IXL Universal Math Screener Technical Manual

August 2026

IXL LEARNING 777 Mariners Island Blvd., Suite 600, San Mateo, CA 94404
650-372-4040 | www.ixl.com



© 2026 IXL Learning. All rights reserved.
IXL and IXL Learning are registered trademarks of IXL Learning, Inc. All other trademarks, registered trademarks, and copyrights are the property of their respective owners.
No part of this document may be modified or further distributed without written permission from IXL Learning.

TABLE OF CONTENTS

INTRODUCTION **4**
- Purpose and Use 5
- Content 5
- Computerized-Adaptive Testing 5
- Test Interface 5
- Test Administration 6
  - Item & Test Time 7
  - Test Length 7
  - Test Repetition 8
  - Accessibility 8
  - Reporting 8

CONTENT DEVELOPMENT **11**
- Principled Assessment Design 11
- Content Specifications 13
- Item Development 14

ITEM & SCALE CALIBRATION **15**
- Background 15
- Measurement Model 15
- Item Calibration 16
- Quality Assurance 17
- Computerized-Adaptive Testing Design & Delivery 17

BENCHMARKS & NORMS **19**
- Benchmarking 19
- Norming 19
  - Sample 20
  - Results 20

RELIABILITY & PRECISION **22**
- Background 22
- Internal Consistency 22
- Test-Retest Reliability 23
- Standard Error of Measurement 24
- Decision Accuracy & Consistency 26

VALIDITY **28**
- Background 28
- Evidence Based on Test Content 28
- Evidence Based on Internal Structure 28
  - Unidimensionality 28

Bifactor Analyses..... 29
Confirmatory Factor Analyses..... 30
Differential Item Functioning..... 31
Subject-Matter Expert Review..... 32
Evidence Based on Relations to Other Variables.....33
Concurrent Validity..... 33
Predictive Validity.....33
Classification Accuracy..... 34
REFERENCES..... 37
APPENDIX A: TEST CONSTRAINTS..... 41
APPENDIX B: NORMING TABLES..... 45
APPENDIX C: RELIABILITY BY GRADE AND STRAND..... 57
APPENDIX D: FACTOR LOADINGS.....58

List of Tables

Table 1. Average number of items and time spent by grade..... 8
Table 2. Subject-Matter Expert Sample and National Demographics..... 12
Table 3. Gwet's AC1 for ALDs and items by grade.....13
Table 8. Demographic representation for the norming sample..... 20
Table 9. Classifications and percentile ranks by grade and window.....21
Table 4. Marginal reliability by grade.....23
Table 5. Test-retest reliability between delivery windows by grade.....24
Table 6. Mean and Median standard errors by grade..... 26
Table 7. Classification accuracy and classification consistency by grade..... 27
Table 10. Bifactor analyses results by grade.....30
Table 11. Confirmatory factor analyses results by grade..... 31
Table 12. Concurrent and Predictive analysis by grade..... 34
Table 13. Classification accuracy by grade and delivery window..... 36
Table A1. Kindergarten min and max item delivery by strand and window..... 41
Table A2. First grade min and max item delivery by strand and window.....41
Table A3. Second grade min and max item delivery by strand and window..... 42
Table A4. Third grade min and max item delivery by strand and window..... 42
Table A5. Fourth grade min and max item delivery by strand and window.....42
Table A6. Fifth grade min and max item delivery by strand and window..... 43

Table A7. Sixth grade min and max item delivery by strand and window....43
Table A8. Seventh grade min and max item delivery by strand and window....43
Table A9. Eighth grade min and max item delivery by strand and window....44
Table B1. Beginning-of-Year norming theta values....45
Table B2. Middle-of-Year norming theta values....49
Table B3. End-of-Year norming theta values....53
Table C1. Marginal reliability by grade and strand....57
Table D1. Factor loadings by grade and strand....58

List of Figures

Figure 1. Example item with timer showing....6
Figure 2. Screenshot showing the creation of a screener window....7
Figure 3. Sample administrator report detailing student performance: Levels view....9
Figure 4. Sample administrator report detailing student performance: National Norms view....10
Figure 5. Example item response functions for three items....16
Figure 6. Standard Error of Measurement across the ability spectrum by grade....25

INTRODUCTION

Purpose and Use

The IXL Universal Math Screener is a brief, accurate assessment designed specifically for identifying students from kindergarten through eighth grade who are experiencing moderate or severe math difficulties. It may be administered up to five times per academic year, with performance expectations automatically adjusting for the beginning, middle, and end of the academic year. By scheduling the first administration early in the academic year, educators can promptly identify students with moderate or severe difficulties. Results are available immediately following the administration so that intervention resources can be deployed quickly to the students who need them. Results are also available by strand, providing insight into the specific areas where each student is struggling.

