feel pressure to attempt answers, even if they do not know the content in the item.

In regard to this last concern, Figures 3c and 3d show the corresponding plots for the College Algebra course. While the correct rates in Figure 3c are roughly similar to the correct rates in Figure 3a, the "I don't know" rates in Figure 3d diverge from the corresponding rates in Figure 3b. In particular, the College Algebra "I don't know" rates reach a maximum of about 0.8, much higher than for College Placement. This would seem to support the hypothesis that students taking the College Placement assessment are motivated to attempt answers, even if items are inaccessible to them, as a higher score may mean placement into a more advanced course. On the other hand, College Algebra students are already enrolled in the course at the time of the assessment, and the purpose of the assessment is to ascertain their knowledge of the course (as opposed to placing them in a different one). The typical College Algebra student will likely have different motivations from the typical College Placement student when taking the initial assessment, and the "I don't know" rates seem to reflect this.

These figures can also be instructive when considering the validity of the item classifications made by the assessment. Specifically, the drop in the "I don't know" rate from the outer layers to the inner layers

[6] In practice, the majority of the uncertain items are in the outer fringe (outer layer 1) of the knowledge state returned by the assessment. As discussed previously, a large portion of these uncertain items are actually known by the student, with the assessment simply lacking enough information to make this classification. As such, including the uncertain items in the outer layers inflates the correct rate, and removing them gives a better estimate of the behavior of the "true" outer layers.

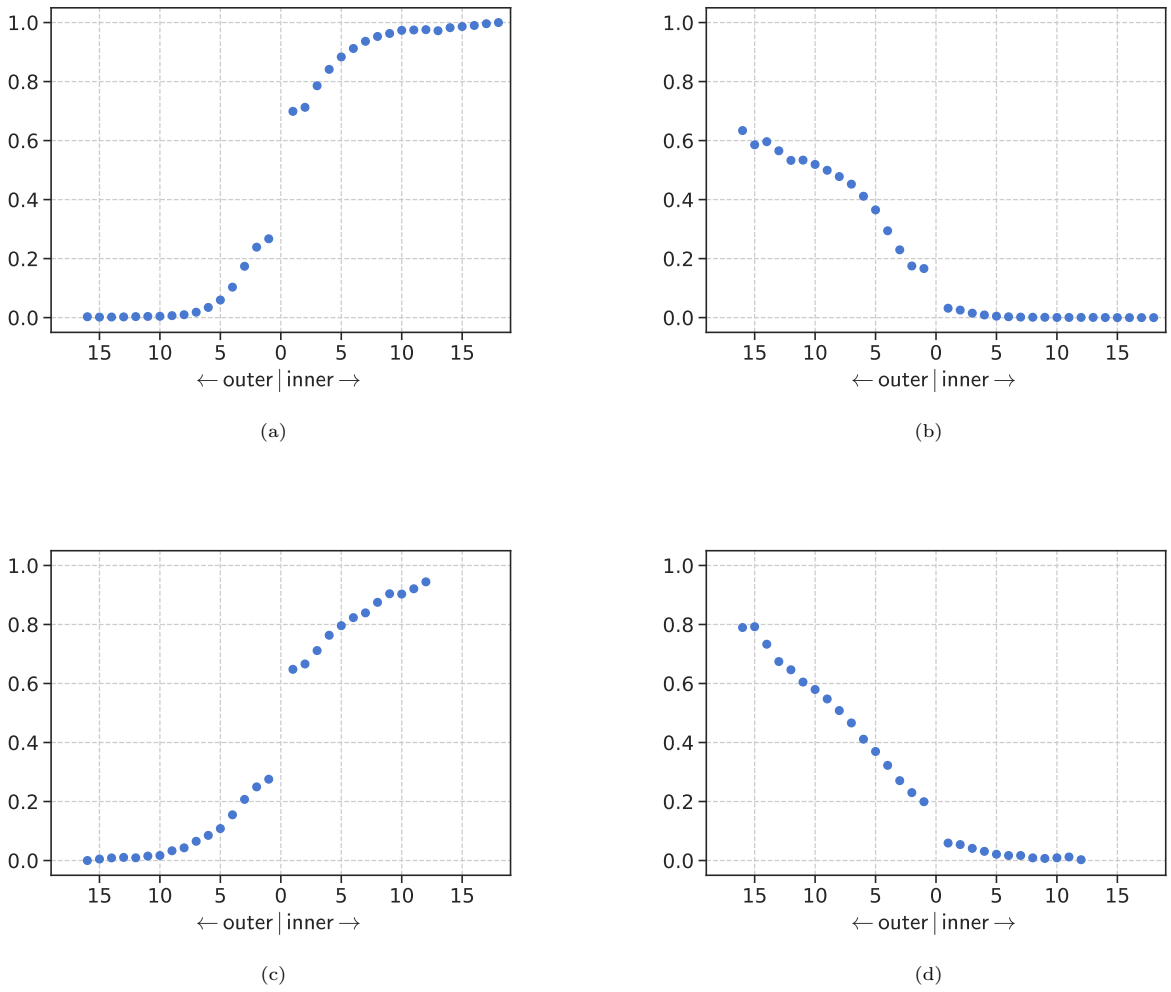


Figure 3: Extra problem correct rate (left) and "I don't know" rate (right) by fringe layer for College Placement (top) and College Algebra (bottom).

provides evidence that the in-state and out-of-state classifications made by the system are justified. That is, since students are not aware of these classifications during the assessment, the observed difference in their behavior when an item is out-of-state (i.e., the outer layers) compared to when it is in-state (i.e., the inner layers) provides empirical validity for these classifications. Additionally, the overall lack of "I don't know" responses for the inner layers seems to indicate that it is relatively rare for a student to truly have no idea how to solve an in-state item.

