International Journal of Artificial Intelligence in Education (2019) 29: 258–282
https://doi.org/10.1007/s40593-019-00176-0

ARTICLE

A Data-Based Simulation Study of Reliability for an Adaptive Assessment Based on Knowledge Space Theory

Christopher Doble[1] · Jeffrey Matayoshi[1] · Eric Cosyn[1] · Hasan Uzun[1] · Arash Karami[1]

Published online: 19 March 2019

Abstract

A large-scale simulation study of the assessment effectiveness of a particular instantiation of knowledge space theory is described. In this study, data from more than 700,000 actual assessments in mathematics using the ALEKS (Assessment and LEarning in Knowledge Spaces) software were used to determine response probabilities for the same number of simulated assessments, for the purpose of examining reliability. Several measures of reliability were examined, borrowed from existing psychometric approaches, with an eye toward developing measures for evaluating reliability for adaptive assessments. The results are compared to analogous results for assessments having mathematics content overlapping that of the ALEKS assessment, and some consequences and future directions are discussed.

Keywords Knowledge space theory · Simulation · Reliability · Adaptive assessment

✉ Jeffrey Matayoshi
jeffrey.matayoshi@aleks.com

Christopher Doble
christopher.doble@aleks.com

Eric Cosyn
eric.cosyn@aleks.com

Hasan Uzun
hasan.uzun@aleks.com

Arash Karami
arash.karami@aleks.com

[1] McGraw-Hill Education/ALEKS Corporation, 15460 Laguna Canyon Road, Irvine, CA 92618, USA

Introduction

Knowledge space theory (KST) is an approach to assessment and learning based on a combinatoric and probabilistic model introduced around 1985 by Jean-Claude Falmagne and Jean-Paul Doignon (Doignon and Falmagne 1985). Since then, hundreds of peer-reviewed journal articles and several books have been published on the subject, most notably the two monographs Doignon and Falmagne (1999) and Falmagne and Doignon (2011). A bibliographic database is maintained by Cord Hockemeyer at the University of Graz in Austria: http://liinwww.ira.uka.de/bibliography/Ai/knowledge.spaces.html .

ALEKS, which stands for "**A**ssessment and **LE**arning in **K**nowledge **S**paces," is a web-based software system based on KST that is designed to assess and instruct students in mathematics. The backbone of the system is an adaptive assessment aimed at determining which specific mathematics topics a student already knows and which she is ready to learn. A specialization of the ALEKS system—ALEKS Placement, Preparation and Learning (ALEKS PPL)—has been developed to determine a student's mastery of high school-level mathematics, offer a recommendation for placement in a post-secondary mathematics course, and remediate the student as needed.

In this paper, we examine reliability for the ALEKS PPL assessment, which is an adaptive assessment. By *reliability* we will be referring to the test-retest repeatability of results for a student; highly reliable scores are generally free from measurement error and would likely be roughly repeated if the student took the assessment multiple times (assuming no learning or forgetting took place) (Thissen 2000). We detail a study of reliability that introduces a new theoretical idea to KST, namely, the 'layers of a knowledge state', as well as a novel simulation method based on this idea. The simulation is based on data derived from more than 700,000 actual ALEKS PPL assessments; students' actual results are compared to simulated ones, and reliability is measured in several ways. The measures of reliability are borrowed from the area of psychometrics with the intention of eventual comparison to other adaptive assessments, both within and outside of KST. To the best of our knowledge, there is not a standard approach within the artificial intelligence in education (AIED) or intelligent tutoring system (ITS) literature for measuring reliability for adaptive assessments. Our hope is that this study will contribute to an eventual implementation of standard measures in the AIED community for doing so.

The outline of the paper is as follows. We begin with a summary of KST and descriptions of some particulars of ALEKS PPL. This is followed by background information on measures of reliability, along with an introduction to the reliability approach used in this study. We then give the details of the simulation study and its results, and we conclude with a discussion of the consequences of the study and possible directions for future research. There is also an Appendix containing some theoretical results in KST that provide additional background for our simulations.

Basic Concepts: Knowledge Space Theory and ALEKS PPL

Items, Knowledge States, and the Knowledge Structure

At the core of KST is the concept of a 'problem type' or 'item.' ALEKS PPL has 314 items, each covering a discrete skill, concept, or granular topic from the high school mathematics curriculum. The items cover areas such as whole number arithmetic, proportions and percentages, integer arithmetic, linear and quadratic functions, polynomial and rational functions, exponentials and logarithms, and trigonometry, among others. Table 1 shows some items in ALEKS PPL and an example of each. The full list of items is available at the following link: https://web.archive.org/web/20190206222221/https://www.aleks.com/highered/ppl/ALEKS_Placement_Problem_Types.pdf .

An item is actually a collection of examples, called *instances*, each focused on the same, narrow topic, and designed to be equal in difficulty. Another instance of the item "Mixed arithmetic operations with fractions" might ask the student to evaluate $\frac{3}{4} - \frac{1}{6} \div \frac{2}{5}$ instead of $\frac{7}{8} + \frac{1}{4} \div \frac{6}{7}$. Each time the student is presented the item, whether in an instructional setting or in an assessment, a different instance is chosen.

The ALEKS PPL 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.

In addition to the concept of an item, two other vital concepts in KST are a 'knowledge state' and the 'knowledge structure.' A *knowledge state* is a particular set of items that some individual could be capable of solving correctly. The knowledge structure is a distinguished collection of knowledge states. More precisely, denoting the set of all items by Q, a pair $(Q, \mathcal{K})$ is a *knowledge structure* if $\mathcal{K}$ is a family of

Table 1 Some items in ALEKS PPL

Item	Example (instance) of the item
Writing a ratio as a percentage	Ann ran for president of the chess club, and she received 18 votes. There were 45 members in the club. What percentage of the club members voted for Ann?
Mixed arithmetic operations with fractions	Evaluate $\frac{7}{8} + \frac{1}{4} \div \frac{6}{7}$. Write your answer in simplest form.
Solving a compound linear inequality	Solve the compound inequality $3x + 1 > 19$ *or* $4x - 2 \leq 6$ and graph the solution on the number line.
Finding an amount in a word problem on growth or decay	A forest covers an area of $2200\,km^2$. Suppose that each year this area decreases by 5.75%. What will the area be after 14 years? Use the calculator provided and round your answer to the nearest square kilometer.
Sketching the graph of $y = a\,sin(bx)$ or $y = a\,cos(bx)$	Graph the function $y = 2\,cos\left(\frac{3}{4}x\right)$. To draw the graph, plot all points corresponding to x-intercepts, minima, and maxima within one cycle. Then click on the graph icon.

subsets of Q containing all the knowledge states that are feasible, that is, that could characterize some individual in the population. In other words, an individual whose knowledge state is K can, in principle (discounting slips[1] and lucky guesses), correctly solve all the items in K and would fail to correctly solve any item not in K. In particular, $\mathcal{K}$ 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).

In what follows, we often shorten the term "knowledge state" to simply "state."

Typically in KST, including in the case of ALEKS PPL, 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 in mathematics. For instance, some items are prerequisites required to master other items: It would be very unlikely, or even impossible, for a student to have mastered some particular items without having mastered other particular ones. Assuming item b could not be mastered without having mastered item a, all of the subsets of Q containing b but not a are absent from the knowledge structure; only those subsets containing b that also contain a are candidates to be knowledge states. Formally, the knowledge structure (Q, $\mathcal{K}$) is constructed so as to satisfy the two axioms below, making (Q, $\mathcal{K}$) a *learning space* (Cosyn and Uzun 2009; Falmagne and Doignon 2011).

AXIOM 1: 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 $1 \leq i \leq n$.
AXIOM 2: 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 $L \cup \{q\} \in \mathcal{K}$.

These axioms are designed to be sensible from a pedagogical standpoint. Axiom 1 requires that, if the state K of the student is included in some other state L containing n more items, then there is a sequence of items $q_1, \ldots, q_n$ that are learnable one at a time, leading the student from state K to state L. Axiom 2 requires that, if item q is learnable from state K, then it is also learnable from (or already learned in) any state L that includes K, so that knowing more does not make one less capable of learning something new.