Content

The IXL Universal Math Screener is designed to assess the most critical math skills for success in grade-level math content. The content covers five strands: Numbers & Operations, Algebra & Algebraic Thinking, Fractions, Geometry, and Data & Measurement. These strands reflect the structure and content across states' standards. The number of items a student receives from any strand is determined by the grade level and time of year the test is delivered.

Computerized-Adaptive Testing

A computerized adaptive test (CAT) is an assessment comprising questions that adapt to each student's ability. Item difficulty increases as students respond correctly and decreases as they respond incorrectly until a stable estimate of ability can be calculated. By adapting to each student's ability, a CAT can significantly reduce testing time, increase testing precision, reduce test-taker fatigue, improve test-taker experience, facilitate repeated testing, and improve test security, alongside other benefits.

Test Interface

The test interface invokes a minimalist design to allow students to focus on the presented question with as few distractions as possible. The question resides inside a white box with a dark blue header that sits on top of a light blue background. A small, unobtrusive timer appears in the upper right corner of the item box when a student is close to reaching the item response time limit described in the next section (see Figure 1). Only one question will appear at a time, and backward navigation is disabled so that students are not able to return to previous questions and change

answers. Responses are not forced, and students are able to enter a blank response; however, they are provided a warning to avoid accidental blank submissions.

The IXL Universal Math Screener is supported on most devices and operating systems, such as Windows, Mac OS, and Chrome OS (including Chromebooks). Minimum requirements include an up-to-date version of Chrome, Safari, or Edge. Detailed technical requirements can be found at the following website: https://www.ixl.com/help-center/article/1850087/what-are-the-technical-requirements-for-ixl.

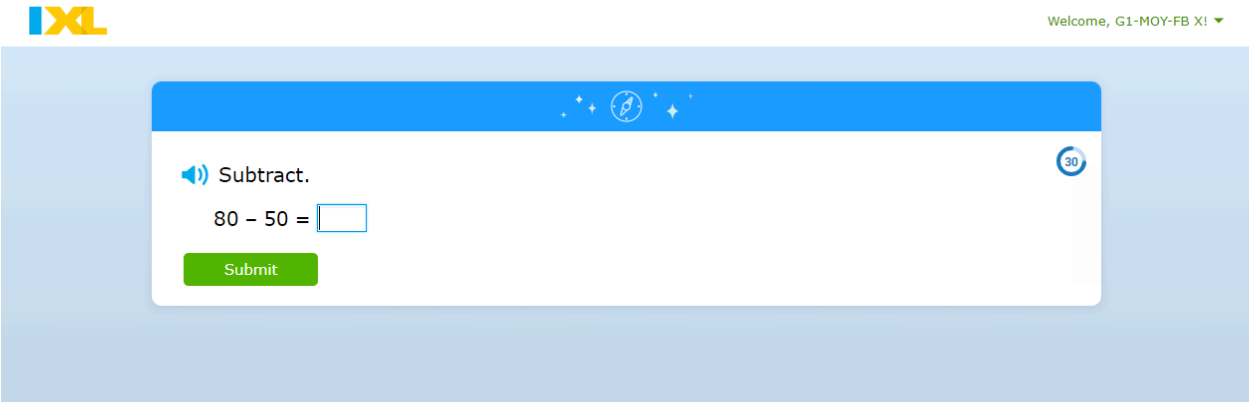


Figure 1. Example item with timer showing

Test Administration

IXL Universal Math Screener administration is simple. Once an administrator has set the screening window, students are prompted to complete the screener automatically when they sign in to IXL. Administrators can monitor student progress in real time, with results available immediately after a student completes the assessment. Screening windows can be specified by grade level and school. To create screener windows, a primary or lead IXL administrator can sign into their account and navigate to the administrator settings page (see Figure 2). From there, they can set, edit, and delete screener windows.

