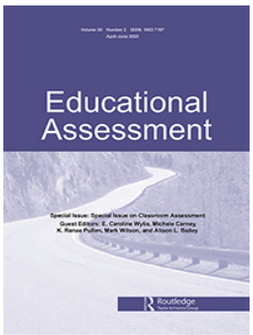

Routledge
Taylor & Francis Group

Educational Assessment

ISSN: 1062-7197 (Print) 1532-6977 (Online) Journal homepage: www.tandfonline.com/journals/heda20

Looking Beyond a Single Score: Examining Student Mathematical Strengths Using ALEKS Data

Christopher G. Lechuga, Jeffrey Matayoshi & Shayan Doroudi

To cite this article: Christopher G. Lechuga, Jeffrey Matayoshi & Shayan Doroudi (11 Apr 2025): Looking Beyond a Single Score: Examining Student Mathematical Strengths Using ALEKS Data, Educational Assessment, DOI: 10.1080/10627197.2025.2485928

To link to this article: https://doi.org/10.1080/10627197.2025.2485928

Published online: 11 Apr 2025.

EDUCATIONAL ASSESSMENT
https://doi.org/10.1080/10627197.2025.2485928

Routledge
Taylor & Francis Group

Looking Beyond a Single Score: Examining Student Mathematical Strengths Using ALEKS Data

Christopher G. Lechuga[a,b], Jeffrey Matayoshi[b], and Shayan Doroudi[a]

[a]University of California, Irvine, CA, USA; [b]McGraw Hill ALEKS, Irvine, CA, USA

ABSTRACT

Although popular educational theories regard ability as being intrinsically multidimensional, academic ability is typically measured with a single, overall score. In this paper, we examine data from the adaptive tutoring system ALEKS to compare three metrics that measure different constructs of mathematical ability that vary in dimensionality. We find that rankings on the abilities associated with teacher-created modules show substantially more variation than rankings based on estimates of overall ability. For example, using a multidimensional metric, we find that 80% of students (and more than half of the students with an incoming overall score in the bottom quartile) had above-median rankings on at least one module. We discuss the potential impact of our findings related to instructional practices such as ability grouping and teaching practices that value naming and recognizing student mathematical strengths within the classroom.

Introduction

Over the years, psychologists have proposed alternative theories of intelligence and ability that emphasize these constructs as multidimensional rather than as a single entity (Gardner, 2011; Guilford, 1982; Sternberg, 1996). Multidimensional models of ability have also become increasingly popular in psychometric research (Briggs & Wilson, 2003; Kang et al., 2022; McMullen et al., 2020; Walker & Beretvas, 2000). Nonetheless, consistent with Spearman's early theory of general intelligence (Spearman, 1904; Sternberg, 2010), in the United States, statewide standardized testing in school systems today typically measure student proficiency of an entire subject (like English Language Arts or Math) with a single scaled score. From this, we observe an incongruence in how some contemporary psychologists and education researchers conceptualize intelligence and how academic ability is measured with standardized testing. Robert Sternberg (1984) articulated the following concern with using unidimensional measures of intelligence:

> although we often need to make comparative judgments of people's intelligence or other skills, we ought to keep in mind that we are placing on a unidimensional scale attributes that are intrinsically multidimensional, with the result that the comparisons, although pragmatically useful, are not wholly valid. (p. 309)

Conscious of this tension of dimensionality, researchers have incorporated Spearman's early theory of general intelligence into more complex models that explore intelligence using several dimensions. One well-known and widely adopted model of intelligence for scientific use is the Cattell–Horn–Carroll (CHC) model, which models intelligence in a hierarchical structure consisting of three strata. Stratum III consists of general intelligence (g), Stratum II consists of abilities more specific than g such as verbal intelligence, spatial reasoning, fluid intelligence, and processing speed, and Stratum I consists of

CONTACT Christopher G. Lechuga lechugc1@uci.edu School of Education, University of California, 401 E. Peltason Drive Suite 3200, Irvine, CA 92617 USA

skills that can be improved with practice and instruction such as quantitative reasoning, visual memory, lexical knowledge, and perceptual speed (McGrew, 2009; Warne, 2016). A significant contribution of the CHC model is that it reconciled opposing theories of intelligence of whether it is unidimensional versus multidimensional in nature. Carroll (1993) showed that both perspectives could co-exist.

Item response theory (IRT) is another framework used to model and measure student ability, and it is commonly applied in achievement tests such as statewide standardized assessments. An IRT model may either be unidimensional (which measures one latent trait) or multidimensional (which measures multiple latent traits). Unidimensional IRT models are widely used in many well-known assessments, such as the SAT, ACT, and NAEP, as well as computerized adaptive assessments such as the SBAC and the Praxis exams[1](ACT, 2022; College Board, 2023; ETS, 2023; NAEP, 2023; Smarter Balanced, 2022). Yet, in spite of the prevalence of unidimensional IRT models, researchers have found that the misapplication of unidimensional models can lead to incorrect inferences about individual student proficiency on a single latent trait when tests and data are known to be multidimensional in nature (Briggs & Wilson, 2003; Kan et al., 2019; Mignani et al., 2006; Sheng & Wikle, 2007; Walker & Beretvas, 2000, 2003). This is mainly a result of the unidimensional model having less information than the multidimensional model on other latent abilities. Specifically, Walker and Beretvas (2003) showed that students who initially had lower estimates on a second dimension of mathematical ability (mathematical communication) in a multidimensional model tended to have lower estimates of general mathematical ability on a unidimensional model than they would have had on the first dimension (general mathematical ability) in the multidimensional model. In other words, the limitation to distinguish mathematical communication from general mathematical ability, led to "under" assessing these students in the unidimensional model.

Building on the previous idea of exploring various dimensional approaches to measuring ability, the present study explores this tension between unidimensional and multidimensional measures in ALEKS for a single-level domain such as middle school math or Algebra 1. ALEKS is an adaptive assessment and learning product, which bases its assessment approach on Knowledge Space Theory (KST) as opposed to IRT. It should be noted that because of the qualities that make KST distinct from IRT, the idea of dimensionality between these theories is not consistent. Yet, because KST and the ALEKS assessment make inferences about a student's knowledge on *every* single skill represented from among the entire curriculum (a concept largely absent in IRT), it is sensible to discuss ALEKS scores in terms of dimensionality, considering one could view the results from ALEKS as possessing a high number of dimensions equal to the number of skills in a course. In particular, the present study focuses on Stratum I of the CHC model, specifically ability[2] measured across hundreds of mathematical skills in an academic course.

Nevertheless, the present study strives to avoid any particular stance as to whether mathematics ability is *theoretically* one construct or many. As Sternberg (1984) highlights, a unidimensional view may often have pragmatic use for making comparisons, and thus, possesses value in certain contexts; yet, because such comparisons may not be entirely valid, it might be worth considering a multidimensional perspective. The point of departure of this paper is that even if various ability constructs (e.g., students' scores on different subtopics) tend to be highly correlated—as suggested by general intelligence (g)—it might still provide value to tease out student ability into multiple dimensions. The overarching research question we seek to investigate is to what extent does the distribution of student classroom rankings change under metrics used to measure different constructs of ability that vary in dimensionality. More specifically, we are interested in how ability rankings on subtopics of the curriculum compare to ranking students based on

[1]NAEP = National Assessment of Educational Progress, SAT = Scholastic Aptitude Test, ACT = American College Testing, SBAC = Smarter Balanced Assessment Consortium, Praxis = teacher certification exams written and administered by the Educational Testing Service (ETS).

[2]Because the term "ability" is most frequently used in the context of psychometrics and education measurement as well as in teaching practices such ability grouping and tracking, we use the word "ability" to express what is typically measured and reported in these contexts. This is akin to *crystallized intelligence*, or learning *knowledge*, as opposed to fluid intelligence, which is seen as being relatively independent of education and experience (J. L. Horn, 1967, 1968).

a single score at the start of the class or ranking students based on their overall score at different points in time. While we aim to demonstrate a distinction between such rankings, in broad terms, our study serves as an investigation into how dimensionality of various ALEKS scores provides evidence for specific interpretations and uses of such scores. In particular, we discuss how measures of ability on a subtopic of the curriculum could potentially have an impact on teacher perceptions of ability, on instructional practices that groups students based on ability, and on teaching practices that place value on naming and recognizing student mathematical strengths within the classroom.

To conduct our investigation, we use online student performance data from different classes on the adaptive learning platform, ALEKS. We compare three metrics that measure different constructs of mathematical ability using ALEKS data: (1) initial ability, (2) overall ability across time, and (3) module-specific ability. We formally introduce these metrics in the Section 'Research Design and Methods', but importantly, the first is a unidimensional measure, while the latter two are multidimensional. Even though the second and third metrics mathematically have the same number of dimensions, we expect the third metric to be "more multidimensional" in that the dimensions are less correlated with one another than for the second metric. We find that taking a multidimensional view of student mathematical ability (i.e., module-specific abilities) paints a different picture than a unidimensional view of ability. For example, students who might be traditionally perceived as having low ability may occasionally possess strengths beyond a majority of their classmates on some subtopics, and vice versa, suggesting that the notion of some students being universally "low ability" and others being universally "high ability" may be an oversimplification.

This paper is organized as follows. In the Section 'Literature Review', we will give a brief overview of ability measurement uses in the U.S. educational school system today. We will review the influences of ability measurement in prevailing education practices such as tracking and ability grouping, which have been shown to have an effect on academic achievement and students' self-concepts. Moreover, we will highlight works on how teachers' implicit theories of intelligence and classroom structure influence students' perceptions of their own ability as well as their perceptions of their classmates' ability. In Section 'The ALEKS System', we will provide a brief overview of the adaptive tutoring system ALEKS, the system from which we collect our data for our investigations. In the Section 'Research Design and Methods', we will outline our research design and methods in context. In the Section 'Results', we will present the results to our main analyses as well as results to supporting analyses that examine the reliability of our data. In the Section 'Discussion', we will conclude by discussing implications of the present work in the educational context as well as various limitations and potential future investigations.

Literature review

Standardized assessments

In response to the No Child Left Behind Act of 2001 (NCLB) and its successor the Every Student Succeeds Act (ESSA) in 2015, today, all students in the U.S. are required to take standardized assessments in grades 3–8 and once in high school to ensure quality education. These assessments, which measure ability for an entire subject such as English Language Arts or Mathematics, have a variety of forms of reporting. Statewide assessments typically report student performance with a unidimensional scaled score for an entire subject, which are often accompanied with proficiency levels and/or percentile ranks. Scores are also provided on different subtest domains of a test battery (e.g., verbal and quantitative subtests for the SAT and ACT) or on subdomains from a single test. Related to the latter, studies have shown that subscores are often less reliable than overall scores (AERA, APA, & NCME, 2014; Sinharay, 2010; Sinharay et al., 2018; Smarter Balanced, 2022) and because of this, as in the case of the SBAC, subscores are not reported as observed scaled scores, but rather as performance levels (Smarter Balanced, 2022). Nonetheless, a unifying characteristic is that

many standardized assessments are based on a unidimensional IRT model, even when multiple scores are provided.

Educational practices surrounding assessment and ability

Statewide standardized assessments, including interim and formative assessments potentially play a significant role in educational practices. In recent years, in large part due to ESSA, which provided for the use of federal funding by the state and local educational agencies to support personalized learning (Gross et al., 2018; Heinrich et al., 2020), standardized testing has placed additional emphasis on informing practices of instruction in the classroom to address student-specific needs. This aligns with a surge of interest on formative assessment and educational tools designed to deliver a personalized learning experience. Specifically, public education has looked to private educational technology companies who offer digital tools and services with a focus on assessment and personalization. In a large survey of 4,600 teachers, the Bill & Melinda Gates Foundation (2015) found that 93% of teachers use digital tools, including digital assessments tools. Teachers reported that these tools help them gather data about individual students to guide either whole-class instruction or small-group differentiated instruction where teachers group students based on their ability level.

Assessments have often been used to make ability judgments for supporting teaching practices such as tracking and ability grouping. The practice of "tracking" segments students of the same grade level into different classrooms based on student ability for the purpose of providing instruction and feedback that is tailored to the ability level of the class. The similar practice of ability grouping[3] involves dividing a classroom consisting of students of widely different abilities into small homogenous groups of like ability, often for the purpose of collaborative learning and/or differentiated instruction. In either case, these strategies are meant to improve teaching and learning through personalization and meeting the student where they are in terms of their development. Research has shown that periodic formative assessments at the skill level are strongly recommended when forming homogenous groups of like ability (Slavin, 1987).