We next look at the extra problem correct rates for a few example items of varying difficulty. For context, Figure 4 displays a histogram of the correct rates for the College Placement items. Figure 5a then shows the correct rates by layer for an item with an overall correct rate of 0.15; from the histogram, we can see that a correct rate of 0.15 puts the item on the more difficult side. The difficulty of the item can also be seen in Figure 5a, where the majority of the data are from the outer layers, with inner layers 1 and 2 being the only inner layers with any significant amount of data; in other words, the item is most often categorized as being out of the student's knowledge state, indicating that it is relatively rare for students to know the item.

Figure 5b shows the layers for an item with a correct rate of 0.47, putting the item somewhere in the middle range of difficulty; this relative level of difficulty can also be seen in the figure, as the data points are balanced between the inner and outer layers. In Figure 5c we then have an example of an easier item, whose correct rate is 0.94. Once again, the difficulty of the item is indicated by the layer information, as the figure contains data points only for the inner layers.

We conclude this subsection by pointing out the contrast between the correct rates by layer presented above and the correct rates predicted by the *basic local independence model (BLIM)* (Definition A.7). Under the BLIM, the extra problem correct rate for an item q should equal the lucky guess rate η_q for all outer layers, and it should equal 1 minus the careless error rate, $1 - \beta_q$, for all inner layers. Figures 5a–5c show plots of the extra problem correct rates by layer for three individual items. These plots do not have enough data points for all layers. Nevertheless, all three plots look different from the 1-step functions predicted by the BLIM. If the predictions of in-state and out-of-state made by the ALEKS system are accurate, then these plots may serve as evidence that the lucky guess and careless error parameters vary by the knowledge state of the student, and therefore are not independent of the student's knowledge state as assumed for the BLIM.

Recall Figure 3a, which is the plot of the extra problem correct rates with all items aggregated. It has a characteristic "S" shape. The "S" shape is also present in similar graphs presented in Doble et al. (2019), where items are aggregated according to difficulty level. Traces of the "S" shape are also found in Figures 5a–5c. These plots suggest that a way of modeling the correct rates using information from the knowledge state (via the layers) is by using generalized sigmoid curves, also known as Richards curves (Richards, 1959; Lei and Zhang, 2004). These curves would require 3 to 5 parameters per curve, and they would not cause an

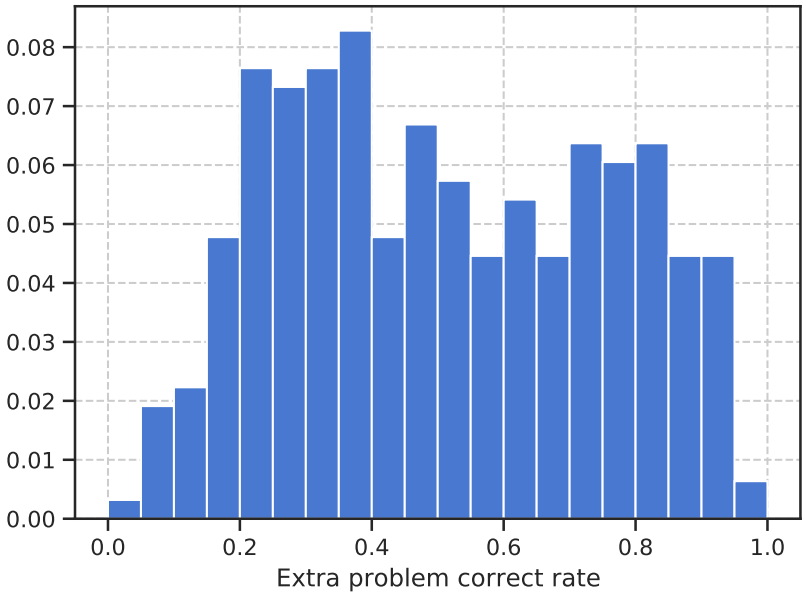


Figure 4: Relative frequency histogram of the extra problem correct rate for College Placement.

explosion in the number of parameters (a concern raised in Remark 11.1.3 in Falmagne and Doignon (2011)).