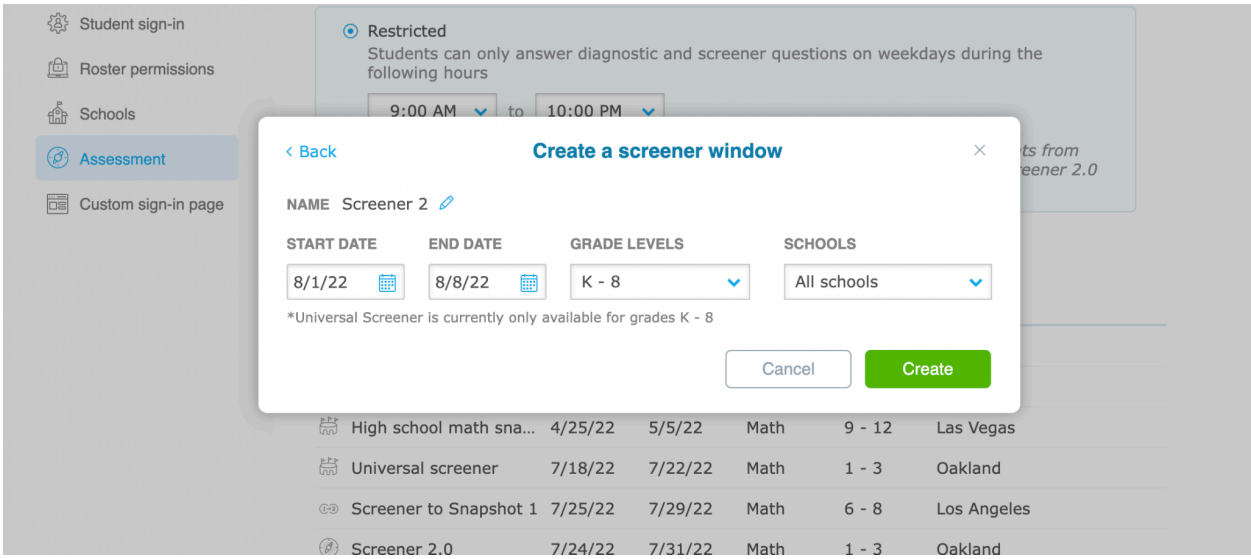


Figure 2. Screenshot showing the creation of a screener window

Item & Test Time

The IXL Universal Math Screener has an overall time limit of 30 minutes as well as a 3-minute time limit for each item. A warning timer appears in the upper right-hand corner of the screen when 30 seconds remain for the current item. When the timer expires, the item is submitted with the currently selected answer (if there is one) or a blank answer. Blank answers are marked as incorrect. This is done so that students do not get stuck and spend too much time on any single item. Items are written so that they can be completed in less than 3 minutes.

Test Length

The IXL Universal Math Screener is a variable-length CAT. As described previously, a CAT selects questions based on a student's responses in order to obtain a precise estimate of the student's ability in a short period of time. The variable-length nature of the IXL Universal Math Screener ends a student's session once a sufficiently precise estimate is obtained rather than delivering a fixed number of questions. In this way, the amount of time students spend testing is reduced while still maintaining appropriate content balancing. Table 1 provides a grade-level overview of the average number of items, the standard deviation of the number of items, the average amount of time spent testing (in minutes), and the standard deviation of the amount of time spent testing.

Table 1. Average number of items and time spent by grade

		Items		Time (min.)	
Grade	*N*	*M*	SD	*M*	SD
Kindergarten	13079	16.36	6.13	9.72	6.34
Grade 1	19313	21.93	4.03	15.16	6.89
Grade 2	23047	22.18	3.14	17.43	7.18
Grade 3	28450	21.70	3.34	18.65	7.06
Grade 4	28390	22.64	3.37	19.95	7.14
Grade 5	30898	22.22	3.81	19.96	7.33
Grade 6	37628	22.32	3.66	20.78	7.29
Grade 7	37224	21.84	3.36	20.11	7.31
Grade 8	32660	21.57	3.40	19.40	7.14

Test Repetition

The IXL Universal Math Screener can be delivered as needed throughout the year, with any individual student completing the screener up to five times during an academic year. To ensure that mathematics ability is the only construct measured, rather than a student's ability to memorize or recall individual items, an individual student will not be presented with any items that appeared on the student's previous four administrations.

Accessibility

The IXL Universal Math Screener supports several assistive technologies in order to provide a platform that is equally accessible to all students. These features include screen reader compatibility, audio support, keyboard shortcuts, browser zoom up to 200 percent, and high contrast ratios. In addition, teachers and administrators can enable Spanish translation, which allows students to complete the entire assessment in Spanish.

Reporting

After students complete the screener, administrators and teachers have access to key data for planning instruction. Administrators have access to the **Universal Screener Results** report (see Figure 3), where students are grouped into the following classifications:

- On or above grade level
- Below grade level
- Far below grade level

- In progress
- Not enough data
- Did not finish
- Not started
- Not applicable

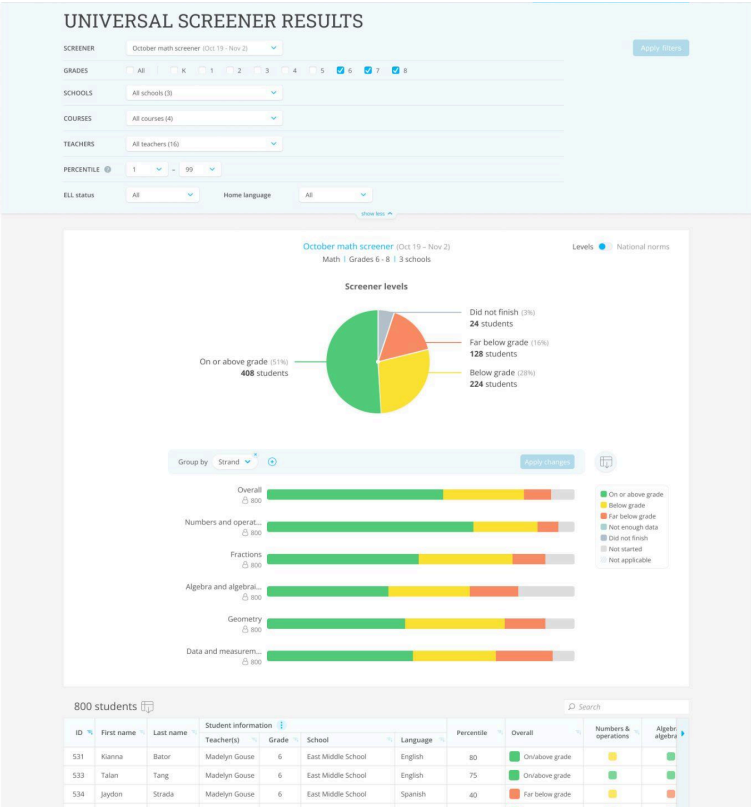


Figure 3. Sample administrator report detailing student performance: Levels view

See the Benchmarking section for details on the first three classifications. **In progress** indicates that the student is currently completing the screener. In rare instances, a student may reach the 30-minute time limit without submitting enough data for the screener to make a confident classification; such students are classified as **Not enough data**. **Did not finish** indicates the student started the screener but abandoned it before completion. **Not started** indicates that the student has not started the screener. **Not applicable** indicates that a strand is not supported for the student's grade (e.g., Fractions for kindergarten students).

The **Universal Screener Results** report for administrators also provides a percentile for each student. Underneath the visual summary in Figure 3, the Percentile column in the student data table indicates the percentage of peers nationally that the student outperformed. This information is summarized visually by setting the toggle in the upper right corner to National Norms as shown

in Figure 4. Administrators see the distribution of percentile ranks across all students, with the number and proportion of students at or above the 50th percentile nationally highlighted in light green. Distribution of percentile ranks grouped by grade level, school, course, or demographics is also available. The Overall column provides the student's overall grade level classification as described above. Administrators can export reports with both overall percentiles and overall levels.

Teachers can access data for individual students on the screener landing page. For each student, teachers see their overall percentile and overall level as well as their percentile and level in each strand. Teachers also have the option to export the available data into a CSV file.

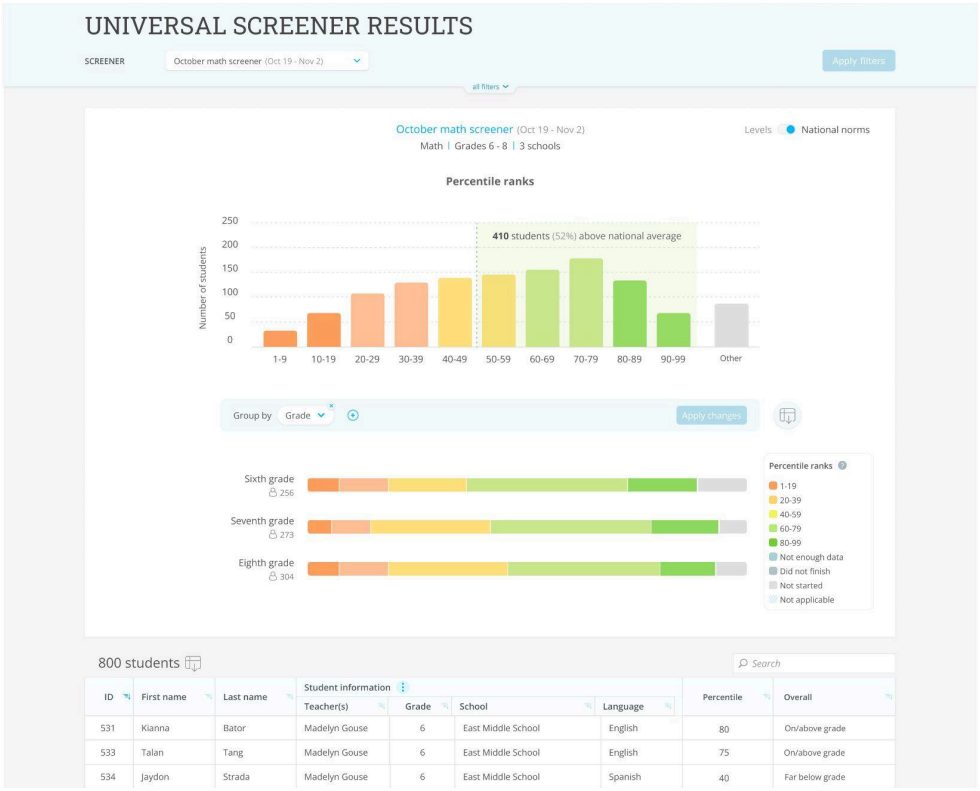


Figure 4. Sample administrator report detailing student performance: National Norms view

CONTENT DEVELOPMENT

Principled Assessment Design

Principled assessment design is a general framework for designing, developing, and implementing an assessment to support the ongoing accumulation and synthesis of evidence to support the validity claims made by the assessment (e.g., Ferrara et al., 2016). This framework requires assessment targets to be clearly defined at the beginning of the process. These assessment targets drive the entire assessment development plan, and continuous focus on targets ensures that all subsequent decisions are consistent with providing evidence to support validity claims.