Importance of supporting practices based on ability

A large body of work has studied the effects of tracking and ability grouping. This literature has explored a variety of themes concerning effects on overall student achievement (C. L. C. Kulik & Kulik, 1982, 1984; J. A. Kulik & Kulik, 1992; Slavin, 1987, 1990), including effects on specific ability groups (Dawson, 1987; Lleras & Rangel, 2009; Oakes, 2005; Rowan & Miracle, 1983), implications concerning social and racial discrimination (Cipriano-Walter, 2015; Gallardo, 1994; Oakes, 2005) as well as teacher expectations and quality of instruction (Finley, 1984; Kelly, 2004; Oakes, 2005; Trimble & Sinclair, 1987). Researchers have also pointed to the lack of mobility patterns, which often keep students in low-ability groups from which they cannot escape (Boaler, 2005; Castle et al., 2005; MacIntyre & Ireson, 2002; Rowan & Miracle, 1983). Moreover, evidence has suggested that students may even be misplaced in ability groups. These studies report that while different ability groups have statistically significantly different assessment scores on average, there is still a substantial overlap in their score distributions, giving rise to doubts about whether groupings are homogeneous in terms of ability (MacIntyre & Ireson, 2002; Rosenbaum, 1980).

Teaching practices are also influenced by teacher perceptions and judgments on student ability, which play a role in the placement and expectations of students (Hoover & Abrams, 2013; Meissel et al., 2017). In particular, teacher perceptions of intelligence may lead to differential treatment on different ability groups of students. For example, Lee (1996) found that teachers holding an entity (or fixed) view of intelligence tended to evaluate and provide

[3]The term "ability grouping" in some contexts is used as an umbrella term to represent whenever students are being grouped by ability. Therefore, sometimes tracking is viewed as one kind of ability grouping. However, in this paper, ability grouping is used to represent *within-class grouping* of students on the basis of ability.

feedback based on perceived ability of the student, which was usually consistent with low expectations. Studies have shown that teachers' implicit theories of intelligence have been known to have effects on student self-concepts, aspirations, and achievement depending on whether teachers hold a fixed view of intelligence versus a malleable view of intelligence (Blackwell et al., 2007; Canning et al., 2019; Lee, 1996; Muenks et al., 2020).

In light of unreliable group compositions and potential negative impacts associated with teacher perceptions of ability, innovations toward measuring ability, with an emphasis on skill-level ability, are particularly critical for supporting teaching practices that aim to tailor instruction to students' specific needs. Depending on the context, when determining ability groups, educators often take into account a variety of factors such as gender, learning perspectives, attitudes toward group work, personality traits, engagement, and motivation (Donovan et al., 2018, Kanika et al., 2022; Sanz-Martínez et al., 2019). More directly related to ability, educators often use assessments that target more specific skills to group students into flexible groupings that can be changed over time as students are reassessed with formative assessments (Missett et al., 2014; Slavin, 1987; Tieso, 2003). Researchers have advocated for this; for example, Slavin (1987) suggests that ability grouping is most effective when it meets several criteria including "when it greatly reduces student heterogeneity in a specific skill [and] when group assignments are frequently reassessed." Nonetheless, when discussing "specific skill," Slavin (1987) largely refers to scores on reading tests and mathematics tests as opposed to IQ and general achievement tests. Thus, the information needed to achieve specific-skill ability likely requires frequent classroom assessments and vigilant monitoring of student performance, which can be quite difficult for educators juggling a multitude of responsibilities. As such, the present study offers an innovative solution for measuring student mathematical ability, which may potentially be useful for identifying specific-ability strengths and weaknesses to support instructional practices focused on personalized learning.

Perceptions of ability

Teacher perceptions of student ability, as well as students' perceptions of their own and classmates' ability, have also been studied through the lens of how teachers structure their classrooms (Rosenholtz & Wilson, 1980; Simpson, 1981; Rosenholtz & Simpson, 1984a, 1984b). In particular, students tend to form different conceptions of ability depending on whether they belong to classrooms with a "unidimensional structure" versus a "multidimensional structure." Researchers define unidimensional classrooms as ones that typically have an undifferentiated curriculum and instruction; put little value on student autonomy; rely on whole-class instruction or groups formed on the basis of ability; and put emphasis on frequently grading assignments. In contrast, multidimensional classrooms typically individualize curriculum and instruction; give students more autonomy; have students work individually or in groups that are not formed on the basis of ability; and put less emphasis on grading assignments or summative evaluations. According to Simpson (1981) and Rosenholtz & Simpson (1984a, 1984b), unidimensional classrooms tend to have an increased amount of stratification of perceived abilities. That is, among student perceptions in the classroom, there is a greater consensus on who is perceived to have low versus high ability. On the other hand, multidimensional classrooms often generate perceptions of ability that are more dispersed. That is, there is less consensus within the classroom on who is perceived as having low versus high ability.

The present study draws inspiration from Simpson (1981) and Rosenholtz & Simpson (1984a, 1984b). However, instead of looking at how dimensional features of classroom structure influence students' (subjective) perceptions of ability, we examine *how different metrics used to measure different constructs of ability influence the objective distributions of student rankings*. Although not tested in the present study, we hypothesize such rankings could in turn influence both teachers' and students' perceptions of ability.

The ALEKS system

Brief description of ALEKS

ALEKS, which stands for Assessment and LEarning in Knowledge Spaces, is an online intelligent learning and assessment system used by millions of students for math and various other STEM disciplines in both K-12 and higher education (About ALEKS, 2021). The system is an instantiation of Knowledge Space Theory (KST), developed by Jean-Paul Doignon and Jean-Claude Falmagne in 1985 (Cosyn et al., 2021; Doignon & Falmagne, 1985) for the purpose of assessing and representing domain-specific knowledge within a course (e.g., Algebra 1). A typical ALEKS course consists of several hundred skills or problem types usually referred to as *items*. An ALEKS item is designed to cover a specific piece of knowledge from the entire curriculum of a course. Additionally, items are organized in a knowledge structure (or mapping) that defines prerequisite and postrequisite relationships among item pairs. The same item can simultaneously be a prerequisite for harder items and a postrequisite for easier items in the course.

At the start of a course, all students take an initial adaptive assessment that determines their *knowledge state* (or *state*) expressed as the set of items the student knows from the full curriculum (i.e., the entire set of items in the course). Unlike typical assessments that measure student knowledge with a single numeric score, ALEKS employs KST to represent a student's knowledge with a knowledge state, which provides a multidimensional view of ability[4] needed for our investigations. After taking the initial assessment, a student's state defines the set of items they are ready to learn next from among the full curriculum. Therefore, the initial assessment determines where the student starts in the course. From this point, the ALEKS system guides each student on a personalized learning path where they increase their knowledge through practice and periodic assessments. Students' learning paths are governed by the knowledge structure, which requires students to learn prerequisite items before working on postrequisites (both of which are a part of the course).

ALEKS also provides teachers and school administrators with administrative and classroom management tools to gauge student progress and allow for flexible instruction through various course customization options. Most notably, teachers may customize their ALEKS class content by selecting or deselecting any item from the entire course (which determines whether the item is assessed *and* taught). Additionally, teachers may sequence the course content into adaptive assignments called *modules*.[5] These assignments are adaptive in the sense that students must first display knowledge of prerequisite items of the module before working on postrequisite items in the module. However, because students will have different states, this learning path of going from prerequisites to postrequisites will not look the same for any two students. Each student will have a unique learning path through the module, but with the same end goal of displaying knowledge for the same set of items in the module that the teacher has selected for the whole class. In other words, students are on different paths moving toward the same destination.

How ALEKS determines a student's knowledge

We examine data from ALEKS to compare three metrics that measure different constructs of mathematical ability. Each metric depends heavily on the results of the ALEKS initial assessment. Here we provide some brief details regarding the ALEKS assessment to better contextualize our

[4]We remind the reader that we use "ability" to mean crystalized intelligence, or learned knowledge, which is why we describe a knowledge state as a multidimensional representation of ability. We use "knowledge" and "ability" interchangeably for the remainder of the paper.

[5]ALEKS modules make up and sequence the entire item set of the course. It is on this item set that the student is assessed on when taking the initial assessment. While the initial assessment cannot ask the student every item in the course, we say that the student is assessed on the entire course because the assessment is designed to determine what the student knows from the entire course. We provide more details on how this is done in the subsection 'How ALEKS determines a student's knowledge'.

measures. For additional and technical details regarding the mechanisms of the ALEKS assessment, the reader is encouraged to refer to Section 1.3 of Cosyn et al. (2021).

The ALEKS initial assessment is designed to efficiently determine, with relatively few items (or questions), a student's state from the entire course consisting of several hundred items. The items asked in the assessment are among the items in the course and will be in one of the modules if the course is sequenced into modules. The student's state is expressed as a subset of all the items in the course that the student knows (or can solve correctly on their own). Even though the assessment stops at typically 29 questions, it is possible for the student to display knowledge of more than 29 items due to the inferences made on each question. The assessment is adaptive in that subsequent items are given to the student based on their previous answers with the goal that each item is informative about the student. With each student answer, the likelihood of each possible knowledge state is adjusted (increased or decreased), resulting in a few likely states for the student at the end of the assessment. Among these, the system selects the most likely state for the student, which is referred as the student's initial state.

Research design and methods

The present study consisted of 42 classes and a total of 915 students. A breakdown of classes and number of students by grade/course level is given in Appendix A. We also outline the set of requirements for inclusion into our dataset, which involve the number of students per class, activity in each class, the number of items per module, the number of modules per class, and module due dates.

Measures of ability used

Because a student's state is a fine-grained representation of the student's knowledge of the entire ALEKS course, it becomes difficult to compare states for many students. For this reason, we examine three metrics associated with a student's state, which are the metrics we adopt for the present study. The first, which we call the student's *initial ability*, is simply the cardinality of the student's initial state (i.e., the number of items the student knows from the entire course at the start of the course). In other words, it is meant to measure their overall mathematical ability on the material in the course at the beginning before they start. This metric produces one measure (or score) for each student.

The second metric is the cardinality of the student's state across time, which changes over time as the student practices in the system and is periodically assessed in the system. This metric will be referred to as *overall ability across time*. It gives the number of items the student knows from the entire course. These items can be any item from the course: an item in the current module, a prerequisite item from a previous module, or an item in a subsequent module (that would have been determined by the initial assessment). However, after the initial assessment, as a student progresses through the modules, this measure will inevitably consist of more and more items from previous modules that the student learned. Overall ability across time is meant to measure a student's mathematical ability on the material in the course at different times. Consequently, the metric is not fixed, as it produces a distribution of measures (or scores) for each student. As a marker across time, these scores were recorded at the start of every module.

While the student's state is a detailed representation of their ability, a great deal of information about the student's state is lost when the state is summarized by a score. This motivates us to consider our final metric, which we call *module-specific ability*. This too produces measures across time (recorded at the start of each module), but on individual teacher-created modules rather than on the entire course. In other words, in a sense, the measures obtained from this metric is meant to recapture some of the information lost from the previous two metrics by obtaining scores on subtopics of the curriculum that we can assume teachers care about. More specifically, the score for module-specific ability is obtained by the number of items the student already knows in the module. Thus, the score is meant to measure the student's mathematical ability on the module (as opposed to the entire course) *before* the student begins working on

that module. Note that module-specific ability is primarily a function of the student's initial state since a students' understanding of each module was assessed in the initial assessment and students generally cannot practice items in a given module before working on that module; however, learned items can play a role in breaking ties when two or more students have the same module-specific score. A more precise definition is given in Appendix B, along with tie-breaker rules for when two students have the same score, which will allow for ranking students in our analyses.

Procedure for ranking students

For a given metric, within a class each student i was ranked from 0 to $n-1$, where n is the number of students in the class. Because not all classes have the same number of students, the original ordinal rank of the student, r_i, was transformed as a way to standardize the rankings. A common way to transform the original ordinal rank is to compute R_i, the quotient of (a) the difference between r_i and the minimum original rank ($r_{min} = 0$) and (b) the difference between the maximum rank ($r_{max} = n-1$) and minimum rank (Denning et al., 2018, Elsner, 2021; Goulas & Megalokonomou, 2021; Murphy & Weinhardt, 2020).

$$R_i = \frac{r_i - r_{min}}{r_{max} - r_{min}} \tag{1}$$

Note, the larger the ranking (for both r_i and R_i), the higher the student is ranked in their respective class. Also, because R_i is a number between 0 and 1, with 1 being the highest transformed ranking, this value bears close resemblance to a percentile. As such the transformed ranking is often referred to as a percentile in the literature, which is how it is referred to in the present study.