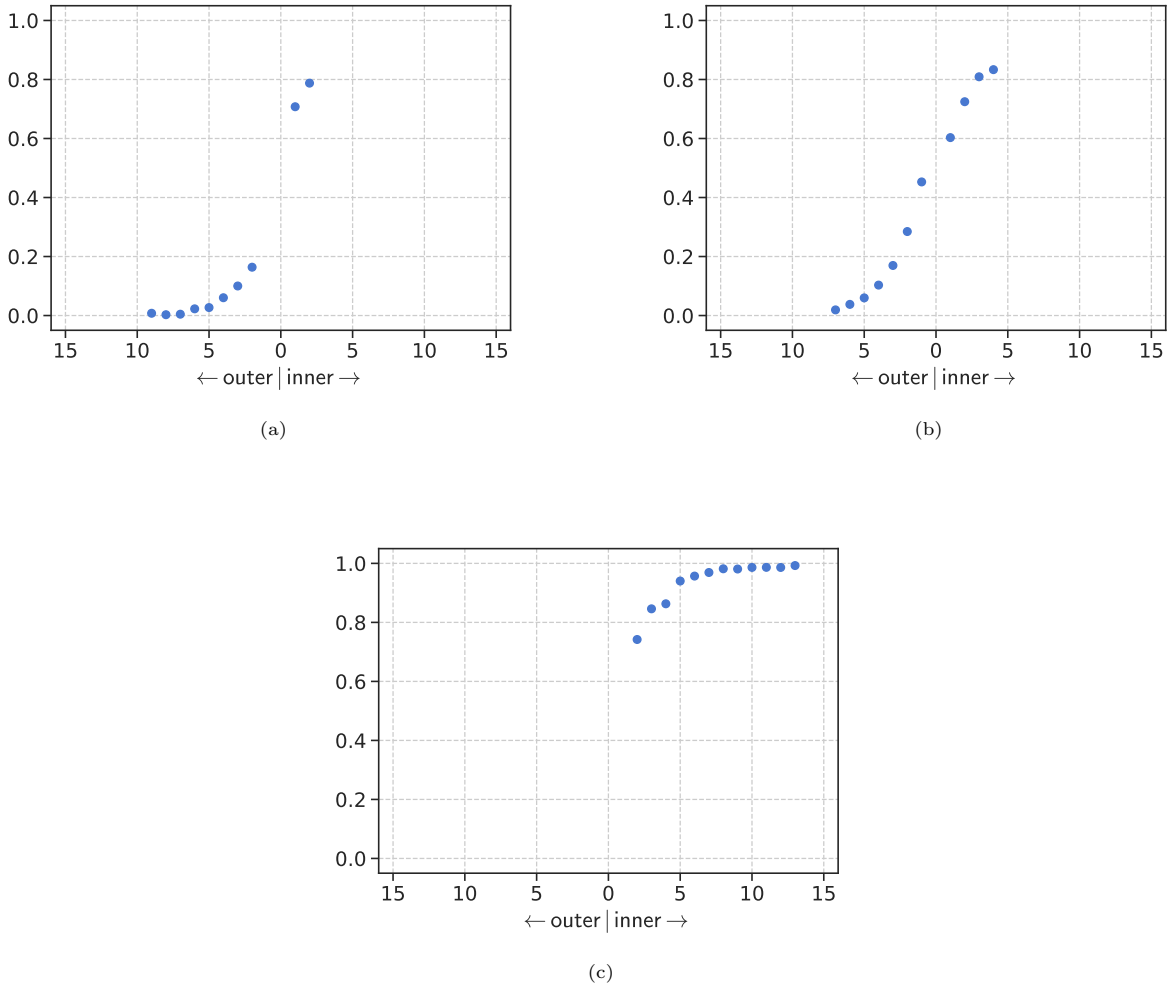
On the other hand, the in-state and out-of-state classifications by the ALEKS system almost certainly are made in the presence of inaccuracies in the knowledge structure and noise in student responses. Some of the evidence against the BLIM in these plots can be attributed to these factors. We did not attempt to mitigate these factors with an eye toward testing the BLIM in the current work, and more careful study is necessary.

3. Evaluating the ALEKS learning mode

To complement the analysis of the ALEKS assessment given in Section 2, we now examine the other principal aspect of the ALEKS system: the learning mode. As described in Section 1.2, once a student finishes the initial assessment, she is placed into the learning mode, during which she practices items in the outer fringe of her knowledge state. After a certain amount of time or items practiced, she is given a progress assessment to verify her learning, and after the assessment the learning mode resumes.

3.1. Learning an item from the student's outer fringe

We outlined in Section 1.2 the scoring system that measures a student's progress while she practices an item from her outer fringe. At the time the ALEKS system prompts the student to take a new assessment,



(a)

(b)

(c)

Figure 5: Extra problem correct rate by fringe layer for a high, medium, and low difficulty item, respectively.

each item that the student attempted to learn since the previous assessment can be classified in one of four categories: (i) the item was considered provisionally learned ('success') and was added to the student's knowledge state; (ii) the learning attempt was considered unsuccessful ('failure') and the student was invited to try another item; (iii) the student practiced the item, but ALEKS had not determined it yet a success or a failure ('incomplete'); and (iv) the student selected the item, but left it immediately after seeing the example explanation and never returned to practice it ('no activity'). This last category is counted as distinct from the 'incomplete' category.