These and similar considerations are employed in the construction of the knowledge structure. In particular, data from past assessments are used to identify and verify the prerequisite relationships between the items. Ensuring that the knowledge structure contains accurate and realistic knowledge states is an important process on which the effectiveness of the ALEKS PPL system rests; while this process will not be detailed further here, the reader is encouraged to see Koppen and Doignon (1990) and Falmagne and Doignon (2011) for more thorough treatments of building the knowledge structure, and Falmagne et al. (2013) and Reddy and Harper (2013) for investigations of the accuracy of the knowledge structure.

[1] A *slip* is an accidental error (see e.g. Desmarais and Baker 2012; Vie et al. 2017), often called a *careless error* in the KST literature (see e.g. Falmagne and Doignon 2011).

For ALEKS PPL, $|Q| = 314$ (there are 314 items in all), but the number $|\mathcal{K}|$ of states in the knowledge structure is considerably smaller than $2^{314} \approx 10^{94}$, the number of subsets of a set with 314 elements. More specifically, for ALEKS PPL, $|\mathcal{K}|$ is about 10^{23}.

The Fringes and Layers of a Knowledge State

Additional concepts from KST that are relevant for our discussion are the 'outer fringe' and the 'inner fringe' of a knowledge state. The *outer fringe* of a state K is the set of all the items q not in K such that $K \cup \{q\}$ is also a knowledge state. The interpretation is that the outer fringe of a student's state is made up of the items the student is 'ready to learn.' The *inner fringe* of a state K is the set of all the 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. A critical theorem (Theorem 4.1.7 in Falmagne and Doignon 2011) says that, in a learning space, the knowledge state of a student is completely determined by the inner fringe and the outer fringe of the state. Expressing a state in terms of its fringes can be very efficient, as the fringes often contain many fewer items than the state does. More importantly, the fringes give significant information for instruction, which is an important aspect of ALEKS PPL, but which is not described in much detail in this paper. See instead section 1.2 of Falmagne et al. (2013) in this regard.[2]

For the current study, we introduce the notions of 'outer layer' and 'inner layer' of a set of items. 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

$$q \precsim r \iff \mathcal{K}_q \supseteq \mathcal{K}_r. \tag{1}$$

It is easily shown that, if $(Q, \mathcal{K})$ is a learning space, then 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

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

[2] As pointed out by a reviewer, the reader will likely be familiar with Vygotsky's 'zone of proximal development' or 'ZPD' (Vygotsky 1978; Chaiklin 2003). Insofar as the outer fringe items comprise the content the student is ready to learn and the ALEKS PPL system provides instruction to help the student master this content, the outer fringe may be seen as an implementation of the notion of ZPD.

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

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

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

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

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 exact relationship between fringes and layers is developed in the Appendix.

The layers give a measure of the difficulty of an item for a particular student. For example, for a student in state K, an item in the 3rd outer layer of K will tend to be further from the student's grasp than an item in the 1st outer layer. Similarly, an item in the 3rd inner layer of K will generally be more solidly mastered by the student than one in the 1st inner layer. States containing many items (states that correspond to advanced students) have few outer layers and many inner layers, while the opposite is true for states containing few items. In ALEKS PPL, for most knowledge states, the sum of the outer and inner layers of the state is about 18 to 20. As described in a later section, the layers are an important component of our procedure for simulating student responses.

Uncovering the Knowledge State: an Assessment

The descriptions in this section borrow heavily from sections 1.3 and 8.7 of Falmagne et al. (2013). The reader is referred also to chapter 2 and section 8.8 of Falmagne et al. (2013) for more details.

The goal of an ALEKS PPL assessment is to uncover, by efficient questioning, the knowledge state of a particular student under examination. The assessment is adaptive, in that the choices of items to be presented depend on the student's previous answers. To begin, each of the knowledge states is assigned a certain a priori likelihood: Data from previously administered ALEKS PPL assessments are used to obtain an estimate of the distribution of knowledge states over the entire population,

and this (estimated) distribution serves as the starting point of the assessment. For the first question in the assessment, an item $\mathbf{p_1}$ is chosen both to be informative about the student and appropriate as a first question. The student's answer is checked by the system, and the likelihoods of all the states are modified according to the following updating rule: If the student gave a correct answer to $\mathbf{p_1}$, the likelihoods of all the states containing $\mathbf{p_1}$ are increased and, correspondingly, the likelihoods of all the states *not* containing $\mathbf{p_1}$ are decreased (so that the overall likelihood, summed over all the states, remains equal to 1). An incorrect answer given by the student has the opposite effect, as then the likelihoods of all the states *not* containing $\mathbf{p_1}$ are increased, and those of the remaining states are decreased. The exact formula for the operator modifying the likelihood distribution will not be recalled here; see Doignon and Falmagne (1999, Definition 10.10). It is proved there that the operator is commutative, in the sense that its cumulative effect in the course of a full assessment does not depend on the order in which the items have been presented to the student.[3] If the student does not know how to solve an item, she can choose to answer "I don't know" instead of guessing. This results in a substantial increase in the likelihoods of the states not containing the item, thereby decreasing the total number of questions required to uncover the student's state.

For the second question of the assessment, item $\mathbf{p_2}$ is chosen by a mechanism similar to that used for selecting $\mathbf{p_1}$, and the likelihood values are increased or decreased according to the student's answer via the same updating rule. Further questions in the assessment are dealt with similarly. In the course of the assessment, the likelihoods of some states gradually increase. The assessment procedure stops when either (i) 29 questions have been asked,[4] or (ii) there is no longer any useful question to be asked (all the items have either a very high or a very low probability of being answered correctly). At that moment, a few likely states remain and the system selects the most likely among them. Note that, because of the probabilistic nature of the assessment procedure, the final state may very well contain an item to which the student gave an incorrect response. Such a response is thus regarded as due to a slip.

Once the student's knowledge state is obtained through the ALEKS PPL assessment, the state is used to place the student into the appropriate mathematics course. To facilitate this placement, it is helpful to summarize the state with a single number that can easily be interpreted by students and instructors; the summary of choice is the student's *percentage score*, which is the percentage of the 314 items that are in the student's state (that is, the cardinality of the student's state divided by 314 and multiplied by 100%). Based on the percentage score, placement into a particular course can be recommended for the student (see Table 4 for an example of placement

[3]This commutativity property is consistent with the fact that, as shown by Mathieu Koppen (see Doignon and Falmagne 1999, Remark 10.11), this operator is Bayesian.

[4]The assessment actually includes up to 30 questions after the addition of a randomly chosen 'extra problem' used for testing purposes, the details of which are explained in a later section. This number of questions strikes a balance between the need to gather enough information about the student's knowledge state and the possibility of overwhelming the student with too many questions. Regarding the latter concern, see Matayoshi et al. (2018) for evidence of a "fatigue effect" experienced by students in ALEKS assessments.

recommendations). Though it is clear that a good deal of information about the student contained in the student's knowledge state is lost when the state is summarized with a single number, such a summary is useful to the post-secondary institutions using ALEKS PPL for placement. The reliability study in the current work is based on this percentage-score summary of the knowledge state. Though the instructional aspects of ALEKS PPL are not studied in this paper, we note for the reader that the full (unsummarized) knowledge state is retained by the ALEKS PPL system and is leveraged when the student works in the system's 'learning mode', where she has the opportunity to learn new material before re-taking the placement test.

Reliability: Background and the Current Study

As described in the next section, a large study was done to examine reliability for the ALEKS PPL assessment. In the absence of data from students taking the assessment multiple times (with no learning in between), simulations were done. These simulations went roughly as follows. First, data from a large number of ALEKS PPL assessments were collected for the purpose of obtaining response probabilities (of correct, incorrect, and "I don't know" responses) from the students' knowledge states and aspects of the ALEKS PPL knowledge structure. Then, using these probabilities, new assessments were simulated for each student. Comparisons of the simulation results to those of the original assessments give a picture of the assessment's reliability, that is, the degree to which the results are free from measurement error.