For every class, each student was ranked on the three metrics of ability. The rank on initial ability remained fixed throughout the course, while the ranks for overall ability across time and modules-specific ability do not remain fixed, but rather produce a distribution of rankings (or percentiles). Figure 1 shows an example of a hypothetical student's ability rankings for each of the three metrics. The pie charts are meant to provide a visual of how the measure of ability is changing across time (or across modules), where each slice (the shaded and non-shaded region) represents a module and where the shaded region represents what the student knows in the module. This student ranked in the 20th percentile after the initial assessment (i.e., at the start of Module 1). As shown in the first row, the percentile remains fixed with a static pie representing their initial ability. The student's overall ability across time changes, as learning occurs on each module. This is depicted in the second row, as the pie is filled in on a slice after the module has ended. The increase in overall ability may or may not result in an increase in percentile rank as other students in the class are presumably learning as well. In the final row, ability is measured on a specific slice (or module) of the pie at the start of the module. As depicted, the percentile rank is based on a specific module. Therefore, while a student may be ranked low overall (as well as on some modules), the student may exhibit a strength as shown in Module 4. Finally, it is worth noting that initial ability contributes in part to module-specific ability since module scores are derived from the student's current state (see Appendix B), and a student's current state builds upon their initial state.

Methods

We examined three statistics to investigate how student rankings differ under our three different metrics of ability. The first is the *range* of a student's classroom rankings, which will provide us with a summary of how much students' rankings change under each metric of ability throughout the class. The second is the student's *maximum* classroom ranking, which will shed light on the extent to which students exhibit relative strengths throughout the class and what metric of ability (of the three) seems best to capture these strengths. The third is a *correlation* computed between student's rankings. This will offer insight as to whether high rankings in one moment in time (or module) tend to produce high rankings in other

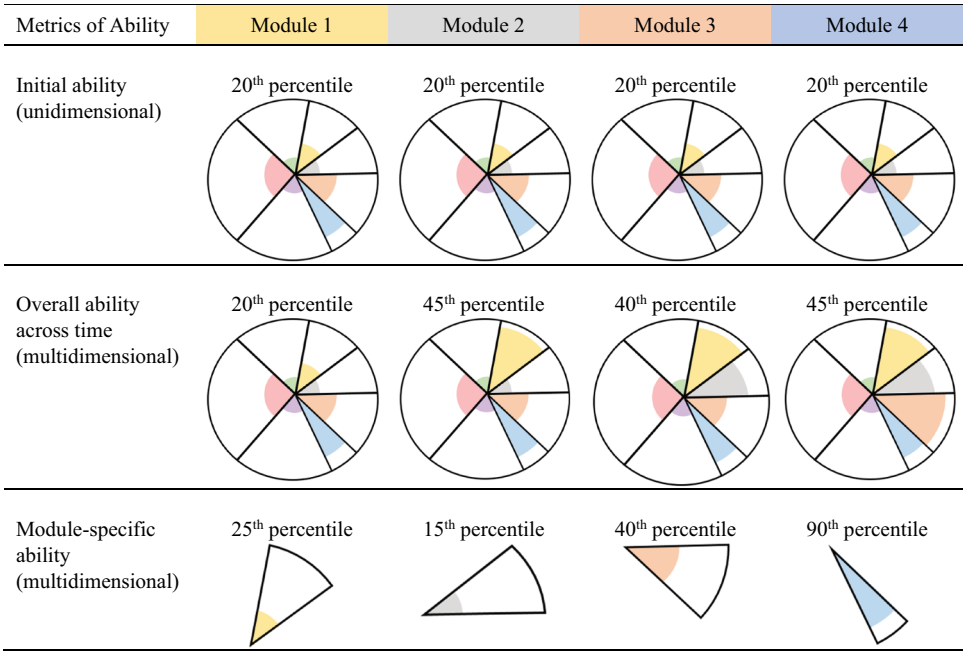

Figure 1. Percentile rankings of a hypothetical student on each of the three metrics of ability.

moments in time (or other modules), a phenomenon consistent with Spearman's theory of general intelligence. The analyses done on these three statistics are described below and summarized in Table 1.

The "Range" column in Table 1 indicates that the range of student rankings was computed for each metric of ability. Note, this is not applicable for initial ability since this measure remains fixed. However, for overall ability across time and module-specific ability, a distribution of ranges was produced for each one of these metrics. For these distributions, a 95% confidence interval (CI) around the mean was computed. We also found students' maximum rankings on all three metrics of ability (represented in the "Maximum" column in Table 1). For initial ability, this is simply the initial ability ranking. By definition, this distribution has a mean of 0.5 and will always be 0.5 for any sample of data. For overall ability across time and module-specific ability, we obtained a distribution of maximum rankings for each one of these metrics. For these distributions, a 95% CI around the mean was computed. We were also interested in the proportion of students with a maximum rank greater than half the class (i.e., greater than 0.5). Once again, this proportion was computed under all three metrics of ability. For initial ability, this proportion by definition is 0.5. For the other two metrics, we computed this proportion along with a 95% CI. All analyses in Table 1 were also done on the subset of students who may be traditionally labeled as low-ability (i.e., those who scored in the lower quartile in the initial assessment).

Table 1. Summary of Statistics and Analyses Performed on Each of the Three Metrics of Ability.

	Statistic		
	Range	Maximum	Correlation
Initial ability	NA	• Mean: 0.5 • Proportion of students with max. ranking greater than 0.5	NA
Overall ability across time	Mean and CI	• Mean and CI • Proportion of students with max. ranking greater than 0.5	Mean and CI
Module-Specific ability	Mean and CI	• Mean and CI • Proportion of students with max. ranking above half the class	Mean and CI

Because students are grouped (or clustered) into classes, there exists some dependency in the data, complicating the computation of confidence intervals. Thus, to handle this dependency, we used the *cluster bootstrap*, a version of the bootstrap specifically designed to work with clustered data (Field & Welsh, 2007). To apply the procedure, we started by randomly sampling classes with replacement from the original data set. Once we had a sample of classes equal to the number of classes in the original data, we then combined all of the student data from those classes to get our bootstrap sample. For that bootstrap sample, we computed our statistic of interest, and then this entire procedure was repeated until we had generated 100,000 bootstrap samples. Lastly, the confidence interval for our statistic was computed using the bias-corrected and accelerated (BCa) method, a procedure introduced by Efron (1987) that adjusts for skewness in the bootstrap distribution.

The third and final statistic computed in our analyses was the correlation between rankings throughout the class (noted in the "Correlation" column in Table 1). For each class, the Pearson correlation coefficient, r, between students' pairwise rankings was computed. This was done for the rankings under overall ability across time and module-specific ability separately. (Note, such analysis is not applicable to initial ability since there are no pairs of rankings for a given student under this metric.) Figure 2 shows an illustration of the pairs of rankings formed, which were used to find a correlation at the class level. The illustration is the same for overall ability across time and module-specific ability. That is, $R_{i,j}$ may represent either the transformed rankings on overall ability at the start of a module or the transformed rankings on module-specific ability, depending on the analysis. For a given class with m modules and n students, there are $\binom{m}{2}$ pairs of rankings per student, producing a total of $n \cdot \binom{m}{2}$ pairs for which a correlation r was found. This was repeated for every class, producing a distribution of correlations for which the average and confidence interval around this average were computed.

Results

Variation (range) in student rankings

The average range in student rankings under each metric of ability is summarized in Table 2. The table captures the averages and CIs for both the full dataset and the filtered dataset of students scoring in the lower quartile of the initial assessment. Figure 3a shows the ranges in percentile rank plotted for the full dataset. The ranges in percentile rank for overall ability across time produced a distribution with a mean of 0.222 (CI $[0.193, 0.250]$). The ranges in percentile rank for module-specific ability produced a distribution with a mean of 0.452 (CI $[0.417, 0.485]$). Thus, when ability is measured on specific modules, the range has an average value that is roughly twice the average value observed for range on overall ability across time; this difference is significant as indicated by the non-overlapping CIs. Figure 3b again shows the ranges in percentile rank, but this time only for those scoring in the lower quartile on the initial assessment. Perhaps

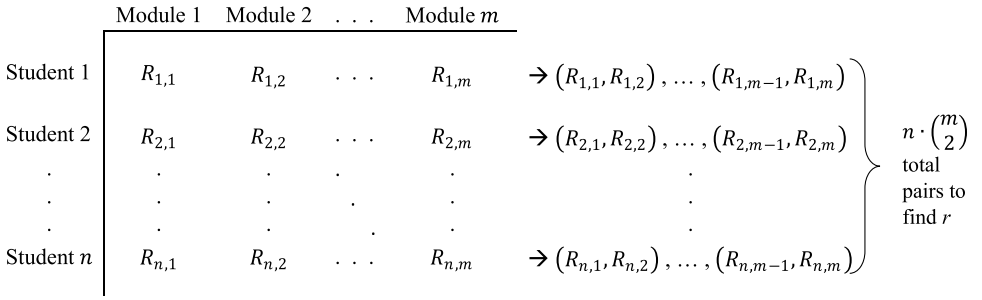

Figure 2. Pairs of rankings used to find a correlation between rankings in each class.

Table 2. Average Range in Student Rankings and CIs under Each Metric of Ability for the Full Dataset and Filtered Dataset of Students Scoring in the Lower Quartile of the Initial Assessment.

	Range	
	Full dataset	Lower quartile
Initial ability	NA	NA
Overall ability across time	0.222 [0.193, 0.250]	0.230 [0.197, 0.262]
Module-specific ability	0.452 [0.417, 0.485]	0.453 [0.411, 0.497]

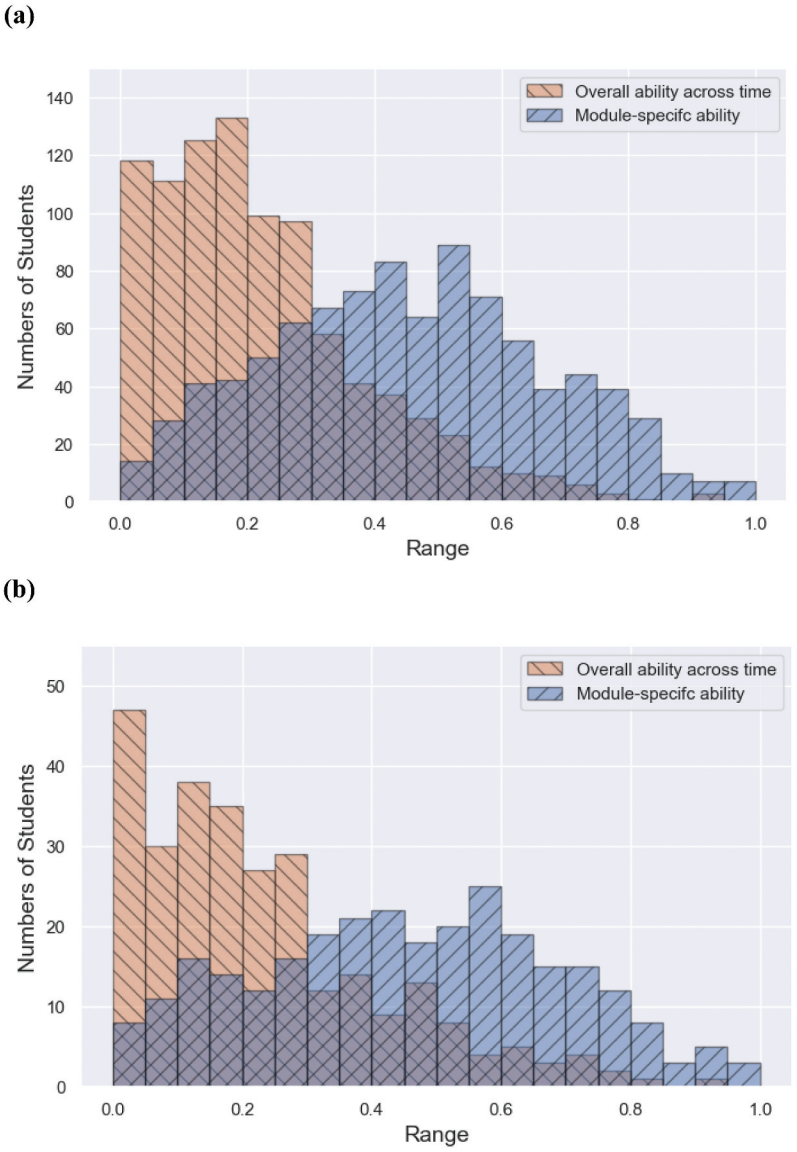

Figure 3. Range in percentile rankings under the three metrics of ability for the full dataset (A) and for students scoring in the lower quartile of the initial assessment (B).

Table 3. Average Maximum Student Ranking and CIs under Each Metric of Ability for the Full Dataset and Filtered Dataset of Students Scoring in the Lower Quartile of the Initial Assessment.