Table 4 reports the proportion of each category for the three ALEKS learning courses under study: Sixth-Grade Math, Algebra I, and College Algebra. The data come from the usage of these three courses over the years $2016 - 2019$. In this table, all learning activity for an item between two consecutive assessments (or between the last assessment and the time all course activity ends) is represented by a single data point. If an item is re-attempted following an assessment, it will generate a second data point. At any time, the student may decide to suspend (or abandon) practicing an item by picking up another item available from the outer fringe. If the student revisits a previously suspended item before a new assessment takes place, the scoring for the item is resumed from when the student left the item. How often the practicing of an item happened without such a revisit is reported in parentheses for each category in Table 4. As the assessment may add or subtract items relative to the student's knowledge state before the assessment, the item scores are reset for all items following the assessment.

Table 4: Relative frequency of the classification of the learning attempts for each course. In parentheses is the proportion of the attempts in the category that did not involve a suspension with revisit.

Product	N	**Success**		**Failure**		**Incomplete**		**No activity**	
Sixth-Grade Math	38,928,733	0.809	(0.845)	0.041	(0.516)	0.103	(0.701)	0.047	(1.000)
Algebra I	21,769,811	0.841	(0.884)	0.030	(0.578)	0.085	(0.726)	0.044	(1.000)
College Algebra	15,749,089	0.943	(0.927)	0.015	(0.613)	0.032	(0.725)	0.010	(1.000)

In a general sense, the success rate can be seen as measuring the appropriateness of the outer fringes, and so, indirectly, of the underlying knowledge structures. We observe that the success rate increases with the grade level of the course (the age of the students), whereas each of the other three categories decrease with the grade level. On the other hand, assessment results in Tables 2 and 3 do not point to Sixth-Grade Math as having a less appropriate knowledge structure than College Algebra. It is reasonable to assume that the results in Table 4 reflect instead behavioral differences between the populations. A large fraction of College Algebra students are adults who enroll at two-year colleges with the goal of gaining higher-education credentials to improve their employment prospects. By contrast, Sixth-Grade Math and Algebra I are compulsory courses taken by minors of varying degrees of maturity and motivation. As the populations

age and become more self-selected, we observe an increased level of dedication and focus on learning. This interpretation is also consistent with the greater proportion of uninterrupted learning attempts from Sixth-Grade Math to College Algebra.

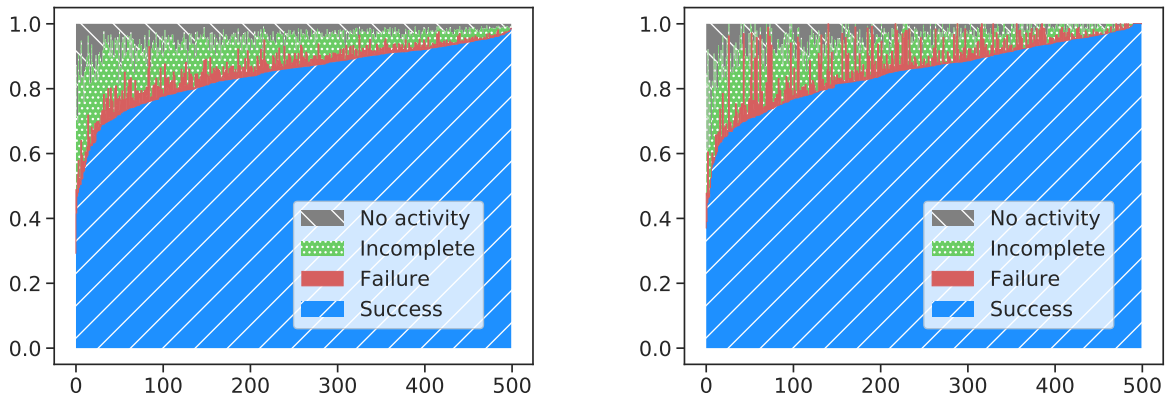
We now take a more detailed view of the Algebra I data from Table 4. Algebra I is chosen for this analysis as the 'middle' of the three courses. There are about 1100 items available to the instructor for inclusion in the Algebra I course. The most commonly used 500 items have more than 10,000 learning attempts each, and together account for more than 90% of the data. Figure 6a shows the relative frequency of each category (success, failure, incomplete, no activity) for these 500 most commonly used items. The items are ordered on the horizontal axis by increasing rate of success; 92.4% of the items have a success rate of 0.70 or higher, and the median success rate is 0.863. A small number of items, however, exhibit low success rates, with 14 of them falling below 0.60. There are several factors that may negatively impact the success rate of an item. One factor is a possible deficiency of the knowledge structure with respect to the item: the item may not belong in the outer fringe of the knowledge states from which the learning is attempted. Other factors may come from the item itself. For example, the item's explanation may need improvement, or there may be instances that are not representative of the item and are problematic for the student to answer successfully. Improving the low success rate of an item may require resolving these respective issues.

The data for Algebra I in Table 4 come from about 196,000 students, 77% of whom had at least 20 learning attempts over the length of the course. Figure 6b is similar to the previous figure, but this time the horizontal axis shows a random sample of 500 students with at least 20 learning attempts. (A sample was used, rather than the entire data set, for clarity in the figure.) The median student in the sample had a success rate of 0.865. The figure illustrates not only the differences in performance (that is, in success rate) among students, but also the differences in behavior. For instance, two students with similar success rates can differ tremendously in their rates of incomplete versus failure.