How this reliability should be measured seems an open question. Our search through the AIED and ITS literature did not yield any peer-reviewed studies of a similar nature to the current work. Thus, to the extent of our knowledge, there is not a standard approach within the AIED and ITS literature for measuring reliability for adaptive assessments. Because of this, we instead turn to the specific area of psychometrics for inspiration. In classical test theory, the measurement error (the standard error of measurement, or SEM) is estimated from the variance of the observed scores and some estimate of test-retest correlation (Green et al. 1984). Such an approach, however, is based on the assumption that the measurement error is the same for all students, regardless of ability. This dubious assumption is avoided in item response theory (IRT) (see, e.g., Green et al. 1984; Thissen 2000), for which the SEM is obtained via the information function, giving an error estimate for each value of the ability variable (Weiss 2011). However, how best to present measures of reliability for IRT and to compare them across studies remains an area of research (e.g., Dimitrov 2002; Weiss 2011; Woodruff et al. 2013).

The situation is especially murky when it comes to computerized adaptive testing (CAT), for which a test may differ in content (differ in the questions given) based on the test-taker's performance. Such adaptive content presents additional difficulties, as many common techniques for measuring reliability rely on having a fixed set of test questions; examples of these techniques include the split-half method (Zhu and Lowe 2018) and Cronbach's α (Cronbach 1951). What is widely accepted for CAT is that, as long as the test's stopping rule is not determined directly by the error

variance (Green et al. 1984), estimating measurement error based on the test score is preferred over having a single, overall error estimate (Green et al. 1984; Nicewander and Thomasson 1999; Thissen 2000; Weiss 2011; ACT 2012). This means having a *conditional* standard error of measurement, or CSEM, is preferred.

Turning to KST, in this study we have developed an original approach to measuring reliability for an assessment such as ALEKS PPL. The approach contains three key features. First, the percentage score is used. As mentioned, the outcome of an ALEKS PPL assessment is a knowledge state—the set of items the student is purported to have mastered—and this knowledge state is then used to compute a percentage score for placement purposes. While it is true that, in comparison to the actual knowledge state, the percentage score represents a loss of information, we believe that using the percentage score is appropriate in the study of reliability for ALEKS PPL, as it is this score that ultimately determines the course placement of the student, and importantly, from this score standard measures of reliability may be computed. Second, as mentioned above and described below, simulations are done to examine the degree to which the percentage scores from actual assesments match the percentage scores from subsequent (simulated) ones. Third, in the spirit of the background on reliability just presented, multiple measures of reliability are given (rather than a single measure). One is a correlation between actual and simulated scores, another is a conditional standard error of measurement giving an error estimate for each of ten score categories, and a third is a measure of the consistency of the actual and simulated scores in recommended course placement. These measures are detailed in the next section.

A Study of Reliability for the ALEKS PPL Assessment

Preparation for the Simulations

The data from 742,851 ALEKS PPL assessments—one assessment from each of 742,851 students—were used as the starting points of the simulations aimed at evaluating reliability for the ALEKS PPL assessment. 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. Each of these assessments was the first assessment the student took in ALEKS PPL; the data set comprises all such first-time placement assessments taken via ALEKS PPL between March 2012 and March 2017. (Students were able to take subsequent assessments in ALEKS PPL, but these subsequent assessments almost invariably came after the student learned new material through the system's 'learning mode' and were not used for this study.) Roughly 85% of the assessments were taken by students of four-year colleges or universities, and the remainder were taken by students at other institutions, such as two-year colleges, with more than 400 institutions represented in all. Further demographic information about the students, such as information regarding gender, age, or academic major, is not available. However, because these placement assessments typically were required for all incoming students at their institutions, it is safe to say

that the demographics of the students in the data set reflect those of students at a large sample of two- and four-year colleges in the United States.

As mentioned, the result of each ALEKS assessment is a knowledge state; it is assumed in this study that a student's knowledge state is the one obtained from the actual ALEKS PPL assessment (not a simulated one, and not obtained from any other information about the student). In what follows, we will call this the student's *assessed state*. Recall that a knowledge state is the set of items the student is purported to have mastered, so that, discounting lucky guesses and slips, the student would answer each item in her state correctly, and she would answer each item not in her state incorrectly. Of course, in practice, lucky guesses and slips may play a role, and they may vary based on the ability of the student and/or the difficulty of the question for the student, both of which should be modeled in the simulations. To this end, first each student was placed in one of ten categories based on the number of items in his assessed state. We will call these ten categories *knowledge categories*. The knowledge categories were based on 10 intervals of equal width, so that the first category contained percentage scores of 0% to 10%, the second category contained percentage scores of 10% to 20%, etc. Then, for each knowledge category, probabilities of the respective responses of correct, incorrect, and "I don't know" were estimated, conditioned on the layer of the item relative to the student's assessed state.

The following procedure was used to obtain these probability estimates. In each assessment, an *extra problem* was selected uniformly at random from all of the available items. The student was then presented this extra problem as a typical question during the assessment. While the student's response did not affect the result of the assessment, the data point from the response was grouped with data from other students in the same knowledge category. The knowledge category was then further divided based on the layer of the extra problem relative to the student's assessed state, and the statistics for each of these layers were compiled. For example, a student who had 100 items in his assessed state would have mastered $\frac{100}{314} \approx 32\%$ of the possible items, putting him in the 30%–40% knowledge category. Students in this category were given an extra problem in their 3rd inner layer 8411 times during their 107,721 assessments; they answered correctly $\frac{6714}{8411} \approx 79.8\%$ of the time, incorrectly $\frac{1528}{8411} \approx 18.2\%$ of the time, and "I don't know" $\frac{169}{8411} \approx 2.0\%$ of the time. (Note that these three fractions sum to 1, as must be the case.) The value .798 (the proportion of correct responses) is plotted in Fig. 1, and the value .020 (the proportion of "I don't know" responses) is plotted in Fig. 2. More generally, the figures show, for each of the ten knowledge categories, the estimated probabilities of a correct response (Fig. 1) and an "I don't know" response (Fig. 2) for each of the layers the students in the category encountered. The estimated probability of an incorrect response is not shown directly in the figures but can be computed easily by adding the probabilities of a correct response and an "I don't know" response, and then subtracting the sum from 1.

It is worth noting the similarity between the plots in Fig. 1 and the sigmoid item characteristic curves of IRT (see, for instance, Hambleton et al. 1991). The interpretation, however, is very different. In IRT, both students and items are located on a unidimensional latent trait θ, measuring the "ability" of the former and the "difficulty" of the latter. The probability that a student of ability θ_s gives a correct