	Maximum	
	Full dataset	Lower quartile
Initial ability	0.5 NA	0.157 [0.129, 0.225]
Overall ability across time	0.612 [0.597, 0.627]	0.385 [0.340, 0.437]
Module-specific ability	0.729 [0.711, 0.748]	0.545 [0.499, 0.603]

surprisingly, the distribution of ranges for students who may be traditionally labeled as low-ability (Figure 3b) is very similar to the distribution for all students (Figure 3a). The means (and CIs) for the lower quartile students were also very similar to the means for all students: 0.230 (CI $[0.197, 0.262]$) and 0.453 (CI $[0.411, 0.497]$) for overall ability across time and module-specific ability, respectively.

Maximum student rankings

The average maximum student ranking under each metric of ability is summarized in Table 3. The table captures the averages and CIs for both the full dataset and the filtered dataset of students scoring in the lower quartile of the initial assessment. Figure 4a shows the distributions of maximum rankings students achieved under the three metrics of ability. For initial ability, because students are ranked once, we expect the maximum ranking distribution to be somewhat uniform. Indeed, this is what we see. It is not perfectly uniform because of ties as well as because the transformed rank, R_i, has some dependency on the number of students in the student's class. By definition, however, the average maximum rank for initial ability is 0.5. In comparison, the overall ability across time produced a distribution that is slightly left-skewed with a mean of 0.612 (CI $[0.597, 0.627]$). This suggests that students' overall ability tended to fluctuate in ordinal rank across time. Moreover, the peak occurred at $R = 1$, with 80 students (9%) ranked as a "top performer" in their respective class for at least one moment in time. This is almost double the number of top performers under initial ability (42 students, 5%). We also note that roughly 20% of students ranked above the 90th percentile at least one moment in time. For module-specific ability, the maximum ranking distribution was even more left-skewed than overall ability across time. This distribution had a mean of 0.729 (CI $[0.711, 0.748]$), meaning that with high confidence, we can conclude that a student ranks higher than 70% of their classmates on at least one module on average. We also observe a considerable proportion of students (15%) who achieved a maximum percentile rank of 1.0, triple the number of top performers under initial ability. Additionally, we note that the average maximum rankings for module-specific ability were significantly higher than the average maximum rankings for overall ability across time (as indicated by the non-overlapping CIs).

Figure 4b shows the same plots as Figure 4a, but for those scoring in the lower quartile on the initial assessment. For this sample, we observe some noticeable differences compared to those from the full dataset. The first difference we see is that the shapes of the graphs are no longer roughly uniform (for initial ability) and no longer left-skewed (for overall ability across time and module-specific ability). The average maximum rankings for initial ability, overall ability across time, and module-specific ability were 0.157 (CI $[0.129, 0.225]$), 0.385 (CI $[0.340, 0.437]$), and 0.545 (CI $[0.499, 0.603]$), respectively. While these are considerably lower than the averages seen in the full dataset, it is interesting to see that some students may still achieve a relatively high rank on a specific module even among those who may be traditionally perceived as having low ability. Table 4 summarizes the proportion of students (and CIs) who displayed a relative strength in their class under the three metrics of ability. We define the occurrence of a relative strength when a student had a maximum rank greater than 0.5. All proportions and CIs were done for the full dataset as well as on the sample of students who scored in the lower quartile on the initial assessment. Indeed, we see that the large majority of students (80.8%,

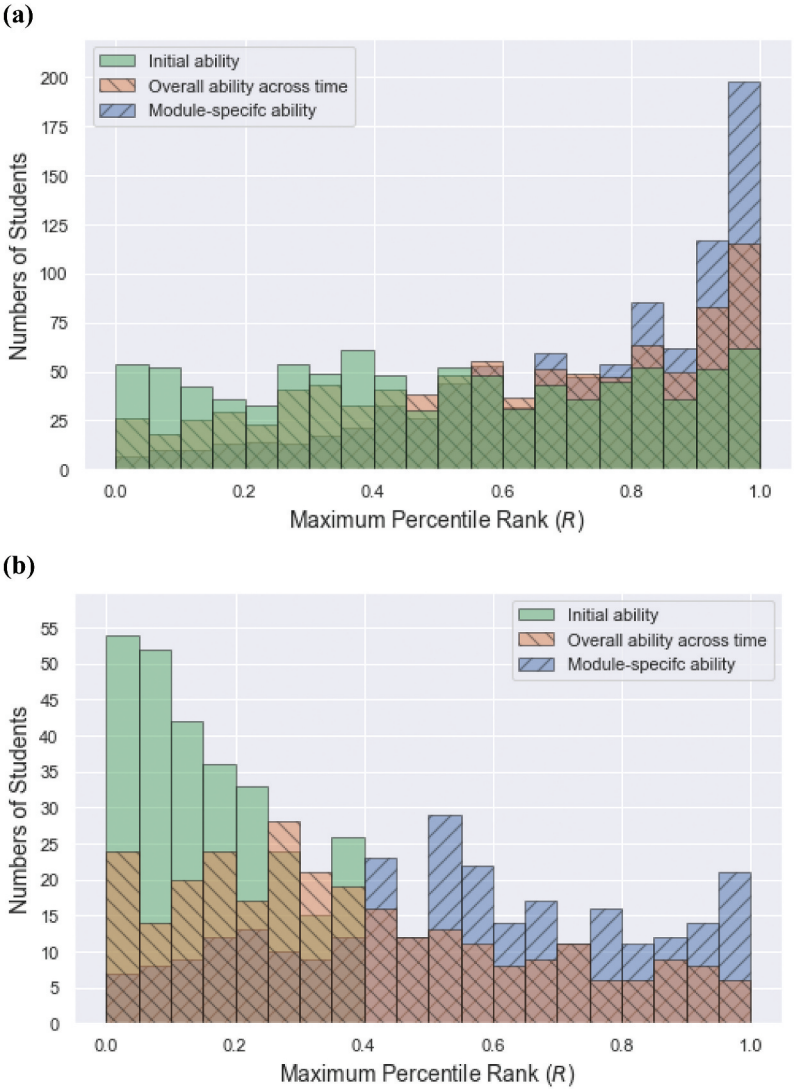

Figure 4. Maximum percentile rankings under the three metrics of ability for the full dataset (A) and for students scoring in the lower quartile of the initial assessment[6] (B).

CI $[0.781, 0.835]$) displayed a relative strength on a specific module for the full dataset. And perhaps surprisingly, among those in the bottom quartile in initial ability, 56.7% (CI $[0.488, 0.650]$) displayed a relative strength on a specific module. Interestingly, there were three classes (boxed in Figure 5) with everyone in the class achieving a maximum module rank greater than 0.50.

Correlation of student rankings

The average class correlation of student rankings under each metric of ability is summarized in Table 5. The distribution of Pearson correlation coefficients under overall ability across time and module-specific ability is plotted in Figure 6a. These distributions had a mean of 0.886 (CI

[6]The data in the sample consisted of students scoring in the lower quartile of the initial assessment. Nonetheless, we see that the initial ability distribution contains measures above 0.25. This is due to ties in certain classes, which means there is more than 25% of the full dataset captured in this sample.

Table 4. Proportion of Students (and CIs) Who Had a Maximum Rank Greater than 0.5 for the Full Dataset and Filtered Dataset of Students Scoring in the Lower Quartile of the Initial Assessment.

	Proportion with maximum rank greater than 0.5	
	Full dataset	Lower quartile
Initial ability	0.479 [0.444, 0.491]	0.000 [0.000, 0.000]
Overall ability across time	0.636 [0.611, 0.664]	0.291 [0.213, 0.386]
Module-specific ability	0.808 [0.781, 0.835]	0.567 [0.488, 0.650]

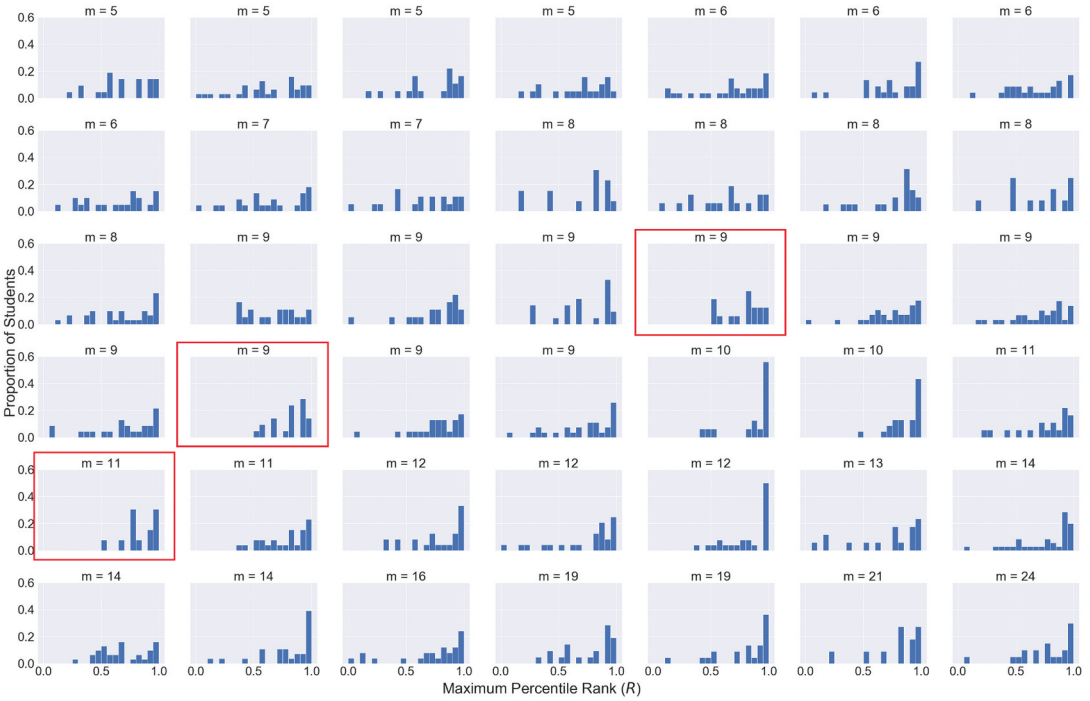

Figure 5. Distribution of maximum percentile rankings for module-specific ability for each class.

Table 5. Average Class Correlation (and CIs) of Student Rankings under Each Measure of Ability for the Full Dataset and Filtered Dataset of Students Scoring in the Lower Quartile of the Initial Assessment.

	Correlation	
	Full dataset	Lower quartile
Initial ability	NA	NA
Overall ability across time	0.886 [0.854, 0.917]	0.795 [0.755, 0.835]
Module-specific ability	0.696 [0.657, 0.735]	0.441 [0.365, 0.517]

[0.854, 0.917]) and 0.696 (CI [0.657, 0.735]), respectively. Based on these results, we observe that there is a strong correlation between students' pairwise rankings for both metrics of ability, thus exhibiting evidence consistent with Spearman's general intelligence, *g*. In other words, students who rank high at the start of one module tend to rank high at the start of other modules. Likewise, students who tend to

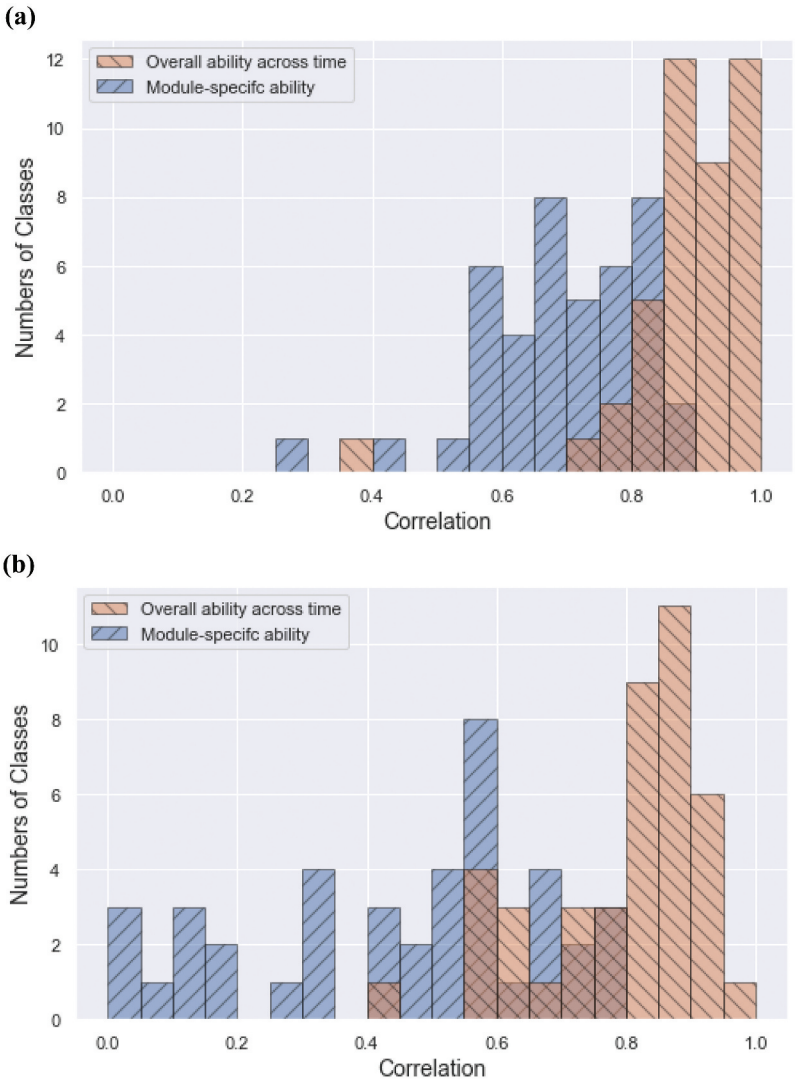

Figure 6. Distribution of correlations between pairwise percentile rankings on each class: (A) full dataset of students used, (B) students ranking in the lower quartile in the initial assessment used.