The assessment targets for the IXL Universal Math Screener are achievement level descriptors (ALDs) derived from all 50 states' standards. Achievement level descriptors are statements that describe expectations of what students at specific achievement levels should know and be able to do. As the IXL Universal Math Screener is designed to identify students who are experiencing moderate or severe difficulties in math, the achievement level was conceptualized as the threshold between on-grade achievement and below-grade achievement. Therefore, the ALDs were written for each educational standard to describe what a minimally competent student would know and be able to do.

The completed set of ALDs were reviewed by an external panel of experienced educators to ensure that they accurately describe a minimally competent student. The subject-matter experts (SMEs) were in-service educators who had significant experience teaching or supervising the grade level(s) they evaluated. Considerable care was taken to ensure that their demographics matched the demographics of elementary and middle school teachers in the U.S. Table 2 provides a comparison of sample demographics to those of the U.S. as a whole using data from the U.S. Department of Education, National Center for Education Statistics, Common Core of Data.

Table 2. Subject-Matter Expert Sample and National Demographics

Variable	Sample (n = 110) %	United States %
Gender		
Male	8	23
Female	91	77
Race/ethnicity		
White	72	80
Black	10	6
Hispanic	1	9
Asian	2	2
Pacific Islander	3	0
American Indian/Alaska Native	2	0
Two or more races / Other	10	2
Educational attainment		
Postbaccalaureate degree	89	60
Teaching certificate or advanced degree	99	98
Years of teaching experience		
Less than 3 years	4	7
3-9 years	30	29
10-20 years	48	37
More than 20 years	18	26

Subject-matter experts (i.e., elementary and middle school math teachers) evaluated both the ALDs for each grade level and sample screener items corresponding to each ALD. When evaluating ALDs, raters assessed whether each ALD aligned to the grade level and were instructed to select "yes" if the descriptor accurately described a performance criterion a proficient student should be able to meet by the end of the target grade. When evaluating screener items, raters viewed an item with its corresponding ALD. Raters then identified whether each sample screener item aligned to the ALD, selecting "yes" if the screener item accurately assessed part or all of the ALD.

We assessed inter-rater reliability separately for ALDs and sample screener items by grade level using Gwet's first-order agreement coefficient (AC1). We selected Gwet's AC1 as the measure of inter-rater reliability because it accounts for agreement due to chance, can accommodate ratings from more than two individuals, and is less skewed than other reliability measures (e.g., generalized Kappa) when rates of agreement are high. For more information about Gwet's AC1,

see Gwet (2008) and (2014). AC1 values for ALDs and screener items by grade are presented in Table 3.

Table 3. Gwet's AC1 for ALDs and items by grade

	ALDs		Items	
Grade	AC1	95% CI	AC1	95%CI
Kindergarten	0.87	[0.83, 0.92]	0.89	[0.84, 0.95]
Grade 1	0.91	[0.88, 0.95]	0.89	[0.84, 0.93]
Grade 2	0.87	[0.82, 0.92]	0.93	[0.90, 0.97]
Grade 3	0.91	[0.87, 0.96]	0.94	[0.91, 0.97]
Grade 4	0.80	[0.75, 0.86]	0.95	[0.92, 0.97]
Grade 5	0.83	[0.78, 0.88]	0.95	[0.92, 0.97]
Grade 6	0.86	[0.82, 0.91]	0.92	[0.89, 0.96]
Grade 7	0.83	[0.79, 0.88]	0.88	[0.84, 0.92]
Grade 8	0.94	[0.91, 0.97]	0.92	[0.88, 0.95]

Various benchmarking scales exist for the qualitative evaluation of agreement coefficients (e.g., Landis-Koch, Altman, Fleiss). Although there is variability among these scales as to what constitutes "good," "very good," or "excellent" agreement, in general, values greater than 0.6 are considered acceptable and values greater than 0.8 are deemed very strong. Based on these guidelines, inter-rater reliability was very high across grade levels for both ALDs and screener items.

Content Specifications

Both achievement level descriptors and their corresponding items are organized into familiar and meaningful mathematical domains. As previously noted, five strands were selected:

- Numbers & Operations,
- Algebra & Algebraic Thinking,
- Fractions,
- Geometry, and
- Data & Measurement.

Due to the adaptive nature of the IXL Universal Math Screener, the total number of items delivered to students varies based on several factors. However, lower and upper bounds on the number of

items from any content strand are set based on the student grade and administration window (i.e., beginning, middle, or end of year). Appendix A shows the relationship between the time of year and test constraints.