3.2. Retention in progress assessments and forgetting curves

We now examine the relationship between the retention of knowledge and the layers of a knowledge state. Specifically, we look at the famous Ebbinghaus forgetting curves (Averell and Heathcote, 2011; Ebbinghaus, 1885) in relation to different knowledge state layers. Previous work has shown that the retention of knowledge in ALEKS changes as a function of the time since the item was learned in the learning mode (Matayoshi et al., 2018, 2019). We define *retention* as the act of answering an item correctly when it appears as an extra problem at a point in time after the item is learned. We then say that the *retention rate* is the correct answer rate on these extra problems.

Starting from a dataset composed of the complete ALEKS learning and assessment profiles of 6,701,233 students drawn from the entire spectrum of ALEKS courses over the years $2016-2019$, we extract 8,352,006 extra problems that fit our definition of retention given in the previous paragraph (i.e., the items were



(a) (b)

Figure 6: Learning categorizations for the 500 most commonly used Algebra I items and for a random sample of 500 Algebra I students, respectively. In each figure, the relative frequencies of the four categories are stacked and add to 1.

learned in the ALEKS system before appearing as an extra problem in an assessment). For each of these extra problems we determine which inner layer the item appeared in when the item was encountered as the extra problem, and then we plot the retention rate for each layer as a function of the time since the item was learned.

The results are shown in Figure 7, where we can see that the curves are monotonically increasing as a function of the inner layer. Additionally, there are some interesting differences in the shapes of the forgetting curves. The first inner layer curve (Layer 1) has a steep drop within the first several days, and then flattens out at a correct rate of roughly 0.6. The second inner layer curve (Layer 2) has less of an initial decline, and then stays much higher, ending at about 0.67. In the subsequent inner layer curves this initial decline continues to lessen, with the final curve (representing inner layers greater than or equal to five) being mostly flat and ending with a rate of about 0.8. Figures 8 and 9 show the analogous curves for "I don't know" and incorrect responses, where we can again see that the response rates are very much dependent on the layer.

Thus, it is clear that the position of an item in the inner layers is strongly related to the retention and forgetting of that item. This makes sense as, all else being equal, the "deeper" an item gets in the knowledge state, the more potential there is for the state to contain related material that builds on the item. This related material would, presumably, have the effect of solidifying the knowledge associated with the item, resulting in higher retention. (See Matayoshi et al. (2020) for an analysis of how the learning of related material can act as a type of retrieval practice.)

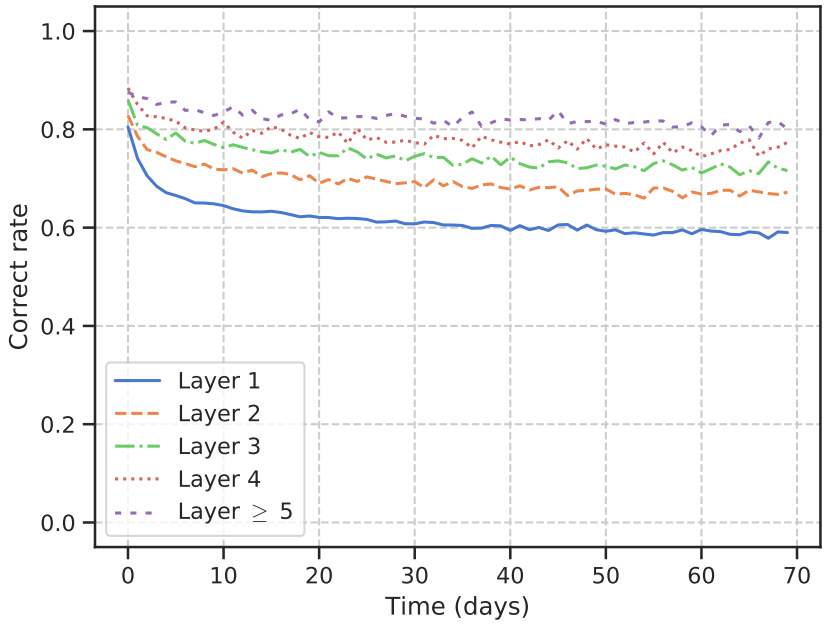


Figure 7: Forgetting curves (i.e., correct response curves) conditioned on the inner layer.