rank low at the start of one module tend to rank low at the start of other modules. While this is more pronounced for overall ability across time, it is still visible for module-specific ability, even though it is the metric that is most multidimensional and that mostly reveals students' relative strengths as shown in Figure 4a and Table 4. We also computed the correlations only using the dataset consisting of students who scored in the bottom quartile of the initial assessment. These distributions are plotted in Figure 6b. Here, the average correlation under overall ability across time was 0.795 (CI $[0.755, 0.835]$) and the average correlation under module-specific ability was 0.441 (CI $[0.365, 0.517]$). We notice that the average correlation decreased considerably from the full dataset to the filtered dataset for both metrics. This is to be expected since students in the filtered dataset tend have relatively low rankings on average. For the filtered data, this meant that there was a high density of data points with low pairs of rankings and a low density of data points with high pairs of rankings, thus contributing to the smaller overall correlation.

Reliability

The present study has relied on scores and distributions of scores to provide insight into the variability of classroom rankings under different ability metrics. Thus, it is necessary to discuss the reliability of such scores. We begin by discussing the scores obtained from initial ability and overall ability across time, as these both give the number of items the student knows from the entire course. For the reliability of these scores, we heavily rely upon Doble et al. (2019), which examined the reliability of the ALEKS assessment scores that assessed students' mastery of high school-level mathematics for the purpose of recommending placement in a post-secondary mathematics course. This study focused on the test–retest repeatability results for a student. In the absence of data from students taking the assessment multiple times and no standard approach for measuring the reliability of adaptive assessments (which present different questions based on the test-taker's performance during the assessment), researchers borrowed inspiration from psychometrics using simulated assessments based on the assessment model to estimate the reliability of scores. Indeed, Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) recommends such a method for determining the reliability for adaptive tests. The study compared scores from over 700,000 actual assessments with scores obtained by simulated assessments. Results showed that actual scores and simulated scores were highly correlated. Additionally, the conditional standard error of measurement (CSEM), a recommended measure for reliability for an adaptive test (ACT, 2012; Green et al., 1984; Nicewander & Thomasson, 1999; Thissen, 2000; Weiss, 2011), showed comparable results to other well-known assessments, including the ACT Computer-Adaptive Placement Assessment and Support System (COMPASS).

While the analysis of Doble et al. (2019) strongly support the reliability of scores from the first two metrics, our third metric, presents a different situation. In particular, module-specific ability is somewhat akin to test subscores, because they are based on teacher-created modules presumably focused on certain subtopics of the curriculum. According to Standards for Educational and Psychological Testing (SEPT), when tests provide subscores, the distinctiveness and reliability of such scores should be demonstrated to determine if they possess added value over a total score. Though, historically, it has been shown that subscores tend not to possess added value over a total score (Sinharay, 2010; Sinharay et al., 2007, 2018). Nonetheless, Sinharay (2010) provides some insights into when subscores have a greater chance of producing added value. Specifically, in an investigation examining multiple tests, including a simulation study employing a multidimensional IRT model, Sinharay (2010) found that subparts of a test typically needed at least 20 items for subscores to have added value. Findings also indicated that the level of correlation between items within a particular subtopic has an influence. That is, less correlation between items (i.e., greater distinctiveness between the items) increases the chances of subscores exhibiting added value. In analyzing the interaction between number of items and correlation, it was noted that fewer items required less correlation between the items and more items could support higher levels of correlation between the items for there to be a high chance of added value on the subscore.

In light of the aforementioned discoveries by previous researchers, we remind the reader that in order to adhere to the Knowledge Space Theory framework (by which ALEKS is based on), ALEKS items are designed to be distinct from one another, each representing a discrete piece of knowledge from the entire curriculum. Therefore, although it is not tested in the present study, we hypothesize that this would only enhance the prospect of ALEKS subscores having added value over an overall score. Furthermore, motivated by the findings in Sinharay (2010) and the recommendation provided by Sinharay et al. (2018), we recomputed our main analyses (presented in Tables 2–5) on modules with at least 20 items. This reduced our dataset by almost a third, reducing the number of unique modules from 431 to 297. We found that our recomputed results (outlined in Appendix C, Tables A1–A4) did not change very much, providing no alternative interpretation to our findings.

Reproducing our results using modules of at least 20 items is encouraging. Nonetheless, we recognize a limitation of our analyses, which is we did not directly test the reliability of the percentile rankings obtained from module-specific ability. On this matter, it remains equally relevant to highlight that scores obtained in the present study for module-specific ability are not true test subscores, as modules are created assignments and differ from class to class even among classes using the same ALEKS course. Therefore, not all students were "assessed" on the same module-specific abilities across the dataset. Moreover, these scores are standardized rankings (akin to percentiles), which reflects a student's position in the module relative to a reference group (i.e., a class) and is always roughly uniformly distributed for the class on a particular module. Consequently, the characteristics of our data make it difficult for testing the reliability of such "subscores." Indeed, because of the complexity of our data, characterized by the presence of different module-specific abilities, different number of modules per class, and standardized percentile rankings, further research is required to develop robust strategies for assessing the reliability of our module-specific scores.

As a first step to investigating the reliability of module-specific scores, we did examine what our results would have been under random modules (i.e., a random collection of items in each module), since it should first be confirmed that a module (or subtopic) contains a distinguishable set of items worthy of a "subscore." With this goal in mind, we were interested in how our main results would potentially differ if modules were comprised of random collections of items chosen from the entire curriculum, which would essentially be devoid of any such subtopic quality. For this, we ran multiple simulations creating random modules and performed analogous analyses to our main analyses outlined in Tables 2–5 for module-specific ability. Our results are detailed in Appendix C. We found that rankings on random modules produced results on the three summary statistics (range, maximum, and correlation) that lie somewhere between overall ability across time and module-specific ability and that the differences from the true module-specific ability statistics (as well as overall ability across time) were statistically significant. These results were expected, namely, that they approach the results of overall ability across time since a random module could be thought of as a representative sample of the entire (overall) curriculum. Ultimately, these results exhibit strong evidence that the actual teacher-created modules were inherently different than those that were created randomly. Nonetheless, this analysis suggests that subscores that differ from overall scores can still be deceptive, after all, our results showed that rankings on random modules still produced distinct results from overall ability across time. In other words, item selection in subcategories (not just *length*) is paramount when considering subscores.

Discussion

Research summary

The present study examined how students rank in their math class under three ways of measuring different constructs of mathematical ability. We found that module-specific ability (i.e., judging student ability on various teacher-created modules) exhibited the most variation in student rankings. Moreover, scores obtained from this metric showed that students tended to possess a relative strength on some subtopic of the curriculum, even among those who may be traditionally labeled as low-ability. However, we also saw that student module rankings still had a strong correlation, a result consistent with Spearman's theory of general intelligence. These results combined suggest that both unidimensional and multidimensional notions of ability may co-exist, a finding consistent with Carroll's (1993) investigations using the Cattell-Horn-Carroll (CHC) model.

Interpretation and future work

The objective in ALEKS is to learn all the items that make up the full curriculum of a course. To capture this for both students and teachers, the platform uses a universally known concept of filling up one's pie until it is completely filled. Therefore, it is natural for teachers to monitor students' progress with the number of items learned or the pie percentage of items learned in the course over time. This is precisely our second metric, overall ability across time. While this metric gives a quick snapshot of a student's overall knowledge of the course, it provides little to no insight to students' strengths and weaknesses on specific skills or modules. Thus, this kind of metric is limited in its use of informing data-driven practices on a regular occurrence that seeks to personalize instruction based on students' evolving needs. Moreover, an overall course measure does not provide any insight into what extent students exhibit strengths in their math class. The results of the study indicate that students' rankings are fairly consistent throughout the course. Nevertheless, a vast majority of students display a relative strength on at least one module of the course. Having this awareness, and more importantly knowing when this happens is principally vital in achieving a personalized educational experience for students.

A common pedagogical practice for fostering personalization is grouping students for the purpose of delivering small group instruction tailored to the group's specific needs. Our findings suggest that how teachers group students together could look very different depending on whether a unidimensional or a multidimensional metric of ability is applied, a finding consistent with Lechuga and Doroudi (2022). If the goal is to form groups of like ability so that adequate scaffolding and appropriate instruction may be administered to a particular group, it may be more suitable to form flexible groupings under a multidimensional view of ability. For example, if a teacher wishes to form groups at the start of a module based on abilities associated with that module, an overall course score may misplace students in the incorrect group where they could potentially receive unnecessary scaffolding or perhaps not enough scaffolding. As we saw in our analyses, even those who may be traditionally labeled as low-ability tend to exhibit a strength in a subtopic of the curriculum.

In addition to potentially receiving inappropriate instruction, students who are misplaced in ability groups may also lose the opportunity of collaborating with other classmates who are generally perceived to have a different level of ability. Indeed, the results of the study suggest a practical use for generating flexible groupings that potentially allow students to be regrouped with peers of like ability on a specific subtopic of the curriculum. We propose that these module-specific scores (or rankings) may be used as a *proactive formative assessment tool* that can identify student strengths or needs for the purpose of tailoring instruction according to the group's ability level *before* a lesson is introduced. This deviates from the classical formative assessment approach, which is typically *reactive* in response to student performance where feedback or targeted interventions occur after a specific concept or lesson has been covered and after the assessment has been administered.

Nevertheless, the evidence suggested by our findings, including the interpretation and potential instructional uses are heavily reliant on the validity of the subscores generated by module-specific ability. While we provide ample evidence and references for the overall validity of ALEKS scores (including subscores) in a subsequent section in this discussion, we are mindful that we did not examine the validity of subscores generated in our specific application (i.e., rankings on module-specific ability). In pursuit of exhibiting that our subscores adequately estimate true subtopic ability, future work could directly evaluate module-specific rankings by comparing them to other external measures or performance in a yet-to-be-administered module. Specifically, it is possible to examine the incremental validity of subscores beyond the overall score by investigating which more accurately predicts proficiency on an external assessment (Biancarosa et al., 2019). Additionally, this sort of investigation may provide insight into when (during the school year) and/or for what subtopics of the curriculum might the overall score be more appropriate than subscores, and vice versa, when estimating the true ability on a particular subtopic (or module) of the curriculum. In any event, the analyses of the present study tell us the prevalence of students exhibiting strengths in their math class when using ALEKS, which was found to be quite common. Therefore, these results demonstrate the

need for a multidimensional metric, such as module-specific ability, to effectively identify students' strengths and weaknesses. Additionally, this data could potentially enable teachers to proactively personalize instruction according to students' needs.

Practical implications

We propose that our findings could potentially challenge many teachers' beliefs around the nature of intelligence, which could in turn influence student learningeven if teachers do not use ability to group students. For example, teachers who hold a fixed view of intelligence tend to see present ability as one's underlying potential, which is often accompanied with low expectations and underestimating ability for those perceived as having low ability (Lee, 1996). It is also well documented that students' self-concepts are heavily influenced by teachers' implicit theories of intelligence (Blackwell et al., 2007; Canning et al., 2019; Lee, 1996; Muenks et al., 2020). Thus, there is the potential to boost students' self-concepts if teachers are successful at harnessing a malleable view of intelligence for identifying and publicly recognizing students' strengths. Indeed, studies have shown that students develop more positive identities in mathematics when they have teachers who recognize and value their strengths (I. S. Horn, 2017; E. N. Walker, 2012). Prior work has focused on ways teachers might accomplish this feat such as through *Complex Instruction*, an approach to instruction for creating equitable classrooms where teachers make deliberate efforts to recognize the strengths of students (primarily those of lower academic status) in the context of cooperative learning (Cohen et al., 1999). Relatedly and more recently, researchers have studied the art of equitable *teacher noticing* (specifically in mathematics), which among other things, seeks to challenge ideologies that position marginalized students as mathematically deficient by recognizing such students as sense-makers who possess unique strengths and ideas that can support future learning (Louie, 2018).