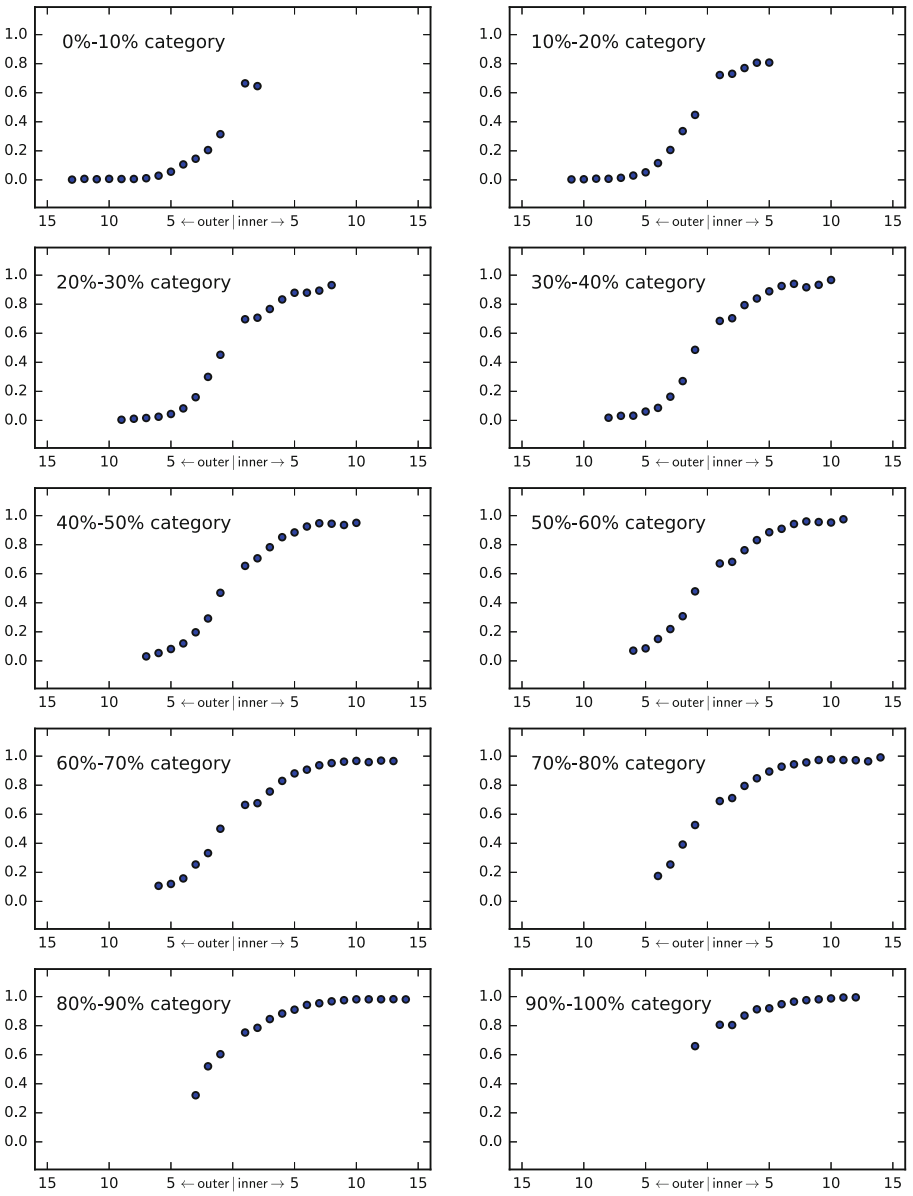

Fig. 1 For each knowledge category, proportion of correct responses to the extra problem as a function of the layer. The outer layers are in decreasing order, followed by the inner layers in increasing order. Only values with 500 or more data points are plotted

response for an item of difficulty θ_i is then a sigmoidal function of $(\theta_s - \theta_i)$. In the plots in Fig. 1, the probability that a student in knowledge state K_s gives a correct response for an item i is a function of the layer of i relative to K_s.

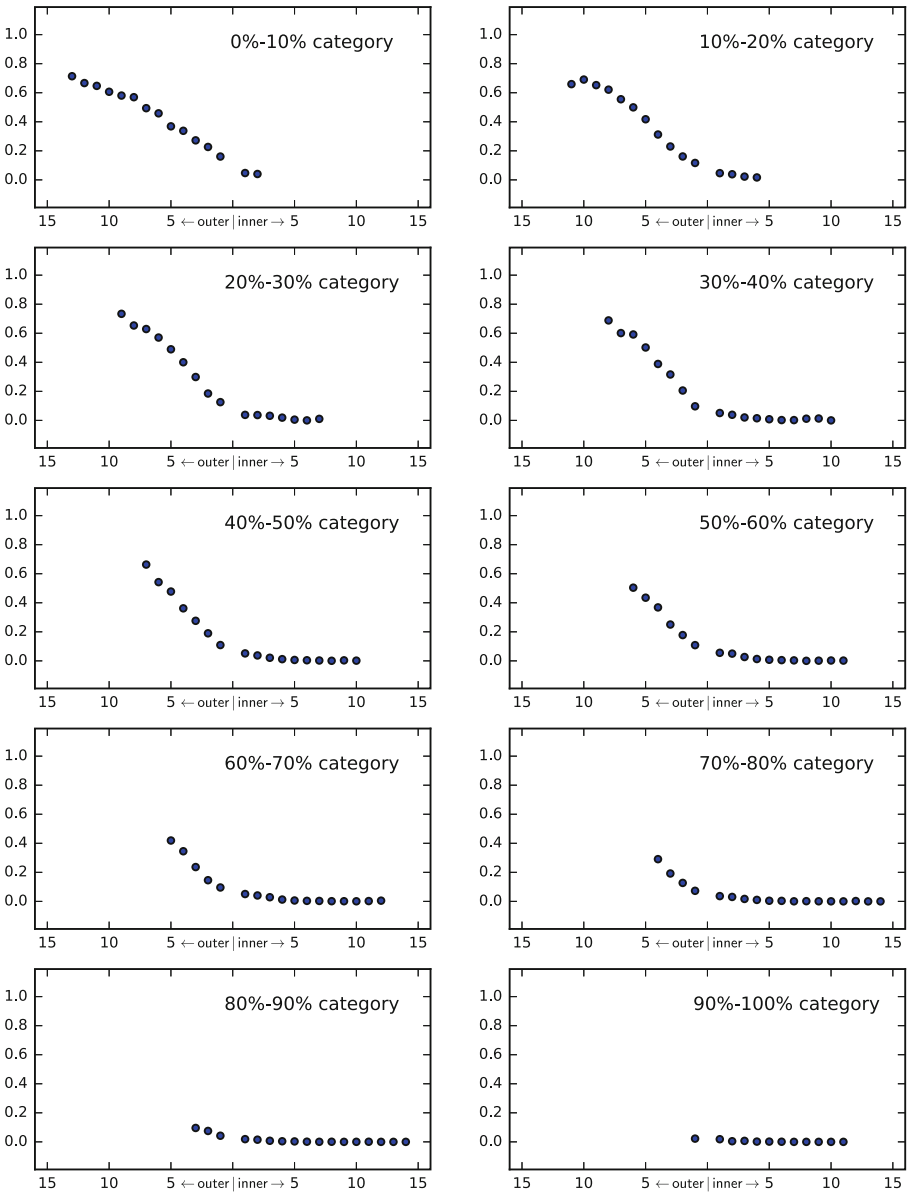

Fig. 2 For each knowledge category, proportion of "I don't know" responses to the extra problem as a function of the layer. Only values with 100 or more data points are plotted

The Simulations

Using the information in Figs. 1 and 2, a total of 742,851 simulated assessments—one for each of the students in the study—were run. For each question of a simulated

assessment, the student's simulated response (correct, incorrect, or "I don't know") was chosen based on the knowledge category of the student's assessed state, the layer relative to the assessed state of the item chosen for the question, and the probabilities in Figs. 1 and 2. For example, suppose a student's assessed state, K, fell into the 30%–40% knowledge category, and that in the simulated assessment for this student, an item was chosen that was in the 3rd inner layer of K. Then the simulated response would be correct with probability .798, incorrect with probability .182, and "I don't know" with probability .020; see Figs. 1 and 2 and the comments above. Each simulated assessment was run according to the procedure described in the section "Uncovering the knowledge state: An assessment". Since the questions for the simulated assessment were chosen according to this adaptive procedure (and were not chosen based on the questions for the student's actual assessment), it was very likely that the set of questions for the simulated assessment was different from the set of questions for the actual assessment. A knowledge state was obtained at the end of the simulated assessment, and this state will be called the student's *simulated state*.

Rarely, the situation occurred in a simulation for which there were not enough data points on which to base the student's response probabilities, given the student's knowledge category and the layer of the item relative to her assessed state. In such a case, response probabilities from the nearest layer for that knowledge category were used.[5]

An important consideration for the methodology described above is the possible bias introduced by using the same 742,851 assessment states for both estimating the response probabilities and running the actual simulations. We address this issue in the Discussion section.

Before turning to the results of the assessment, we note that, among the 742,851 students in the study, there were 729,398 assessed states held. Of the 15,496 students whose assessed state was shared by two or more students, nearly one-third had the empty state (the state with no items in it) as their assessed state. So, it was rare for non-empty states to be shared by two or more students to begin the simulations. As a consequence, the results of the simulations give an idea of the repeatability of the ALEKS PPL assessment for a variety of the possible outcomes of the assessment.

Results

Recall that the outcome of an ALEKS PPL assessment is the set of items the student is purported to have mastered, and that this set of items is converted to a percentage score for the purpose of placing the student in an appropriate mathematics course. (The percentage score is the percentage of the 314 possible items the student has in his state.) We begin the analysis of the results of the simulations by examining the degree to which the percentage scores for the assessed states (the *assessed scores*)

[5]Alternatively, generalized logistic functions (also known as Richards' curves; see Richards 1959; Lei and Zhang 2004) could have been fit to each of the plots in Figs. 1 and 2, and then used to estimate the response probabilities. This approach was examined and its effect was negligible. There was essentially no difference between the results from this approach and the results reported below.

Table 2 Summary statistics for the percentage scores

	Assessed scores	Simulated scores	
N	742,851	742,851	
Mean	50.5	50.6	Pearson correlation, *r*
Median	50.0	48.4	(assessed score versus simulated score):
S.D.	24.1	24.5	.958
Min.	0.0	0.0	
Max.	100.0	100.0	

match the percentage scores for the simulated states (the *simulated scores*). Table 2 shows some summary statistics for these percentage scores.