Item Development

Item writers were selected based on their relevant grade-level and subject-matter experience as math Ph.D.s and K–8 classroom teachers. Item development was not constrained to a single item type, such as multiple-choice questions. Instead, items were developed using the most appropriate item type for assessing the target knowledge, skill, or ability. Therefore, students are presented with a mix of selected-response and constructed-response questions.

Where appropriate, item writers applied the principles of Universal Design (Thompson et al., 2002) to reduce construct-irrelevant variance while simultaneously increasing accessibility. Items were written to target achievement level descriptors while considering important factors such as reading level, cognitive load, vocabulary, and sentence length, among others. An external panel of experienced educators was recruited to review sample items alongside the evaluation of ALDs described previously. The SME panel consisted of in-service educators with significant experience teaching or supervising the grade level(s) they evaluated, and with demographics matching that of elementary and middle school teachers in the U.S. Multiple rounds of internal review were then conducted to ensure all items appropriately targeted the intended achievement level descriptors while meeting the content and item writing specifications.

ITEM & SCALE CALIBRATION

Background

Item calibration is the process of estimating the difficulty of a test item relative to other items and placing that value on a common scale. To ensure the accuracy of these estimates, each item needs a sufficient number of responses across the range of ability. Once items are calibrated and a scale is established, new items are embedded as unscored field test items alongside existing scored items until a sufficient number of responses have been recorded. The field test items are then calibrated, and those that pass quality assurance are added to the scored item pools.

Measurement Model

The IXL Universal Math Screener uses an algorithm based on Rasch Measurement Theory (Rasch, 1960). The Rasch model is a sophisticated statistical approach that places student ability and test item difficulty on a common scale and is mathematically equivalent to a scaled version of the one-parameter logistic item response theory model where the discrimination parameter is set to 1 across all items (Hambelton & Swaminathan, 1985). A student with a high value on this scale has more ability than a student with a low value, and an item with a high value on this scale is more difficult than one with a low value. A student with the same value as an item has a 50% chance of responding to it correctly. Rasch modeling makes it possible to calculate comparable scores for students even though they are administered different test items. It also makes it possible to determine the probability that a student knows a certain skill without specifically administering items related to every skill.

Determining the probability of a correct response is a function of the relationship between student ability and item difficulty. An item response function (IRF) graphically illustrates this relationship. Figure 5 is a plot showing the item response functions for three items of varying difficulty. The horizontal axis shows the range of student ability from low ability (-4.00) to high ability (4.00). The vertical axis represents the model probability of a correct response. An item's difficulty is modeled as the point along the student ability scale at which a student has a 50% chance of answering the question correctly. For example, the figure shows that item A is the easiest item with a difficulty calibration of -2.0, item B is slightly harder with a difficulty calibration of 0.0, and item C is the hardest with a difficulty calibration of 1.5.

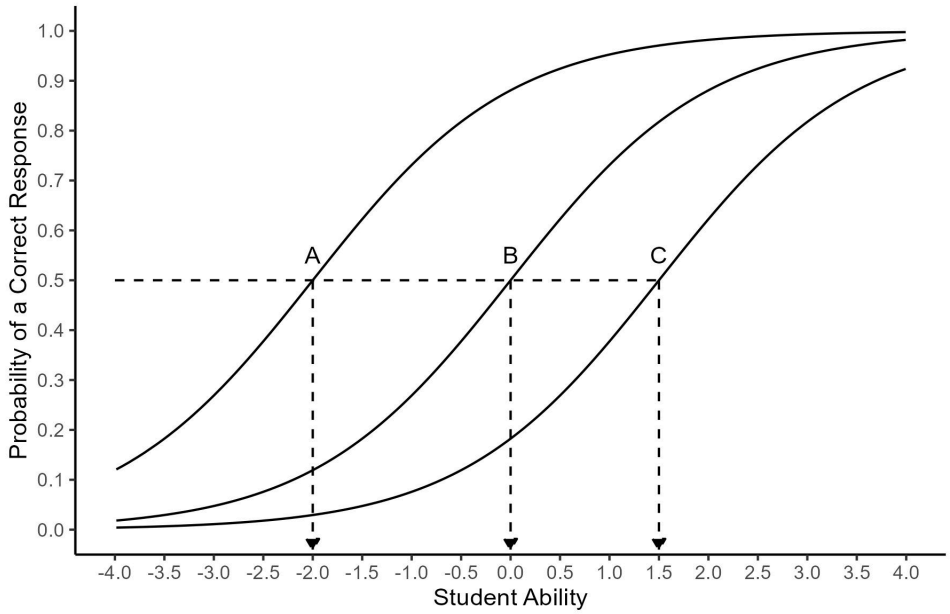


Figure 5. Example item response functions for three items

Item Calibration

The IXL Universal Math Screener utilizes a pre-calibrated item pool (Lord, 1980), where all items have been placed on the same logit scale. In a pre-calibrated item pool, each item's difficulty is estimated and scaled relative to all other items in the pool to ensure fair classification of students taking different test forms built using the same pre-calibrated item pool. Unlike most equating methods that are implemented *after* a student has completed an assessment to adjust for small differences in test form difficulty, item pre-equating allows for the appropriate placement of students along the ability scale even *before* the completion of an assessment, assuming they answer a sufficient number of items. Item pre-equating from a calibrated item pool enables the delivery of computerized adaptive tests (described below) as well as linking field-test items onto the same scale as "live" or scorable items.