Nonetheless, researchers have suggested that naming and recognizing students' strengths, especially in mathematics, can be quite difficult due to the influence of deficit-based thinking rooted in mathematics education and practices such as tracking (Aguirre et al., 2013; Cohen et al., 1999; I. S. Horn, 2017). Thus, there may be a benefit to presenting teachers with multidimensional rankings (as done in the present study) in order to surface student strengths that may otherwise go unnoticed. For example, on learning platforms such as ALEKS, one could imagine dashboards or nudge alerts that inform teachers of students' specific strengths in their math class (even for those who may be perceived as having lower ability). In this, teachers would be supported with a tool, grounded in data, which could help combat dominant ideologies ingrained in mathematics education that tend to hinder teachers' ability to see students' strengths with authenticity. Indeed, Louie (2018) recounts a teacher's struggle in constantly doubting the legitimacy in the "smartness" of her students despite being fully committed to noticing and naming her students' mathematical strengths. Hence, a tool grounded in data, such as the one suggested, could potentially eliminate doubts even among the most well-intentioned and committed teachers who wish to highlight their students' strengths. These hypotheses could potentially be tested in future work that combines such a tool with theoretical frameworks borrowed from Complex Instruction and equitable teacher noticing.

A case for validity

As suggested, one interpretation of our results could be that students typically possess relative strengths within the curriculum and thus have the potential to be meaningful contributors in their math class. While this interpretation is encouraging, we recognize this is heavily reliant on the reliability and validity of our data. While the validity of our subscores is not directly examined in the present work, we think it is still valuable to discuss the validity of ALEKS scores through the lens of previous works and other contextually relevant information about ALEKS. Such examples make a compelling case for the presence of validity in the present work and the potential of evidencing this directly in a future study.

First, we wish to underscore an assertion made by Falmagne et al. (2007), which touches upon the "burden of validation" for the ALEKS assessment. The idea is that assessments based on KST such as ALEKS present a fundamentally different situation than most assessments, which typically aim to represent the degree of competence of an academic subject by a normalized numerical score. Thus, because there is no obvious connection between the numerical score and whether the examinee is able to solve a particular problem, validation is paramount. However, conversely, the collection of all items potentially used in an ALEKS assessment, by design, covers the full curriculum and results indicate whether a specific skill in the curriculum is known by the student.[7] Because of this principal difference, Falmagne and colleagues posit the plausibility that the measurement of reliability is indistinguishable from that of validity, provided the item set is an adequate representation of the full curriculum. Indeed, SEPT highlights this as one form of evidence of validity and proposes that this can be demonstrated by subject matter experts inspecting the alignment of whether the test content appropriately samples curriculum standards. While teachers ultimately determine their ALEKS course content and modules, it should be noted that ALEKS courses are automatically aligned to the Common Core State Standards as well as *all* 50 U.S. states' standards (K-12 Standards, 2023). These alignments are produced by subject matter experts, many of whom are former teachers. Additionally, because an ALEKS assessment draws from the full curriculum and makes inferences at the problem (or skill) level for all problems in the curriculum, construct underrepresentation[8] (as well as construct-irrelevance) is less of a concern in ALEKS. Another form of evidence of validity noted by SEPT deals with *response processes*. We emphasize that most ALEKS items avoid multiple choice and require an open-ended response using input tools that would mimic what would be done with paper and pencil (Cosyn et al., 2021). On the other hand, we recognize that such input tools that support these kinds of student responses do not come without their potential threats to validity as they may require digital skills that go beyond mathematical ability. While ALEKS does support learners with content that is designed to bring innovation in instructional design and learning, such content is excluded from ALEKS assessments and are only experienced during practice and in some cases in re-assessments only *after* the student has displayed familiarity with the tool. In other words, ALEKS aims to strike a balance in its free response input tools without introducing multimedia or interactive features that do not closely mimic paper and pencil experiences. Apart from avoiding lucky guesses, which could potentially negatively affect the reliability and thus the validity of the ALEKS assessment, we submit that the effort to incorporate input tools is substantiated as many state curriculum standards that require students to actively demonstrate a mathematical skill such as "*write/create* an equation/inequality," "*sketch* graphs of functions," "*draw* polygons," or "*create* appropriate displays for numerical data" to name a few.

Perhaps the most common form of evidence of validity is through evaluating test–criterion relationships (AERA, APA, & NCME, 2014); that is, how well test scores predict outcomes that are operationally distinct from the test (e.g., SAT/ACT scores predicting grades in entry-level college courses). In their evaluation of the ALEKS assessment for placing incoming students in a math course at the University of Illinois, Ahlgren Reddy and Harper (2013) examined the underlying hypothesis that the result of the ALEKS assessment is indicative of student performance in their placed math course. The study showed a strong link between ALEKS assessment scores and student grades in their entry-level math course. This relationship was found to be an improvement over the school's former placement exam, the ACT. More recently, in a study conducted at an HBCU, Ayele et al. (2023) found that the ALEKS assessment performed comparatively well as the SAT when predicting entry-level course grades. Notably, ALEKS was a better predictor than the SAT of student performance in the

[7]The accuracy of making classifications on whether a student knows or does not know an item post assessment is addressed in Falmagne et al. (2013) and Cosyn et al. (2021).

[8]Construct underrepresentation refers to the degree to which a test falls short in measuring the construct being examined. Construct-irrelevance refers to the degree that test scores are impacted by unrelated factors (AERA, APA, & NCME, 2014).

Elementary Algebra course (a developmental college course similar to a secondary Pre-Algebra and Algebra 1 course). Consequently, it was hypothesized that the SAT did not adequately capture the ability of lower-performing students in mathematics in a way ALEKS may have. Relevant to the metric of overall ability across time, Ayele et al. (2023) also found that practice in the ALEKS learning system was associated with a 38% chance of placing in a higher course when students retook the ALEKS assessment. Pertaining to subscores (and modules-specific ability), Ahlgren Reddy and Harper (2013) found that performance on subtopics of the ALEKS assessment were correlated with outcomes, especially in courses where the subcategory is foundational to the recommended course. For example, performance on the subtopics equations/inequalities, rational/radical expressions, and exponents and polynomials correlated well to eventual course grades in Business Calculus and Calculus.

Final thoughts

One limitation of the present study is that it did not take into account student usage time in the ALEKS system. Because we re-rank students at different times during the course, it is reasonable to suspect that rankings would be affected by usage. In particular, those who spent more time in the system would likely rank higher and those who spent less time would likely rank lower. While this is a legitimate concern (especially for overall ability), at the same time this concern would be less of an issue for module-specific ability since rankings were formed *before* the start of the module. This means that the knowledge states used to compute rankings are only informed by assessments and student work on *prior* modules, not the current module. Prior modules may contain prerequisite items for upcoming modules, but they would also contain many unrelated items that have no direct influence on the rankings on subsequent modules. On the other hand, even if time usage had a meaningful influence in the rankings, we contend that rankings that are influenced by how long students spend on the platform are still meaningful (and perhaps preferred) as they would potentially capture student effort, which is an important factor to consider when decisions are made based on ability.

Lastly, though it was beyond the scope of the study, we should note that information regarding student demographics and information about the school and class (including adopted practices) was not available. One purpose of our study was to see what ability rankings look like for students who may be traditionally perceived as having low mathematical ability. Therefore, a sensible choice for identifying such students with the available data was to look at students' initial assessment (i.e., those scoring in the lower quartile of the initial assessment). We recognize, however, that perceptions about student ability often go beyond initial performance or performance in general. In particular, information about the student (e.g., gender, race, and socioeconomic status, etc.) as well as school and class information could potentially offer further insights about students who may be perceived as low-ability for other reasons. Additionally, we recognized that the attention given to perceived low-ability students could be expanded to those perceived as having high ability. For example, future work may focus its attention on situations where higher performing students rank lower on specific modules. Investigations of this sort could potentially reveal misconceptions or learning gaps on certain subtopics of the curriculum, benefiting both lower- and higher-ability students. Indeed, this suggestion was recognized by Walker and Beretvas (2003) when examining the misclassification of students into a higher proficiency level under a unidimensional IRT model versus a multidimensional IRT model.

Despite these limitations, we believe the present study may inform future investigations regarding the measurement of student ability with a focus toward fine-tuning education practices and updating perceptions about student ability. Using new methodological approaches to analyzing fine-grained student data (whether from adaptive learning platforms like ALEKS, formative assessments, or multidimensional standardized tests) may give new insights into the nuanced ways in which student ability varies across time and content in various educational settings.

Disclosure statement

The authors Christopher G. Lechuga and Jeffrey Matayoshi are full-time employees at McGraw Hill ALEKS. Author Shayan Doroudi declares no potential or perceived conflicts of interest. The data used for this manuscript was collected, anonymized, and provided by McGraw Hill ALEKS. None of the authors, however, were involved in the collection or the anonymization of the data used for this study.

References

About ALEKS. McGraw Hill ALEKS. Retrieved April 30, 2023, from https://www.aleks.com/about_aleks

ACT. (2012). *ACT compass internet version reference manual*. Retrieved December 20, 2023, from https://act-stage.adobecqms.net/content/dam/act/unsecured/documents/CompassReferenceManual.pdf

ACT. (2022). *ACT technical manual*. Retrieved January 7, 2024, from https://www.act.org/content/dam/act/unsecured/documents/ACT_Technical_Manual.pdf

Aguirre, J., Mayfeld-Ingram, K., & Martin, D. (2013). *The impact of identity in K-8 mathematics: Rethinking equity-based practices*. National Council of Teachers of Mathematics.

Ahlgren Reddy, A., & Harper, M. (2013). Mathematics placement at the University of Illinois. *Primus*, *23*(8), 683–702. https://doi.org/10.1080/10511970.2013.801378

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Ayele, A. D., Carson, Z., & Tameze, C. (2023). An efficacy study of ALEKS-Based placement in entry-level college math courses. *Primus*, *33*(4) 414–430 https://doi.org/10.1080/10511970.2022.2073623.

Biancarosa, G., Kennedy, P. C., Carlson, S. E., Yoon, H., Seipel, B., Liu, B., & Davison, M. L. (2019). Constructing subscores that add validity: A case study of identifying students at risk. *Educational and Psychological Measurement*, *79*(1), 65–84. https://doi.org/10.1177/0013164418763255

Bill & Melinda Gates Foundation. (2015). *Teachers know best: Making data work for teachers and students*. Retrieved November 21, 2023, from https://files.eric.ed.gov/fulltext/ED557084.pdf

Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development*, *78*(1), 246–263. https://doi.org/10.1111/j.1467-8624.2007.00995.x

Boaler, J. (2005, September). The 'psychological prisons' from which they never escaped: The role of ability grouping in reproducing social class inequalities. *The Forum*, *47*(2), 125–134. https://doi.org/10.2304/forum.2005.47.2.2

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement*, *4*(1), 87–100.

Canning, E. A., Muenks, K., Green, D. J., & Murphy, M. C. (2019). STEM faculty who believe ability is fixed have larger racial achievement gaps and inspire less student motivation in their classes. *Science Advances*, *5*(2), eaau4734. https://doi.org/10.1126/sciadv.aau4734

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press. https://doi.org/10.1017/CBO9780511571312

Castle, S., Deniz, C. B., & Tortora, M. (2005). Flexible grouping and student learning in a high-needs school. *Education & Urban Society*, *37*(2), 139–150. https://doi.org/10.1177/0013124504270787

Cipriano-Walter, M. (2015). Falling off the track: How ability tracking leads to intra-school segregation. *Thurgood Marshall Law Review*, *41*, 25.