We see that the summary statistics are similar for the assessed scores as for the simulated scores. The Pearson correlation, $r = .958$, relating the assessed and simulated scores is high.[6]

Figures 3 and 4 give a more detailed picture. These figures show, for each of the ten knowledge categories, the distributions of the differences

$$(\textit{student's simulated score}) - (\textit{student's assessed score}).$$

The figures also show, for each distribution, the number N of students in the category and an estimate of the conditional standard error of measurement (CSEM). As is typically done, we computed the CSEM as the root-mean-square error

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N} e_i^2}$$

where the error e_i is the difference between the simulated and assessed scores of student i. As mentioned above, the CSEM is a recommended measure for reliability for an adaptive test (Green et al. 1984; Nicewander and Thomasson 1999; Thissen 2000; Weiss 2011; ACT 2012). It has the important characteristic of estimating error at multiple test scores (rather than giving a single error estimate). In our case, the CSEM estimates the variability (error) in the assessment's percentage score based on the knowledge category. For convenience, the CSEM values from Figs. 3 and 4 are shown in Table 3.

It is typical to consider CSEM values as determining intervals into which the student's score would be estimated to fall, namely, it would be estimated that the student's score would lie within one CSEM of the assessed score approximately 68% of

[6]Borrowing a measure from classical test theory, we also note that the standard error of measurement (Green et al. 1984) may be estimated, in units of percentage score points, as

$$S.D.\sqrt{1-r} \approx 24.1\sqrt{1-.958} \approx 4.94.$$

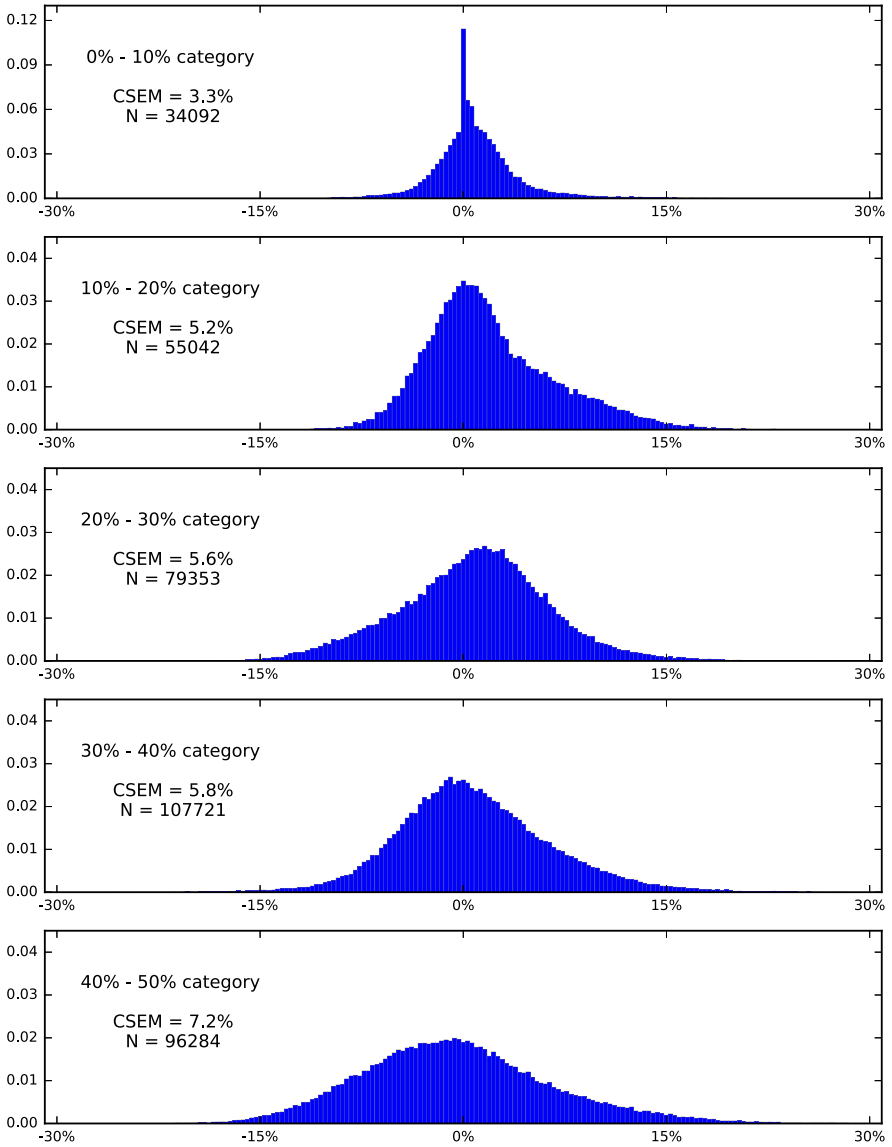

Fig. 3 For each of the first five knowledge categories, distribution of the differences between simulated and assessed scores. The root-mean-square error of the differences is reported as the conditional standard error of measurement (CSEM). For readability, the 0%–10% category has been given a different scale on its vertical axis than the other categories

the time, were the student to re-take the assessment many times (without a change in the student's knowledge state).

The CSEM is used to capture possible differences in variability across the span of possible scores, rather than to assume the same variability for each score. Examining

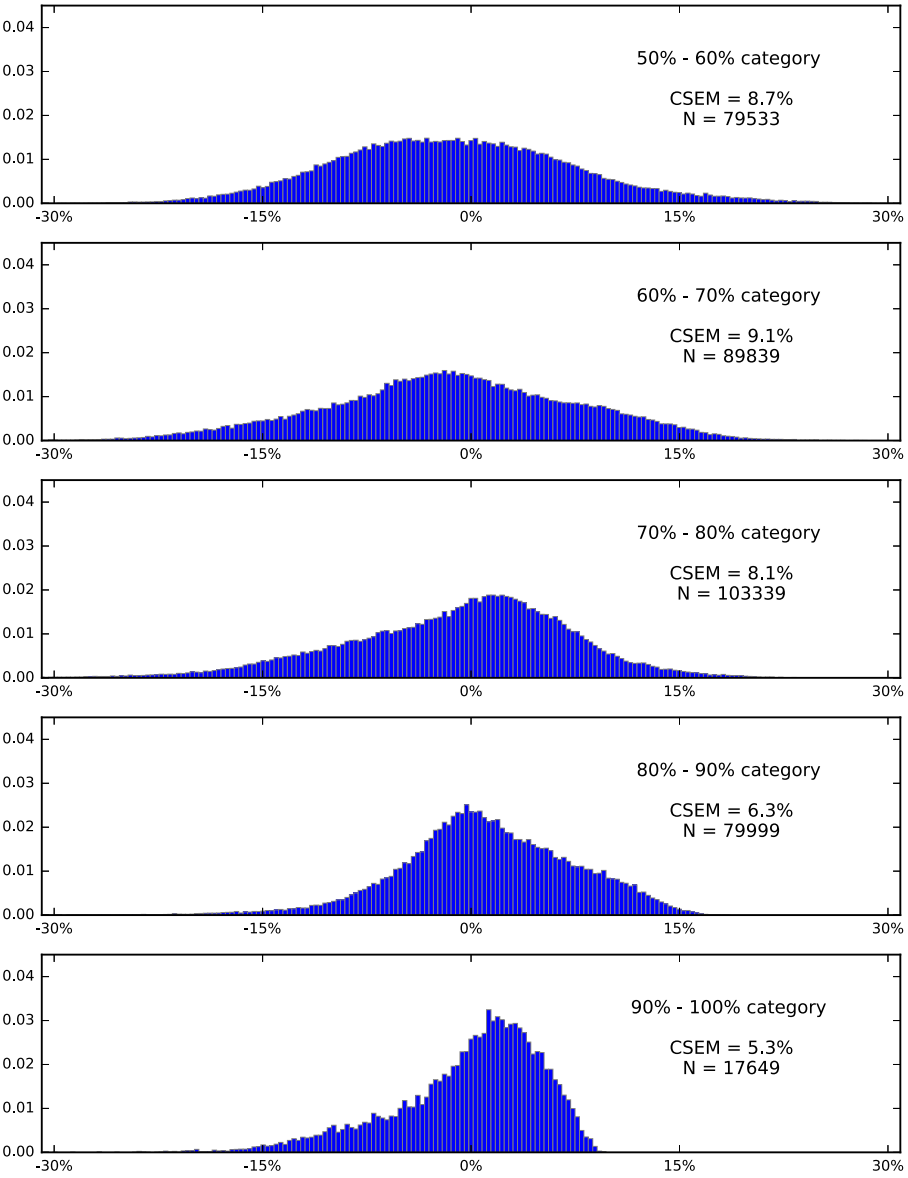

Fig. 4 For each of the last five knowledge categories, distribution of the differences between simulated and assessed scores

this in Table 3 and in Figs. 3 and 4, we see that the CSEM values range from about 3.3 percentage score points (for the 0%–10% knowledge category) to about 9.1 percentage score points (for the 60%–70% knowledge category). We also see that these CSEM values gradually but consistently increase from the first knowledge category (the 0%–10% category) to the seventh (the 60%–70% category), then decrease to the