A common-item linking design with concurrent calibration (see Kolen & Brennan, 2014, pp. 440-442) was used to create a vertical scale for the IXL Universal Math Screener. With this design, students respond to some items that are off-grade level in order to create a vertical scale, which allows for the comparison of student scores across grades and student growth across years.

Psychometric checks to ensure person and item fit to the model were conducted iteratively until the final person and item set exhibited acceptable model fit. Once the item difficulty estimates were set, these parameters were fixed so that the scale would be maintained across years.

Field testing is the process of adding new unscored items to the pool after the initial scale has been set. Once these items have reached the minimum number of exposures, they are then calibrated onto the score scale, and new field test items are deployed. This is an iterative process that is monitored continuously, as described in the following section.

Quality Assurance

Items on the IXL Universal Math Screener are continuously monitored for quality assurance to ensure they function as expected. Similarly, field test items must pass extensive quality assurance checks prior to achieving promotion to the scored item pool. These quality assurance checks include both traditional methods for item analysis as well as Rasch-specific methods.

Rasch models are prescriptive, meaning items must fit the model; therefore, two statistics are examined to ensure proper fit: mean square and standardized (Wright & Masters, 1982). The mean-square fit statistic is a chi-square statistic divided by its degrees of freedom, which provides an expected value close to 1.0. Values above 1.0 indicate unpredictability in the responses, while values below 1.0 indicate overly predictable responses. The standardized fit statistic is a transformation of the mean-square statistic into a unit-normal distribution, similar to a t-statistic with infinite degrees of freedom, which reports the probability of the mean-square statistic occurring by chance. Mean-square values between 0.5 and 1.5 and standardized values between -2.0 and 2.0 are typically considered acceptable.

In addition, several traditional item analysis statistics are calculated. The p-value of an item is the proportion of examinees who answer the question correctly. An item that 20% of the students answered correctly would be said to have a p-value of 0.20, while an item that 80% of the students answered correctly would have a p-value of 0.80. The point-biserial correlation is a special case of the Pearson product-moment correlation, where one variable is continuous and the other is dichotomous. A high point-biserial coefficient is desirable as it indicates that students who answer a given item correctly tend to have higher overall test scores. Conversely, low or negative point-biserial coefficients indicate an inverse relationship with the overall test score and are indicative of items that are not performing as expected. P-values between 0.10 and 0.90 and point-biserial coefficients above 0.20 are typically considered acceptable for use

Computerized-Adaptive Testing Design & Delivery

The IXL Universal Math Screener computerized adaptive testing (CAT) algorithm employs four distinct rules for delivery: (1) a starting rule, (2) a testing rule, (3) a stopping rule, and (4) a final scoring rule. When a student begins the assessment, the algorithm's starting rule randomly selects an item within a range of difficulty that varies by grade level. Each range covers the first quartile to the median difficulty of each grade level's item bank, helping ensure that most students feel successful at the start of the test, regardless of ability.

As a student progresses through the assessment, the algorithm updates its estimate of ability and the standard error of this ability estimate each time the student responds to an item using a conventional Bayesian *expected a posteriori* (EAP) estimation method. As the algorithm becomes more certain in its running estimate of student ability (as indicated by the decreasing standard error of the estimate), it selects items more suitable to the student's ability estimate. That is, the algorithm chooses more difficult items as the student answers correctly and less challenging items as the student answers incorrectly. This adaptivity reduces the number of items needed to measure math ability, thus increasing test efficiency and improving the test experience.

The item selection method used in the testing rule combines a traditional approach (i.e., maximum Fisher information [MFI]) with a weighting method akin to content balancing. Specifically, the maximum priority index (MPI) is a two-stage process (Cheng & Chang, 2009). In the first stage, the item selection algorithm calculates the MFI for each unadministered item and weights this value by the ratio of the remaining unadministered items from each content area over the total minimum number of required items from each content area to calculate an MPI. Then, the algorithm randomly selects an item out of 5 with the highest MPI values conditioned on a student's provisional ability estimate to reduce item exposure and increase item bank coverage similar to the randomesque procedure (Kingsbury & Zara, 1989). After the student has answered the minimum number of items from each content area (i.e., the lower bound constraint is met), MPI moves on to the second stage, which uses the upper bound constraints as targets, and culminates in the stopping rule phase.