Cohen, E. G., Lotan, R. A., Scarloss, B. A., & Arellano, A. R. (1999). Complex instruction: Equity in cooperative learning classrooms. *Theory into Practice*, *38*(2), 80–86. https://doi.org/10.1080/00405849909543836

College Board. (2023). *Assessment framework for the digital SAT suite*. Retrieved January 7, 2024, from https://satsuite.collegeboard.org/media/pdf/assessment-framework-for-digital-sat-suite.pdf

Cosyn, E., Uzun, H., Doble, C., & Matayoshi, J. (2021). A practical perspective on knowledge space theory: ALEKS and its data. *Journal of Mathematical Psychology*, *101*, 102512. https://doi.org/10.1016/j.jmp.2021.102512

Dawson, M. M. (1987). Beyond ability grouping: A review of the effectiveness of ability grouping and its alternatives. *School Psychology Review*, *16*(3), 348–369. https://doi.org/10.1080/02796015.1987.12085298

Denning, J. T., Murphy, R., & Weinhardt, F. (2018). Class rank and long-run outcomes. *The Review of Economics and Statistics*, 1–45. https://doi.org/10.1162/rest_a_01125

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*(2), 258–282. https://doi.org/10.1007/s40593-019-00176-0

Doignon, J. P., & Falmagne, J. C. (1985). Spaces for the assessment of knowledge. *International Journal of Man-Machine Studies*, *23*(2), 175–196. https://doi.org/10.1016/S0020-7373(85)80031-6

Donovan, D. A., Connell, G. L., Grunspan, D. Z., & Wilson, K. J. (2018). Student learning outcomes and attitudes using three methods of group formation in a nonmajors biology class. *CBE - Life Sciences Education*, *17*(4), 60. https://doi.org/10.1187/cbe.17-12-0283

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*(397), 171–185. https://doi.org/10.1080/01621459.1987.10478410

Elsner, B., Isphording, I. E., & Zölitz, U. (2021). Achievement rank affects performance and major choices in college. *The Economic Journal*, *131*(640), 3182–3206. https://doi.org/10.1093/ej/ueab034

ETS. (2023). *Technical manual for the praxis tests and related assessments*. Retrieved January 7, 2024, from https://www.ets.org/pdfs/praxis/technical-manual.pdf

Falmagne, J. C., Albert, D., Doble, C., Eppstein, D., & Hu, X. (Eds.). (2013). *Knowledge spaces: Applications in education*. Springer Science & Business Media. https://doi.org/10.1007/978-3-642-35329-1

Falmagne, J. C., Cosyn, E., Doble, C., Thiéry, N., & Uzun, H. (2007). Assessing mathematical knowledge in a learning space: Validity and/or reliability. *Annual Meeting of the American Educational Research Association (AERA)*, (949). Retrieved Apirl 1, 2025, from https://www.stat.cmu.edu/~brian/NCME07/Validity_in_L_Spaces.pdf

Field, C. A., & Welsh, A. H. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(3), 369–390. https://doi.org/10.1111/j.1467-9868.2007.00593.x

Finley, M. K. (1984). Teachers and tracking in a comprehensive high school. *Sociology of Education*, *57*(4), 233–243. https://doi.org/10.2307/2112427

Gallardo, E. V. (1994). Hierarchy and discrimination: Tracking in public schools. *Chicana/O Latina/O Law Review*, *15*(1), 74. https://doi.org/10.5070/C7151021048

Gardner, H. E. (2011). *Frames of mind: The theory of multiple intelligences*. Basic books.

Goulas, S., & Megalokonomou, R. (2021). Knowing who you actually are: The effect of feedback on short-and longer-term outcomes. *Journal of Economic Behavior and Organization*, *183*, 589–615. https://doi.org/10.1016/j.jebo.2021.01.013

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*, *21*(4), 347–360. https://doi.org/10.1111/j.1745-3984.1984.tb01039.x

Gross, B., Tuchman, S., & Patrick, S. (2018). A national landscape scan of personalized learning in K-12 education in the United States. *iNACOL*. Retrieved April 1, 2025, from https://eric.ed.gov/?id=ED589851.

Guilford, J. P. (1982). Cognitive psychology's ambiguities: Some suggested remedies. *Psychological Review*, *89*(1), 48–59. https://doi.org/10.1037/0033-295X.89.1.48

Heinrich, C. J., Darling-Aduana, J., Good, A. G. (2020). *Equity and quality in digital Learning: Realizing the promise in K-12 education*. Harvard Education Press.

Hoover, N. R., & Abrams, L. M. (2013). Teachers' instructional use of summative student assessment data. *Applied Measurement in Education*, *26*(3), 219–231. https://doi.org/10.1080/08957347.2013.793187

Horn, I. S. (2017). *Motivated: Designing math classrooms where students want to join in*. Heinemann.

Horn, J. L. (1967). Intelligence-why it grows, why it declines. *Trans-Action*, *5*(1), 23–31. https://doi.org/10.1007/BF03180091

Horn, J. L. (1968). Organization of abilities and the development of intelligence. *Psychological Review*, *75*(3), 242. https://doi.org/10.1037/h0025662

K-12 Standards. McGraw Hill ALEKS. Retrieved December 30, 2023, from https://www.aleks.com/k12/standards

Kan, A., Bulut, O., & Cormier, D. C. (2019). The impact of item stem format on the dimensional structure of mathematics assessments. *Educational Assessment*, *24*(1), 13–32. https://doi.org/10.1080/10627197.2018.1545569

Kang, C., Liu, N., Zhu, Y., Li, F., & Zeng, P. (2022). Developing college students' computational thinking multi-dimensional test based on life story situations. *Education and Information Technologies*, 1–19. https://doi.org/10.1007/s10639-022-11189-z

Kanika, C., Chakraborty, S., Chakraborty, P, M., & Chakraborty, P. (2023). Effect of different grouping arrangements on students' achievement and experience in collaborative learning environment. *Interactive Learning Environments*, *31*(10), 6366–6378. https://doi.org/10.1080/10494820.2022.2036764

Kelly, S. (2004). Are teachers tracked? On what basis and with what consequences. *Social Psychology of Education*, *7*(1), 55–72. https://doi.org/10.1023/B:SPOE.0000010673.78910.f1

Kulik, C. L. C., & Kulik, J. A. (1982). Effects of ability grouping on secondary school students: A meta-analysis of evaluation findings. *American Educational Research Journal*, *19*(3), 415–428. https://doi.org/10.3102/00028312019003415

Kulik, C. L. C., & Kulik, J. A. (1984, August). Effects of ability grouping on elementary school pupils: A meta-analysis. Paper presented at the Annual Meeting of the American Psychological Association (92nd, Toronto, Ontario, Canada, August 24-28, 1984). (ERIC Document Reproduction Service No. Ed 255 329)

Kulik, J. A., & Kulik, C. L. C. (1992). Meta-analytic findings on grouping programs. *The Gifted Child Quarterly*, *36*(2), 73–77. https://doi.org/10.1177/001698629203600204

Lechuga, C. G., & Doroudi, S. (2022). Three algorithms for grouping students: A bridge between personalized tutoring system data and classroom pedagogy. *International Journal of Artificial Intelligence in Education*, 1–42. https://doi.org/10.1007/s40593-022-00309-y

Lee, K. (1996). A study of teacher responses based on their conceptions of intelligence. *The Journal of Classroom Interaction*, 1–12. https://www.jstor.org/stable/23870415

Lleras, C., & Rangel, C. (2009). Ability grouping practices in elementary school and African American/Hispanic achievement. *American Journal of Education*, *115*(2), 279–304. https://doi.org/10.1086/595667

Louie, N. L. (2018). Culture and ideology in mathematics teacher noticing. *Educational Studies in Mathematics*, *97*, 55–69. https://doi.org/10.1007/s10649-017-9775-2

MacIntyre, H., & Ireson, J. (2002). Within-class ability grouping: Placement of pupils in groups and self-concept. *British Educational Research Journal*, *28*(2), 249–263. https://doi.org/10.1080/01411920120122176

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, *37*(1), 1–10. https://doi.org/10.1016/j.intell.2008.08.004

McMullen, J., Lewis, R. W., & Bailey, D. H. (2020). Latent classes from complex assessments: What do they tell us? *Learning & Individual Differences*, *83*, 101944. https://doi.org/10.1016/j.lindif.2020.101944

Meissel, K., Meyer, F., Yao, E. S., & Rubie-Davies, C. M. (2017). Subjectivity of teacher judgments: Exploring student characteristics that influence teacher judgments of student ability. *Teaching & Teacher Education*, *65*, 48–60. https://doi.org/10.1016/j.tate.2017.02.021

Mignani, S., Monari, P., Cagnone, S., & Ricci, R. (2006). Multidimensional versus unidimensional models for ability testing. In S. Zani, A. Cerioli, M. Riani, & M. Vichi (Eds.), *Data analysis, classification and the forward search. Studies in classification, data analysis, and knowledge organization*. Springer. https://doi.org/10.1007/3-540-35978-8_38

Missett, T. C., Brunner, M. M., Callahan, C. M., Moon, T. R., & Price Azano, A. (2014). Exploring teacher beliefs and use of acceleration, ability grouping, and formative assessment. *Journal for the Education of the Gifted*, *37*(3), 245–268. https://doi.org/10.1177/0162353214541326

Muenks, K., Canning, E. A., LaCosse, J., Green, D. J., Zirkel, S., Garcia, J. A., & Murphy, M. C. (2020). Does my professor think my ability can change? students' perceptions of their STEM professors' mindset beliefs predict their psychological vulnerability, engagement, and performance in class. *Journal of Experimental Psychology General*, *149*(11), 2119. https://doi.org/10.1037/xge0000763

Murphy, R., & Weinhardt, F. (2020). Top of the class: The importance of ordinal rank. *Review of Economic Studies*, *87*(6), 2777–2826. https://doi.org/10.1093/restud/rdaa020

NAEP. (2023). *Technical documentation: NAEP assessment IRT parameters*. Retrieved January 7, 2024, from https://nces.ed.gov/nationsreportcard/tdw/analysis/scaling_irt.aspx

Nicewander, W. A., & Thomasson, G. L. (1999). Some reliability estimates for computerized adaptive tests. *Applied Psychological Measurement*, *23*(3), 239–247. https://doi.org/10.1177/01466219922031356

Oakes, J. (2005). *Keeping track: How schools structure inequality*. Yale University Press.

Rosenbaum, J. E. (1980). Chapter, 8: Social implications of educational grouping. *Review of Research in Education*, *8*(1), 361–401. https://doi.org/10.3102/0091732X008001361

Rosenholtz, S. J., & Simpson, C. (1984a). Classroom organization and student stratification. *Elementary School Journal*, *85*(1), 21–37. https://doi.org/10.1086/461389

Rosenholtz, S. J., & Simpson, C. (1984b). The formation of ability conceptions: Developmental trend or social construction? *Review of Educational Research*, *54*(1), 31–63. https://doi.org/10.3102/00346543054001031

Rosenholtz, S. J., & Wilson, B. (1980). The effect of classroom structure on shared perceptions of ability. *American Educational Research Journal*, *17*(1), 75–82. https://doi.org/10.3102/00028312017001075

Rowan, B., & Miracle, A. W., Jr. (1983). Systems of ability grouping and the stratification of achievement in elementary schools. *Sociology of Education*, *56*(3), 133–144. https://doi.org/10.2307/2112382

Sanz-Martínez, L., Er, E., Martínez-Monés, A., Dimitriadis, Y., & Bote-Lorenzo, M. L. (2019). Creating collaborative groups in a MOOC: A homogeneous engagement grouping approach. *Behaviour & Information Technology*, *38*(11), 1107–1121. https://doi.org/10.1080/0144929X.2019.1571109

Sheng, Y., & Wikle, C. K. (2007). Comparing multiunidimensional and unidimensional item response theory models. *Educational and Psychological Measurement*, *67*(6), 899–919. https://doi.org/10.1177/0013164406296977

Simpson, C. (1981). Classroom structure and the organization of ability. *Sociology of Education*, 120–132. https://doi.org/10.2307/2112356

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, *47*(2), 150–174. https://doi.org/10.1111/j.1745-3984.2010.00106.x

Sinharay, S., Haberman, S., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement Issues & Practice*, *26*(4), 21–28. https://doi.org/10.1111/j.1745-3992.2007.00105.x

Sinharay, S., Puhan, G., Haberman, S. J., & Hambleton, R. K. (2018). Subscores: When to communicate them, what are their alternatives, and some recommendations. Edited by Zapata-Rivera, Diego. In *Score reporting research and applications* (pp. 35–49). Routledge.