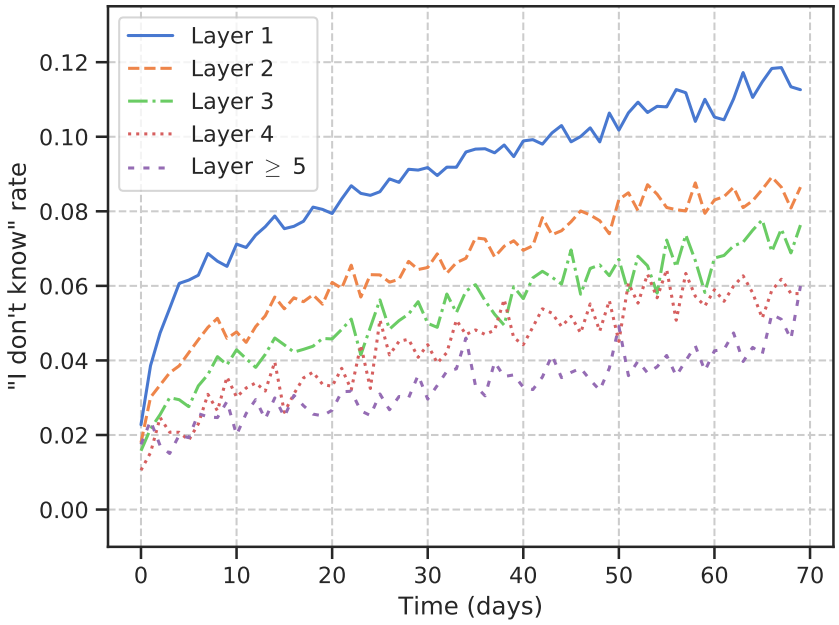


Figure 8: "I don't know" response curves conditioned on the inner layer.

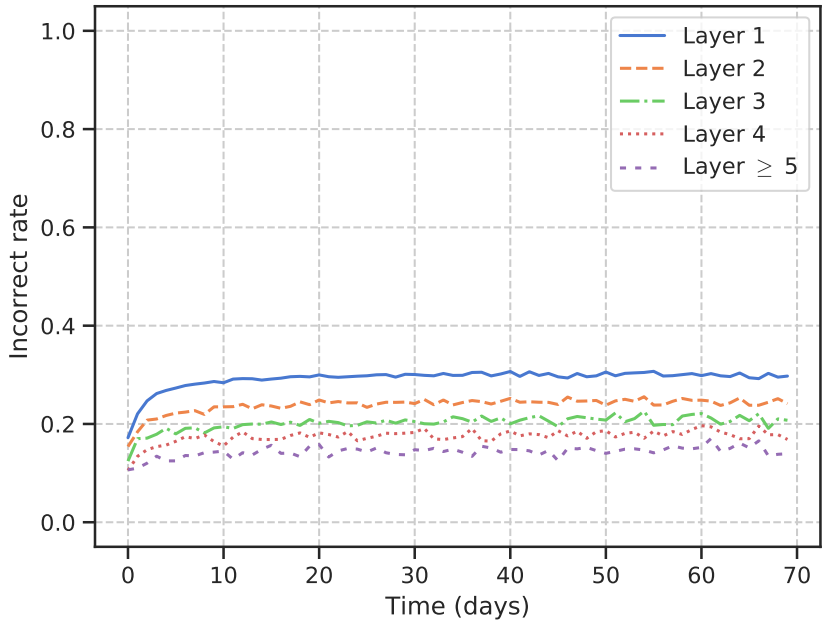


Figure 9: Incorrect response curves conditioned on the inner layer.

4. Discussion

As an implementation of knowledge space theory, the ALEKS system had to solve problems of scale not directly addressed by the theory. One issue is the scale of the domain of knowledge. Comprehensive content coverage of a typical course requires hundreds of items. Another issue is the scale of usage. ALEKS serves millions of students yearly. At any given time, tens of thousands of them can be simultaneously engaged in a knowledge assessment or in the learning of an item. Solving these issues remains an ongoing process. While this paper leaves out the technical and engineering aspects brought by these challenges, it outlines some of the solutions as they relate to the theory. There are also features of ALEKS that lie outside the direct scope of KST. The notion of outer fringe drives which items are accessible to the student for learning, but what constitutes the actual learning of an item is shaped by evolving heuristics. Similarly, the cycle of learning and assessment is not intrinsically a KST concept, but is supported by findings from cognitive psychology (Dunlosky et al., 2013) and the emerging field of learning science (Feldman, 2020). Critically, however, all design decisions were made as to remain consistent with the formal framework of KST.

The work presented here leverages the large amount of available data to analyze the extent to which practice conforms to theory. Section 2 evaluates the performance of the ALEKS initial assessment with respect to several measures, one of which relies on the concept of layers of a knowledge state. The data may

be viewed as questioning the assumptions of the basic local independence model, a mainstay of the research in KST. According to the BLIM, the conditional answer to an item given the knowledge state of the student is only determined by two parameters, the careless error rate and the lucky guess rate of the item, with no dependence on the knowledge state. The data instead suggest a high sensitivity to the actual knowledge state. But they also suggest that this dependency can be simply captured by the *relative* position of the item to the state as defined by its layer. Such a dependency can be modeled by a sigmoid function having as few as three parameters. Exploring this path would be the object of further study. In Section 3, we drew again on the notion of layers to analyze the retention of learned items. We found a monotonic dependency between rate of retention of an item and its inner layer relative to the student's knowledge state at the time retention is tested. Specifically, when items are positioned in the deeper inner layers of a knowledge state, there is an apparent reinforcement effect at work, in which the items are retained at an increasingly higher rate. Along the way, we also noticed effects in the data better explained by behavioral differences between student populations. Usage of the "I don't know" answer, introduced purely for the efficiency of the knowledge assessment, turned out to be indicative of the perceived stake students place in the assessment: a higher stake assessment such as College Placement is associated with a lower usage of the option. Similarly, the aggregate success rate for learning an item in one's outer fringe increases monotonically with the course grade.