Table 3 Conditional standard errors of measurement

Knowledge category	CSEM
0%–10%	3.3%
10%–20%	5.2%
20%–30%	5.6%
30%–40%	5.8%
40%–50%	7.2%
50%–60%	8.7%
60%–70%	9.1%
70%–80%	8.1%
80%–90%	6.3%
90%–100%	5.3%

The 0%–10% interval is left and right inclusive. The other intervals are left exclusive and right inclusive

final category (the 90%–100% category). So, the greatest variability is seen in the middle to high categories, that is, for percentage scores in the 40%–80% categories.

As previously mentioned, we are not aware of any reliability studies for adaptive assessments in the AIED/ITS literature; thus, directly comparing the CSEM scores of ALEKS PPL to any other system is difficult. Additionally, the documentation that does exist comes in the form of user manuals or unpublished technical reports that, presumably, have not been peer-reviewed; furthermore, these documents focus on IRT-based tests, of which the majority are not adaptive. However, with these caveats in mind, we can extract some useful information from such comparisons. For example, while the CSEM values for ALEKS PPL are greatest in the middle to high categories, this stands in contrast to some (non-adaptive) assessments of secondary mathematics, including the ACT Explore assessment for eighth- and ninth-graders (ACT 2014, page 41); the end-of-course assessments in Algebra II and Integrated Mathematics III for the Delaware Comprehensive Assessment System (DCAS) (American Institutes for Research 2014, pages 10 and 11); and the end-of-course assessments in Algebra I and Algebra II for the State of Texas Assessments of Academic Readiness (STAAR) (Texas Education Agency 2016, Appendix B: STAAR 2016 Raw Score to Scale Score (RSSS) Conversion Tables and Conditional Standard Error of Measurement (CSEM), pages 23–24 and 34–35). These latter assessments have their smallest CSEM values throughout the wide range of middle scores, and they have relatively large CSEM values as the scores become extreme (very high or very low). A study for another non-adaptive assessment, the ACT College Readiness Assessment (Woodruff et al. 2013, page 32), found small CSEM estimates for scores in the middle of the range of possible scores, large CSEM estimates for scores slightly more extreme, then small CSEM estimates for scores more extreme still.

While the above results are interesting, it should be emphasized that none of the aforementioned assessments are adaptive. The closest comparison to ALEKS PPL we have encountered in a reliability analysis comes from the ACT Computer-Adaptive Placement Assessment and Support System (COMPASS). COMPASS is an adaptive

Table 4 ALEKS PPL placement recommendations

Placement category	Cut score	Course placement
1	<14%	Basic Math/Prealgebra
2	≥14%	Beginning Algebra
3	≥30%	Intermediate Algebra
4	≥46%	College Algebra
5	≥61%	Precalculus/Business Calculus
6	≥76%	Calculus I

test that, similarly to ALEKS PPL, is used for placement purposes in a college or university setting and covers a wide range of content (ACT 2012). While not a perfect comparison, as the COMPASS test uses multiple choice questions and consists of five separate subtests, a relevant simulation study was perfomed in ACT (2012) to evaluate reliability. As described there, the variability is analogous to ALEKS PPL in that the largest CSEM values for the COMPASS test are found for scores near the middle of the range of possible scores; furthermore, each subtest of the COMPASS has CSEM values that appear comparable in size and variation to those in ALEKS PPL (see ACT 2012, Part 3, pages 7–10). However, it should be noted that additional differences between COMPASS and ALEKS PPL complicate the comparison. That is, while each COMPASS subtest asks fewer questions than a typical ALEKS PPL assessment, the scope of the content covered in each subtest is also much narrower.

For a final comment on Figs. 3 and 4, we note that the distributions are relatively more symmetric for the middle knowledge categories, especially the knowledge categories in the 20% to 70% range, and are relatively less symmetric for the extreme knowledge categories, especially the 10%–20% and 90%–100% categories. The asymmetries in extreme categories are likely due, in large part anyway, to floor and ceiling effects.

Rather than examining variability from the standpoint of CSEM values for percentage scores, we could instead look at variability in course placement. Consider Table 4, which shows the "cut scores" giving recommendations for placement in college courses based on ALEKS PPL percentage scores. (Each college or university may have its own cut scores. The default cut scores for ALEKS PPL are being used here for illustration.)

According to these cut scores, for example, a student with a percentage score of 34% would receive the recommendation to enroll in Intermediate Algebra.

Supposing each student were placed in a category based on her score and the Table 4 cut scores, the assessed and simulated placement categories of the student are determined by her assessed and simulated placement scores, respectively. Assuming the placement categories form an interval scale from 1 to 6, Fig. 5 gives, for each of the 6 categories, the distribution of the differences

(*student's simulated placement category*)−(*student's assessed placement category*).

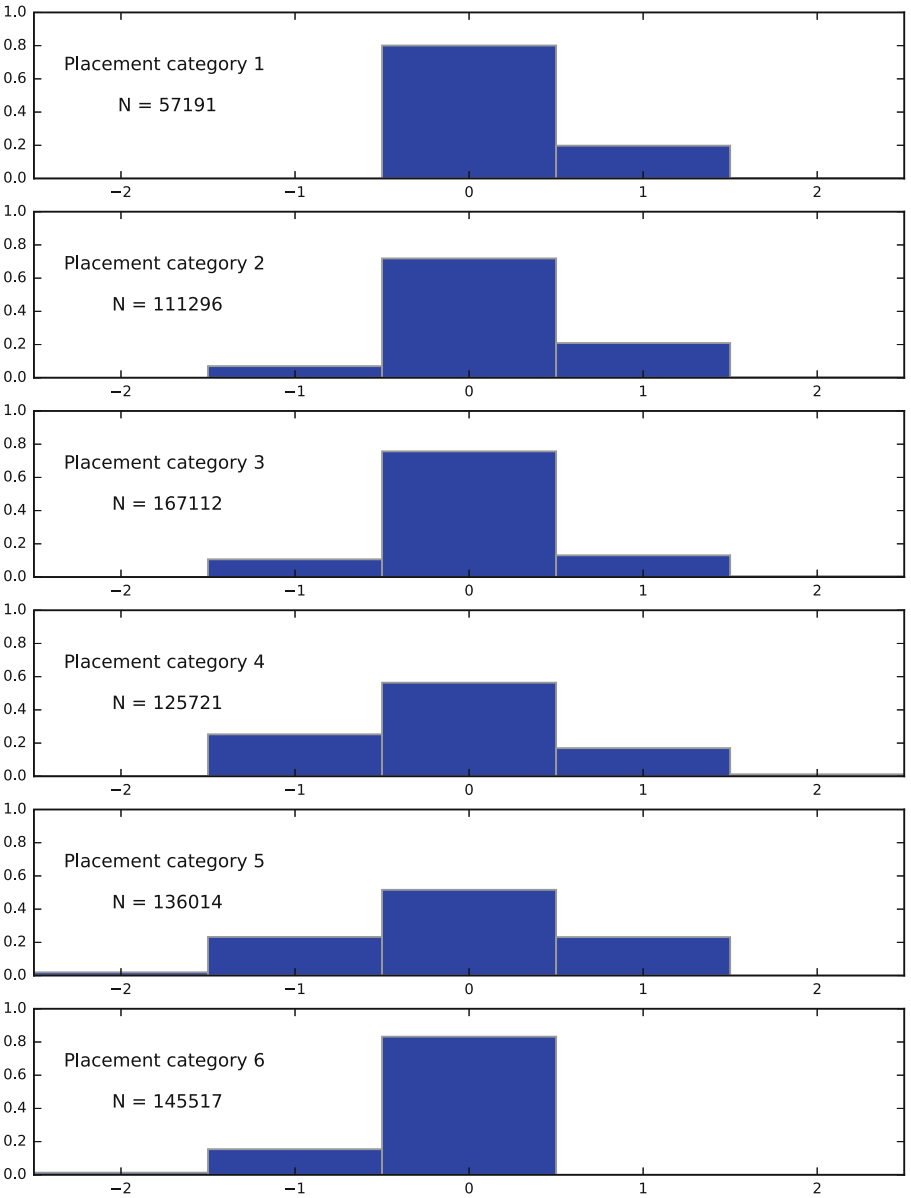

Fig. 5 For each placement category, distribution of the differences between simulated and assessed placement categories