Slavin, R. E. (1987). Ability grouping and student achievement in elementary schools: A best-evidence synthesis. *Review of Educational Research*, *57*(3), 293–336. https://doi.org/10.3102/00346543057003293

Slavin, R. E. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research*, *60*(3), 471–499. https://doi.org/10.3102/00346543060003471

Smarter Balanced, (2022). *2020-21 Summative technical report: Chapter 5 scores, scales, and norms*. Retrieved January 7, 2024, from https://technicalreports.smarterbalanced.org/2020-21_summative-report/_book/scores-scales-norms.html

Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201–292. https://doi.org/10.2307/1412107

Sternberg, R. J. (1984). Toward a triarchic theory of human intelligence. *Behavioral and Brain Sciences*, *7*(2), 269–287. https://doi.org/10.1017/S0140525X00044629

Sternberg, R. J. (1996). Myths, countermyths, and truths about intelligence. *Educational Researcher*, *25*(2), 11–16. https://doi.org/10.3102/0013189X025002011

Sternberg, R. J. (2010). Intelligence. In P. Peterson, E. Baker, & B. McGraw Eds. *International encyclopedia of education* (3rd ed. pp. 184–190). Elsevier. https://doi.org/10.1016/B978-0-08-044894-7.00482-6

Thissen, D. (2000). Reliability and measurement precision. In *Computerized adaptive testing* (pp. 159–184). Routledge.

Tieso, C. L. (2003). Ability grouping is not just tracking anymore. *Roeper Review*, *26*(1), 29–36. https://doi.org/10.1080/02783190309554236

Trimble, K. D., & Sinclair, R. L. (1987). On the wrong track: Ability grouping and the threat to equity. *Equity & Excellence in Education*, *23*(1–2), 15–21. https://doi.org/10.1080/1066568870230104

Walker, C. M., & Beretvas, S. N. (2000). Using multidimensional versus unidimensional ability estimates to determine Student proficiency in mathematics. Paper presented at the 2000 Annual Meeting of the American Educational Research Association, New Orleans, LA

Walker, C. M., & Beretvas, S. N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement*, *40*(3), 255–275. https://doi.org/10.1111/j.1745-3984.2003.tb01107.x

Walker, E. N. (2012). Cultivating mathematics identities in and out of school and in between. *Journal of Urban Mathematics Education*, *5*(1), 66–83. https://doi.org/10.21423/jume-v5i1a173

Warne, R. T. (2016). Five reasons to put the *g* back into giftedness: An argument for applying the cattell-horn-carroll theory of intelligence to gifted education research and practice. *The Gifted Child Quarterly*, *60*(1), 3–15. https://doi.org/10.1177/0016986215605360

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. https://doi.org/10.2458/v2i1.12351

Appendix

Appendix A

McGraw Hill ALEKS provided data for 191 classes containing modules with due dates occurring between 2017 and 2019. The level of the classes consisted of middle school math (grades 6-8) and Algebra 1. From this pool, several attributes (listed below) were considered for the class being included in the study.

Attributes:

(i) Number of active students at the start of the class
(ii) Number of modules with at least 10 items
(iii) Average number of items learned
(iv) Non-overlapping due dates

With regard to (i), because the study ranks students and interprets these rankings similar to percentiles, the class sizes must be large enough for meaningful interpretation. Also, because students were sometimes added mid-year, we considered the number of active students at the start of the class. So, we required at least 10 active students at the start of the class. This requirement happened to lie in the 32nd percentile among the 191 classes. With regard to (ii), a threshold of 10 items was chosen because there needed to be an adequate number of items per module for producing meaningful rankings of students. Otherwise, if the number of items is too small, ability levels and thus rankings would be more difficult to differentiate. The Section 'Reliability' also speaks to the appropriateness of 10 items per module. While the literature on subscore reliability recommends a length of 20 items, we note that our main study results, as well as our reliability results (presented in Section 'Results'), changed very little on a smaller sample of data consisting of modules of at least 20 items. For this reason, we elected for setting this threshold at 10 items to include more data in our sample.

When considering the number of modules for a class, we also desired a meaningful number of subdivisions of the entire curriculum. Since a typical ALEKS class consists of hundreds of items, having too few modules might not express a meaningful subtopic and might include several subtopics where there would be no chance of observing relative strengths exhibited in student ability. We note that many market-leading textbooks for this level of mathematics typically subdivide the curriculum into 10–12 chapters. Although, we thought requiring 10+ modules would be a bit too rigid, as many state frameworks, including the Common Core State Standards for Mathematics, divide their state-standards into five domains (or strands). Thus, for (ii) we required that a class have at least five modules, which happened to lie in 25th percentile for the 191 classes.

With regard to (iii), because we were interested in ability over time (our second metric detailed in Section 'Measures of Ability Used'), we desired classes with student activity exhibited by students learning items in the system. Otherwise, if there was no activity and no learning in the system, ability over time would be no different than ability measured at the start of the class (which is our first metric detailed in Section 'Measures of Ability Used'). That said, because it is unknown how many items should be learned in order to see meaningful differences in ability over time, the amount of data also played a role in determining the threshold for (iii). We noticed that the 25th percentile for the average number of items learned was 26.73. We considered this a reasonable choice for two reasons: (a) this did not exclude too much of the received data, and (b) it required classes to have some learning activity (albeit this was a conservative choice as the number of items required was relatively low compared to the total number of items typically observed in an ALEKS class). Finally, with regard to (iv), in order to delimit when modules start and end, a requirement was made that module durations were distinct and not overlapping. Classes satisfying all four criteria were included in the study, resulting in a total of 42 classes and 915 students. A breakdown of classes and number of students by grade/course level is given in Table 6.

Table 6. Breakdown of the Number of Classes and Students for Each Grade/Course Level in Data.

Grade/Course	Number of classes	Number of students
Grade 6	11	224
Grade 7	9	180
Grade 8	16	384
Algebra 1	6	127
Total	42	915

Appendix B

Module-specific ability is measured on a subset of the curriculum. This makes it more likely for multiple students in the same class to have the same measure, especially at the start of the module. Therefore, to distinguish ability on a specific module more effectively, preparedness on the module was taken into consideration. Specifically, there were three components for computing a student's module-specific ability: (a) what the student already knows in the module (i.e., the student's current state intersected with the set of items in the module), (b) the student's preparedness for the module (i.e., the student's current state intersected with the set of prerequisite items of the module), and (c) the student's overall current ability represented by their current state. So, a student's module-specific ability was measured by the *raw module score* (RMS) given by,

$$RMS = |S \cap M| + \left|S \cap M_{prereqs}\right| \cdot 0.001 + |S| \cdot 0.00001,$$

where S is the student's current state, M is the set of items in the module, and $M_{prereqs}$ is the set of prerequisite items for M.[9]

Note, *RMS* is constructed in such a way as to minimize ties. However, while unlikely, there could still be rank ties for module-specific ability. The same is true for initial ability and overall ability across time. Whenever this happened, students who tied were assigned a corrected rank, which was the average of their ordinal position. For example, if two students were tied for seventh, this would mean they occupied the seventh and eighth positions. Thus, they were each given a corrected rank of 7.5. This was then used as the value for r_i in Equation 1 (in Section 'Procedure for Ranking Students') for each student, thus producing an average percentile rank.

Appendix C

Main Analyses Redux

Tables A1–A4 are repeats of Tables 2–5, which display the results of our main analyses. These repeats give the results for a subset of data obtained from modules that have at least 20 items. So, the column labeled "Full dataset" is within the context of the already filtered data consisting of modules of at least 20 items. The column labeled "Lower quartile" refers to data obtained from modules of at least 20 items *and* obtained from students scoring in the lower quartile of the initial assessment.

Table A1. Repeat of Table 2 for Data Consisting of Modules with At Least 20 Items.

	Range	
	Full dataset	Lower quartile
Initial ability	0 NA	0 NA
Overall ability across time	0.196 [0.168, 0.227]	0.203 [0.172, 0.238]
Module-specific ability	0.392 [0.353, 0.430]	0.394 [0.350, 0.443]

Table A2. Repeat of Table 3 for Data Consisting of Modules with At Least 20 Items.

	Maximum	
	Full dataset	Lower quartile
Initial ability	0.5 NA	0.157 [0.129, 0.225]
Overall ability across time	0.599 [0.584, 0.615]	0.372 [0.327, 0.426]
Module-specific ability	0.698 [0.679, 0.719]	0.505 [0.456, 0.566]

[9]By using small coefficients (0.001 and 0.00001), we guarantee that a student's preparedness on a module, $\left|S \cap M_{prereqs}\right|$ and a student's overall current ability, $|S|$, are only used as tie-breakers when two or more students have a tie in one of the preceding terms in the *RMS*.

Random module runs

We performed the analyses outlined in Research Design and Methods for module-specific ability with random modules, while preserving every other aspect of our data and methods. That is, all class and student data were preserved, including the set of items making up the overall curriculum for the class as well as student knowledge states at each moment in time. The only difference in our simulated data was the manipulation of each module for each class. Specifically, from the set of items used in the class, random modules (consisting of the same number of items as the actual modules) were formed by randomly sampling items with replacement. This was repeated 100 times, thus producing 100 sets of module rankings for each student.

Tables A5–A8 are repeats of Tables 2– 5 with an appended row, *module-specific ability (random)*, containing the average statistic among the 100 runs. For example, in Table A5, we see that the average mean range for the 100 runs was 0.356 (CI [0.355, 0.357]) on the full dataset. We notice that this value lies between the means found for overall ability across time and module-specific ability. Indeed, this trend is the same for all statistics across Tables A5–A8, including for the filtered dataset consisting of only those scoring in the bottom quartile of the initial assessment. In retrospect, this is what we should expect considering the nature of random modules. Namely, since modules were formed randomly from the entire curriculum, we should expect that each module is a representative sample of the entire curriculum to some degree. Therefore, ranking a student's ability on a random module is somewhat akin to ranking that student on the entire curriculum (i.e., ranking on overall ability). Yet, the results would seem to indicate that ability on a subset of the curriculum still has a higher degree of dimensionality than overall ability on the entire curriculum. Lastly, as seen by the non-overlapping CIs between the metrics module-specific ability and module-specific ability (random), we observe evidence that these two metrics result in statistically different values on each of the statistics reported across all tables.

Table A3. Repeat of Table 4 for Data Consisting of Modules with At Least 20 Items.

	Proportion with maximum greater than 0.5	
	Full dataset	Lower quartile
Initial ability	0.479 [0.444, 0.491]	0.000 [0.000, 0.000]
Overall ability across time	0.617 [0.593, 0.646]	0.280 [0.203, 0.378]
Module-specific ability	0.762 [0.732, 0.789]	0.486 [0.407, 0.572]

Table A4. Repeat of Table 5 for Data Consisting of Modules with At Least 20 Items.

	Correlation	
	Full dataset	Lower quartile
Initial ability	NA	NA
Overall ability across time	0.889 [0.856, 0.923]	0.803 [0.758, 0.849]
Module-specific ability	0.699 [0.657, 0.741]	0.444 [0.366, 0.521]

Table A5. Repeat of Table 2 with Appended Row Giving the Average Mean Range for the 100 Random Module Runs.

	Range	
	Full dataset	Lower quartile
Initial ability	NA	NA
Overall ability across time	0.222 [0.193, 0.250]	0.230 [0.197, 0.262]
Module-specific ability	0.452 [0.417, 0.485]	0.453 [0.411, 0.497]
Module-specific ability (random)	0.356 [0.355, 0.357]	0.347 [0.345, 0.348]

Table A6. Repeat of Table 3 with Appended Row Giving the Average of the Mean Maximum Ranking for the 100 Random Module Runs.

	Maximum	
	Full dataset	Lower quartile
Initial ability	0.5 NA	0.157 [0.129, 0.225]
Overall ability across time	0.612 [0.597, 0.627]	0.385 [0.340, 0.437]
Module-specific ability	0.729 [0.711, 0.748]	0.545 [0.499, 0.603]
Module-specific ability (random)	0.678 [0.677, 0.678]	0.471 [0.470, 0.473]

Table A7. Repeat of Table 4 with Appended Row Giving the Average of the Mean Proportion of Students with a Maximum Rank Greater than 0.5 for the 100 Random Module Runs.

	Proportion with maximum rank greater than 0.5	
	Full dataset	Lower quartile
Initial ability	0.479 [0.444, 0.491]	0.000 [0.000, 0.000]
Overall ability across time	0.636 [0.611, 0.664]	0.291 [0.213, 0.386]
Module-specific ability	0.808 [0.781, 0.835]	0.567 [0.488, 0.650]
Module-specific ability (random)	0.728 [0.727, 0.730]	0.421 [0.417, 0.425]

Table A8. Repeat of Table 5 with Appended Row Giving the Average of the Mean Class Correlation between Rankings for the 100 Random Module Runs.

	Correlation	
	Full dataset	Lower quartile
Initial ability	NA	NA
Overall ability across time	0.886 [0.854, 0.917]	0.795 [0.755, 0.835]
Module-specific ability	0.696 [0.657, 0.735]	0.441 [0.365, 0.517]
Module-specific ability (random)	0.800 [0.799, 0.801]	0.642 [0.640, 0.644]