In conclusion, the paper aimed to survey several of the challenges arising in a large-scale application of a tutoring system that must meet the pedagogical expectations of students and instructors. It is our hope that the findings presented here validate knowledge space theory as the model on which the system rests. Moreover we believe that the practical considerations discussed in the paper and the methods introduced to analyze the data will in turn stimulate new research in knowledge space theory.

Acknowledgments

We are grateful to Andrew Rast for useful discussions.

References

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

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

Baxter, R., Thibodeau, J., 2011. Does the use of intelligent learning and assessment software enhance the acquisition of financial accounting knowledge? Issues in Accounting Education 26, 647–656.

Boughorbel, S., Jarray, F., El-Anbari, M., 2017. Optimal classifier for imbalanced data using Matthews correlation coefficient metric. PloS one 12, e0177678.

Cardinal, J., Fiorini, S., Joret, G., Jungers, R.M., Munro, J.I., 2013. Sorting under partial information (without the ellipsoid algorithm). Combinatorica 33, 655–697.

Chicco, D., 2017. Ten quick tips for machine learning in computational biology. BioData mining 10, 35.

Cohen, J., 1983. The cost of dichotomization. Applied Psychological Measurement 7, 249–253.

Craig, S., Anderson, C., Bargagloitti, A., Graesser, A., Okwumabua, T., Sterbinsky, A., Hu, X., 2011. Learning with ALEKS: The impact of students attendance in a mathematics after-school program, in: Artificial Intelligence in Education, pp. 435–437.

Craig, S., Hu, X., Graesser, A., Bargagliotti, A., Sterbinsky, A., Cheney, K., Okwumabua, T., 2013. The impact of a technology-based mathematics after-school program using ALEKS on students' knowledge and behaviors. Computers and Education 68, 495–504.

Desmarais, M.C., Maluf, A., Liu, J., 1995. User-expertise modeling with empirically derived probabilistic implication networks. User modeling and user-adapted interaction 5, 283–315.

Desmarais, M.C., Pu, X., 2005. A bayesian inference adaptive testing framework and its comparison with item response theory. International Journal of Artificial Intelligence in Education 15, 291–323.

Doble, C., Matayoshi, J., Cosyn, E., Uzun, H., Karami, A., 2019. A data-based simulation study of reliability for an adaptive assessment based on knowledge space theory. International Journal of Artificial Intelligence in Education 29, 258–282.

Doignon, J.P., Falmagne, J.C., 1985. Spaces for the assessment of knowledge. International Journal of Man-Machine Studies 23, 175–196.

Dunlosky, J., Rawson, K., Marsh, E., Nathan, M., Willingham, D., 2013. Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. Psychological Science in the Public Interest 14, 14–58.

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

Falmagne, J.C., Albert, D., Doble, C., Eppstein, D., Hu, X. (Eds.), 2013. Knowledge Spaces: Applications in Education. Springer-Verlag, Heidelberg.

Falmagne, J.C., Doignon, J.P., 2011. Learning Spaces. Springer-Verlag, Heidelberg.

Fang, Y., Ren, Z., Hu, X., Graesser, A., 2019. A meta-analysis of the effectiveness of ALEKS on learning. Educational Psychology 39, 1278–1292.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognition Letters 27, 861–874.

Feldman, R. (Ed.), 2020. Learning Science: Theory, Research, and Practice. McGraw-Hill Education.

Hagerty, G., Smith, S., Goodwin, D., 2010. Redesigning college algebra: Combining educational theory and web-based learning to improve student attitudes and performance. PRIMUS 20, 418–437.

Hickey, D., Robinson, J., Fiorini, S., Fang, Y., 2020. Internet-based alternatives for equitable preparation, access, and success in gateway courses. The Internet and Higher Education 44.

Hockemeyer, C., 1997. Using the basis of a knowledge space for determining the fringe of a knowledge state. Journal of Mathematical Psychology 41, 275–279.

Hockemeyer, C., 2020. Bibliography on Knowledge Spaces. Available at `http://liinwww.ira.uka.de/bibliography/Ai/knowledge.spaces.html`.

Hu, X., Luellen, J., Okwumabua, T., Xu, Y., Mo, L., 2007. Observational findings from a web-based intelligent tutoring system: Elimination of racial disparities in an undergraduate behavioral statistics course, in: Conference of the American Educational Research Association.

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

Koppen, M., Doignon, J.P., 1990. How to build a knowledge space by querying an expert. Journal of Mathematical Psychology 34, 311–331.

Lei, Y., Zhang, S., 2004. Features and partial derivatives of Bertalanffy-Richards growth model in forestry. Nonlinear Analysis: Modelling and Control 1, 65–73.

Matayoshi, J., Granziol, U., Doble, C., Uzun, H., Cosyn, E., 2018. Forgetting curves and testing effect in an adaptive learning and assessment system, in: Proceedings of the 11^{th} International Conference on Educational Data Mining, pp. 607–612.

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

Matayoshi, J., Uzun, H., Cosyn, E., 2020. Studying retrieval practice in an intelligent tutoring system, in: Proceedings of the 7^{th} Conference on Learning at Scale, ACM. pp. 51–62.

Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure 405, 442–451.

McGrath, R.E., Meyer, G.J., 2006. When effect sizes disagree: the case of r and d. Psychological Methods 11, 386.

McGraw Hill, 2020. ALEKS course products. Available at `http://www.aleks.com/about_aleks/course_products`.

Mojarad, S., Essa, A., Mojarad, S., Baker, R.S., 2018. Studying adaptive learning efficacy using propensity score matching, in: Companion Proceedings of the 8^{th} International Conference on Learning Analytics and Knowledge (LAK 18), pp. 5–9.

Powers, D.M., 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies 2, 37–63.

Provost, F., Fawcett, T., 1997. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions, in: Proceedings of the 3^{rd} International Conference on Knowledge Discovery and Data Mining, pp. 43–48.

Provost, F., Fawcett, T., Kohavi, R., 1998. The case against accuracy estimation while comparing induction algorithms, in: ICML Conference.

Rawson, K.A., Dunlosky, J., 2011. Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? Journal of Experimental Psychology: General 140, 283.

Reddy, A., Harper, M., 2013. Mathematics placement at the University of Illinois. PRIMUS 23, 683–702.

Richards, F., 1959. A flexible growth function for empirical use. Journal of Experimental Botany 2, 290–300.

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

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

Roediger III, H.L., Karpicke, J.D., 2006b. Test-enhanced learning: Taking memory tests improves long-term retention. Psychological Science 17, 249–255.

Appendix

We recall here a few major definitions and theorems of knowledge space theory. They can all be found in Falmagne and Doignon (2011), sometimes in a slightly different form.

Definition A.1. A *learning space* is a knowledge structure $(Q, \mathcal{K})$ that satisfies the following two conditions.

(i) If $K \subset L$ are two knowledge states in $\mathcal{K}$, with $|L \setminus K| = n$, then there is a chain of states
$$K_0 = K \subset K_1 \subset \cdots \subset K_n = L$$
such that $K_i = K_{i-1} \cup \{q_i\}$ with $q_i \in Q \setminus K_{i-1}$ for $1 \leq i \leq n$.

(ii) If $K \subset L$ are two knowledge states in $\mathcal{K}$, with $q \in Q \setminus K$ and $K \cup \{q\} \in \mathcal{K}$ for some item q, then $L \cup \{q\} \in \mathcal{K}$.

Theorem A.2. *For any $K \in \mathcal{K}$, let $K^{\mathcal{I}}$ and $K^{\mathcal{O}}$ denote the inner and outer fringe of K, respectively. If $\mathcal{K}$ is a learning space, then*
$$\forall K, L \in \mathcal{K} : \ (K^{\mathcal{I}} = L^{\mathcal{I}} \text{ and } K^{\mathcal{O}} = L^{\mathcal{O}}) \iff K = L.$$

Definition A.3. Let $(\mathcal{K}, Q)$ be a knowledge structure and Q' a nonempty proper subset of Q. The family
$$\mathcal{K}_{|Q'} = \{W \subseteq Q' \mid W = K \cap Q' \text{ for some } K \in \mathcal{K}\}$$
is called the *projection* of $\mathcal{K}$ on Q'.

Theorem A.4. *If $(\mathcal{K}, Q)$ is a learning space and Q' a nonempty proper subset of Q, then $(\mathcal{K}_{|Q'}, Q')$ is a learning space.*

Definition A.5. For any $K, L \in \mathcal{K}$, let $d(K, L)$ denote the set-symmetric distance between K and L. For any integer $n \geq 0$, the family $\{L \in \mathcal{K} \mid d(K, L) \leq n\}$ is the *n-neighborhood* of K. For any sub-collection $\mathcal{F}$ of $\mathcal{K}$, the family $\{L \in \mathcal{K} \mid d(K, L) \leq n \text{ for some K in } \mathcal{F}\}$ is the *n-neighborhood* of $\mathcal{F}$.

Definition A.6. A *probabilistic knowledge structure* is a triple $(Q, \mathcal{K}, p)$ in which

(i) $(Q, \mathcal{K})$ is a knowledge structure;

(ii) the mapping $p : \mathcal{K} \to [0, 1] : K \mapsto p(K)$ is a probability distribution on $\mathcal{K}$; thus, for any $K \in \mathcal{K}$, we have $p(K) \geq 0$, and moreover, $\Sigma_{K \in \mathcal{K}} p(K) = 1$.

Definition A.7. The *basic local independence model (BLIM)* is a probabilistic knowledge structure $(Q, \mathcal{K}, p)$ that satisfies the following conditions.

(i) For each $q \in Q$, there are two constants $\beta_q, \eta_q \in [0, 1[$, respectively called (careless) error probability and guessing probability at q.

(ii) For any *response pattern* $R \subseteq Q$ and state $K \in \mathcal{K}$, the probability of observing R for a subject in state K is

$$\left(\prod_{q\in K\setminus R}\beta_q\right)\left(\prod_{q\in K\cap R}(1-\beta_q)\right)\left(\prod_{q\in R\setminus K}\eta_q\right)\left(\prod_{q\in Q\setminus(R\cup K)}(1-\eta_q)\right).$$