Similarly, Table 5 reports the root-mean-square error for the distributions in Fig. 5. The most variability is seen for placement categories 4 (College Algebra) and 5 (Pre-calculus/Business Calculus). This is consistent with the results in Figs. 3 and 4, which show that the most variability comes with percentage scores in the 40% to 80% range.

Table 5 For each placement category, root-mean-square error (RMSE) of the differences between simulated category and assessed category

Placement category	RMSE
1	.452
2	.537
3	.511
4	.691
5	.737
6	.458

Note from Fig. 5 that floor and ceiling effects are again seen for the lowest category (Basic Math/Prealgebra, which is placement category 1) and the highest one (Calculus I, which is placement category 6).

Discussion

In the absence of currently recommended best practices in the AIED/ITS literature for measuring reliability for adaptive assessments, we have started from some concepts borrowed from the area of psychometrics and outlined a procedure for evaluating reliability for the ALEKS PPL assessment. In doing so, we have summarized the results of the assessment with a single numerical score, allowing for the computation of several measures of reliability. The knowledge state resulting from an ALEKS assessment contains very detailed information about the student's knowledge, and it is clear that much information is lost by summarizing this knowledge with a single number. However, while this is a significant drawback, the upside is that relatively standard measures of reliability can be used. Furthermore, though this matter is not investigated directly in the current study, such measures may be general enough to be used for many other adaptive assessments, including those that return a qualitative map of student knowledge (for examples, see Lynch and Howlin 2014; Desmarais and Baker 2012; Falmagne et al. 2013; Vie et al. 2017, and the present study) as well as those that employ numerical scores (such as assessments using IRT; see, for example, ACT 2012; Thissen 2000; Nicewander and Thomasson 1999; Vie et al. 2017).

Our analysis of the repeatability of the ALEKS PPL assessment rests on having a very large number of knowledge states represented—having many unique assessed states—among those being studied. These knowledge states were collapsed into categories by cardinality, and a separate simulation was done for each of the many states that occupied the category. A detailed analysis of results by category could then be carried out. The decision to have ten knowledge categories, as opposed to some other number, was based on the fact that having more categories meant a greater likelihood of not having enough data from actual assessments from which to estimate certain probabilities for simulated responses, and having fewer categories meant less of a chance to capture variety in competence among the population of students.

The reader may wonder about possible bias introduced by using the full data set of 742,851 actual assessments in computing parameters (that is, in computing the respective probabilities of correct, incorrect, and "I don't know" responses) and then using the states from these same assessments to run the simulations. Here is an argument as to why this procedure does not bias the results of our simulations. The parameters used in simulating student responses within a knowledge category were computed exclusively using the extra problems. Each extra problem was chosen uniformly at random from the complete set of 314 items, and as such was chosen independently of the questions and responses used to compute the student's assessed state and the corresponding knowledge category. Thus, the only dependence between the knowledge category of the student and the response to the extra problem was the underlying ability of the student, and we submit that this is the minimal amount of dependence required in any test-retest situation.

Note also that, in contrast to the situations in which cross-validation or some other model validation approach is usually recommended, we are not building a predictive model, nor are we running an optimization over any of our parameters. That being said, to test our argument empirically we tried a hold-out approach in which we used 100,000 of the assessments in the data set for validation and the remaining 642,851 for computing the parameters. The results were very similar to the ones reported above: The Pearson correlation coefficient was identical to that reported above, and among the ten CSEM values, five were identical to those reported in Table 3 above, four differed by 0.1% (for example, a value of 7.2% was obtained compared to a value of 7.1% in Table 3), and one differed by 0.2%. Taking this validation approach even further, we observe that an application of leave-one-out cross-validation would likely give results even more similar to those reported in Table 3.[7]

We now turn again to the results of the present simulation. The results suggest that the ALEKS PPL assessment has variability, as measured by the CSEM values for the percentage scores, that varies by knowledge category. The variability of the CSEM values by category is typical of assessments having a similar level of mathematics content as that of ALEKS PPL. As mentioned, several such assessments have their smallest error estimates—smallest CSEM values—for the middle range of possible scores, while others have their smallest error estimates for the extreme scores. The ALEKS PPL assessment seems to fall into the latter category, with the simulations indicating that ALEKS PPL has its greatest CSEM values for the middle to high percentage scores, especially those corresponding to placement in College Algebra and Precalculus/Business Calculus. It is not clear to us the reason for this, and it is a topic of current study. We can say that College Algebra and Precalculus have a good deal of overlapping content, and that discriminating between them can be difficult. On the other hand, the CSEM values for the highest percentage scores, corresponding

[7]For an illustration of this, there were 186 correct-response probabilities estimated from the data, with an average sample size for these estimates of 3941.0. So, in each step leave-one-out cross-validation would effectively remove exactly one data point (out of an average sample of 3941.0) from exactly one of the 186 correct-response probability estimates.

to placement in Calculus I, are lower, giving higher confidence that students placed in this category generally are being placed appropriately.

The outcome of an ALEKS PPL assessment contains a great deal of information about the student, including the particular items the student is purported to have mastered and those she is most capable of learning next. This information is put to use in the ALEKS PPL 'learning mode,' a central component of ALEKS PPL. In the learning mode, the student practices items in her outer fringe, which are the items she is deemed most capable of learning next. As the student makes progress in mastering these items, ALEKS PPL updates her knowledge state accordingly and presents her with new items from her updated outer fringe. This process allows the student to progress efficiently through the curriculum. If the student is seeking placement in a more advanced mathematics course, she may, depending on her school's regulations, re-take the ALEKS PPL assessment before the term starts for the chance to improve her standing. Leveraging the information in the knowledge state helps provide a smooth transition from course placement to coursework.

The cardinality of the knowledge state—the measure used in the current study—is an especially rough numerical summary of the knowledge state. Other numerical summaries may retain more information about the state. For example, a score that is a weighted sum, with the weights based on the degree to which items are useful in predicting success in future courses, may offer an improvement; see especially Reddy and Harper (2013) in this regard. Alternatively, note that, as described in the "Uncovering the knowledge state: An assessment" section above, the ALEKS PPL assessment engine determines, for each item, a probability the student has mastered the item. Rather than discarding these probabilities in favor of all-or-nothing decisions about items being in or out of the knowledge state, the system could instead report the probabilities, or some function of them, as the outcome of the assessment. Reliability may then be measured by the correlation of these probabilities from one assessment to the next. Another numerical summary that could be used for the purpose of examining reliability would be a set-theoretic distance between knowledge states, such as the 'symmetric difference' (see e.g. Falmagne and Doignon 2011). An examination of these and other measures, which depart from the score reporting currently used in ALEKS PPL, is a possible topic of future work.

Acknowledgements We are grateful to the associate editor and four anonymous reviewers for many helpful comments on a previous draft of this paper.

Appendix

For completeness, we give the following results relating a knowledge state's layers (a concept introduced in this work) to that state's fringes (an established concept in KST, e.g., Falmagne and Doignon 2011).

Let $(Q, \mathcal{K})$ be a learning space. We recall the definition of the *surmise function* σ. It is a mapping from Q to 2^{2^Q} defined, for all $q \in Q$, by

$$\sigma(q) = \{C \in \mathcal{K}_q \mid \forall S \in \mathcal{K}_q,\ S \subseteq C \ \Rightarrow\ S = C\}.$$

In other words, $\sigma(q)$ is the family of the minimal states, with respect to set inclusion, containing q. These minimal states are called the *clauses* of q. A fundamental result (Falmagne and Doignon 2011) is that

$$K \in \mathcal{K} \iff \forall q \in K, \exists C \in \sigma(q) \text{ such that } C \subseteq K. \tag{2}$$

The implication from left to right follows immediately from the definition of σ. The converse implies that any union of clauses is a state. Koppen (1998) moreover showed that, for all $q, r \in Q$,

$$r \in \cap\sigma(q) \iff r \precsim q, \tag{3}$$

where $\precsim$ is the surmise relation defined in (1). Let $K^{\mathcal{O}}$ and $K^{\mathcal{I}}$ denote the outer fringe and the inner fringe of state K, respectively. We show the following statements $(i) - (iv)$ for any state $K \in \mathcal{K}$.

(i) $K^{\mathcal{O}} \subseteq K^{ol}$. Let q be an item in $K^{\mathcal{O}}$. Theorem 1 in Hockemeyer (1997) states that, equivalently, there exists $C \in \sigma(q)$ such that $C \setminus K = \{q\}$. Suppose that there exists $x \notin K$ such that $x \precsim q$. Then $x \in \cap\sigma(q)$ and so $x \in C$. Thus $x = q$ and $q \in K^{ol}$.

(ii) $K^{\mathcal{I}} \subseteq K^{il}$. Let q be an item in $K^{\mathcal{I}}$ and suppose there exists $x \in K$ such that $q \precsim x$ and $x \neq q$. As x belongs to the state $K \setminus \{q\}$, there exists $C \in \sigma(x)$ such that $C \subseteq K \setminus \{q\}$. But $q \precsim x$ implies $q \in \cap\sigma(x)$ and so $q \in C$, a contradiction. So there is no such item x and q is a maximal item with respect to $\precsim$ restricted to K. Thus $q \in K^{il}$.

A case of interest is when $\sigma(q)$ is a singleton for all $q \in Q$. A learning space with this property is called an *ordinal* learning space. If C_q denotes the single clause of item q, (3) becomes

$$r \in C_q \iff r \precsim q.$$

(iii) In an ordinal learning space, $K^{\mathcal{O}} = K^{ol}$. We only need to show the inclusion from right to left. Let q be an item in K^{ol}. Let x be an item in $C_q \setminus K$. In particular, $x \precsim q$, and so $x = q$ since q is minimal with respect to $\precsim$ on $Q \setminus K$. We thus have $C_q \setminus K = \{q\}$ and $q \in K^{\mathcal{O}}$.

(iv) In an ordinal learning space, $K^{\mathcal{I}} = K^{il}$. We only need to show the inclusion from right to left. Let q be an item in K^{il}. Let x be an item in $K \setminus \{q\}$. From (2), we have the inclusion $C_x \subseteq K$. At the same time, $q \notin C_x$, otherwise q would not be maximal with respect to $\precsim$ on K. So $C_x \subseteq K \setminus \{q\}$. We can thus express $K \setminus \{q\}$ as a union of clauses. So $K \setminus \{q\}$ is a state and $q \in K^{\mathcal{I}}$.

Finally, let us observe that, since a learning space is closed under union, $K \cup K^{\mathcal{O}}$ is also a state, since it is the union of all the states $K \cup \{q\}$, where $q \in K^{\mathcal{O}}$. But we do not have in general that $K \cup K^{ol}$ is a state, unless the learning space is ordinal. As an ordinal learning space is closed under intersection (Birkhoff 1937), we have that $K \setminus K^{il}$ is a state in that case, since it is the intersection of all the states $K \setminus \{q\}$, where $q \in K^{\mathcal{I}}$.

References

ACT (2012). ACT Compass Internet Version Reference Manual. http://act-stage.adobecqms.net/content/dam/act/unsecured/documents/CompassReferenceManual.pdf. Accessed January 29, 2019.

ACT (2013–2014). ACT Explore Technical Manual. https://www.act.org/content/dam/act/unsecured/documents/Explore-TechManual.pdf. Accessed April 9, 2017.

American Institutes for Research (2013–2014). State of Delaware End-of-Course (EOC) Technical Report. Evidence of Reliability and Validity (Vol. 4).

Birkhoff, G. (1937). Rings of sets. *Duke Mathematical Journal*, *3*, 443–454.

Chaiklin, S. (2003). The zone of proximal development in Vygotsky's analysis of learning and instruction. In A. Kozulin, B. Gindis, V.S. Ageyev, S.M. Miller (Eds.) *Vygotsky's educational theory and practice in clutural context* (pp. 39–64). New York: Cambridge University Press.

Cosyn, E., & Thiéry, N. (2000). A practical procedure to build a knowledge structure. *Journal of Mathematical Psychology*, *44*, 383–407.

Cosyn, E., & Uzun, H.B. (2009). Note on two necessary and sufficient axioms for a well-graded knowledge space. *Journal of Mathematical Psychology*, *53*(1), 40–42.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Desmarais, C., & Baker, R. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, *22*(1–2), 9–38.

Dimitrov, D. (2002). Reliability: arguments for multiple perspectives amd potential problems for generalization across studies. *Educational and Psychological Measurement*, *5*, 783–801.

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

Doignon, J.-P., & Falmagne, J.-C. (1999). *Knowledge spaces*. Berlin: Springer.

Falmagne, J.-Cl., & Doignon, J.-P. (2011). *Learning spaces. Interdisciplinary applied mathematics*. Berlin: Springer.

Falmagne, J.-Cl., Albert, D., Doble, C.W., Eppstein, D., Hu, X. (Eds.) (2013). *Knowledge spaces: application in education*. Berlin: Springer.

Green, B.F., Bock, R.D., Humphreys, L.G., Linn, R.L., Reckase, M.D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, *4*, 347–360.

Hambleton, R.K., Swaminathan, H., Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park: Sage Press.

Hockemeyer, C. (1997). Using the basis of a knowledge space for determining the fringe of a knowledge state. *Journal of Mathematical Psychology*, *3*, 275–279.

Hockemeyer, C. (2017). The collection of computer science bibliographies: bibliography on knowledge spaces. http://liinwww.ira.uka.de/bibliography/Ai/knowledge.spaces.html. Accessed February 14, 2017.

Koppen, M. (1998). On alternative representations for knowledge spaces. *Mathematical Social Sciences*, *36*, 127–143.

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

Lynch, D., & Howlin, C.P. (2014). Real world usage of an adaptive testing algorithm to uncover latent knowledge. In *7th international conference of education, research and innovation (ICERI2014 Proceedings)*. Seville.

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 11th international conference on educational data mining* (pp. 607–612).

Nicewander, A., & Thomasson, G.L. (1999). Some reliability estimates for computerized adaptive tests. *Applied Psychological Measurement*, *23*(3), 239–247.

Reddy, A.A., & Harper, M. (2013). Mathematics placement at the University of Illinois. *PRIMUS*, *8*, 683–702.

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

Texas Education Agency (2015–2016). Technical Digest. http://tea.texas.gov/Student_Testing_and_Accountability/Testing/Student_Assessment_Overview/Technical_Digest_2015-2016. Accessed April 12, 2017.

Thissen, D. (2000). Reliability and measurement precision. In *Computerized adaptive testing: a primer*, 2nd edn (pp. 159–184). London: Lawrence Erlbaum Associates.

Vie, J.-J., Popineau, F., Bruillard, E., Bourda, Y. (2017). A review of recent advances in adaptive assessment. In A. Peña-Ayala (Ed.) *Learning analytics: fundaments, applications, and trends* (pp. 113–142): Springer International Publishing.

Vygotsky, L.S. (1978). *Mind and society: the development of higher mental processes*. Cambridge: Harvard University Press.

Weiss, D.J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences*, *2*(1), 1–27.

Woodruff, D., Traynor, A., Cui, Z., Fang, Y. (2013). A comparison of three methods for computing scale score conditional standard errors of measurement. ACT Research Report Series No. 7.

Zhu, Q., & Lowe, P.A. (2018). Split-half reliability. In Frey, B.B. (Ed.) *The SAGE encyclopedia of educational research, measurement, and evaluation*. Thousand Oaks: SAGE Publications.