The assessment ends for a student when one of the four stopping rule components is satisfied. The four components of the stopping rule are (1) overall precision, (2) confidence in classification, (3) maximum number of items, and (4) maximum amount of time. The stopping rule will end the assessment if the overall precision of the student’s ability estimate reaches a prespecified level—specifically, if the conditional standard error of measurement (CSEM) is less than 0.4 logits. The stopping rule will also end the assessment when the algorithm is sufficiently confident that the student is correctly classified as **On or Above Grade Level** or **Far Below Grade Level**, which occurs when an ability estimate's 90% confidence interval is completely within one of those two score ranges. In addition to these components, the IXL Universal Math Screener will end when the maximum number of items is reached or the maximum amount of time has passed, regardless of the final ability estimate or CSEM.

Once the algorithm has ended an assessment, the final scoring rule is implemented. This rule re-estimates the student’s ability using EAP with a uniform prior distribution, which is mathematically equivalent to maximum likelihood estimation (see Hambelton & Swaminathan, 1985, p. 93).

BENCHMARKS & NORMS

Benchmarking

Benchmark assessments are intended to evaluate students against specific grade-level standards to identify relative strengths and weaknesses for guiding future instruction rather than to determine content mastery. The purpose of the IXL Universal Math Screener is to identify students who are experiencing moderate or severe math difficulties; therefore, student performance is classified into one of the following achievement levels: **On or above grade level**, **Below grade level**, or **Far below grade level**. Benchmarks are also set relative to the beginning (August 1 – November 30), middle (December 1 – February 28), and end (March 1 – June 1) of the school year.

For each grade, achievement level descriptors for end-of-year performance were derived from all 50 states' standards. Conceptually, the **On or above grade level** classification means that a student exhibits performance that suggests they are at least minimally competent for their grade level. The **Below grade level** classification suggests that the student is performing at a level below the **On or above grade level** cut score but is not more than one academic year below it. The **Far below grade level** classification suggests that the student is performing at a level *more* than one grade level below the student's actual grade.

Because the IXL Universal Math Screener is typically administered in three distinct windows [beginning-of-year (BOY), middle-of-year (MOY), and end-of-year (EOY)], there are different thresholds for benchmarking students based on the time of year. However, the end-of-year threshold for one grade (e.g., third) is used as the BOY threshold for the upward adjacent grade (i.e., fourth). The MOY benchmark is the midpoint between the BOY and EOY benchmarks within a given grade.

The Embedded Standard Setting method (ESS; Lewis & Cook, 2020) was used to determine the cut scores between **On or above grade level** and **Below grade level** for grades K–5. The modified-Angoff standard setting method (Angoff, 1971) was used to determine the cut scores for grades 6–8.

Norming

Assessment norms provide information about the typical levels of performance for an identifiable population of students or schools. For example, a student may achieve the highest test score in her class on a given assessment but still fall below the national average of students at her grade level who have completed the same assessment (i.e., a percentile rank < 50). This information allows educators to compare their students' scores to the scores of students across the United

States who completed the same assessment. Such comparisons are often used to help educators appropriately target resources to maximize student learning and achievement.

Sample

The norming sample included IXL Universal Math Screener assessments administered to 125,281 students from 811 schools in 46 states and the Bureau of Indian Education during the 2024-25 academic year. School-level demographic data were obtained from the National Center for Education Statistics (NCES; https://nces.ed.gov/ccd). After matching the IXL Universal Math Screener sample with school-level demographics, we were able to compare the IXL Universal Math Screener student demographics with the national averages. As seen in Table 8, the IXL Universal Math Screener sample is sufficiently representative of student demographics nationwide.

Table 8. Demographic representation for the norming sample

Variable	National %	IXL %
Gender		
Female	48.7	48.8
Male	51.3	51.2
Race/ethnicity		
American Indian or Alaska Native	1.0	1.7
Asian	5.9	6.2
Black or African American	15.6	14.3
Hispanic or Latino	31.3	22.2
Native Hawaiian or Pacific Islander	0.3	0.2
White	46.2	55.6
Free or reduced-price lunch	53.7	53.0

Results

Table 9 shows the correspondence between IXL Universal Math Screener classifications and percentile ranks by grade. Classification at a given time of the year provides a criterion-referenced inference regarding a student's math ability at that time relative to standards-based expectations for their grade level. Percentile ranks provide a norm-referenced inference of a student's performance relative to their peers; specifically, percentile ranks indicate the percentage of scores that fall below a specific score. This table serves the following three purposes.

First, it provides educators with a high-level breakdown of student performance nationwide. For example, at the beginning of the school year, about 4% of second-grade students were **Far below grade level**, another 22% were Below grade level, and the remaining 73% were **On or above grade level**.

Second, it allows for the differentiation of students who received the same classification. Using the individual student percentile ranks provided within IXL Analytics, an educator can refer to this table and interpret relative performance on the IXL Universal Math Screener. For example, a second-grade student who scored in the 25th percentile at the beginning of the year is close to being on level, whereas one who scored in the 6th percentile is further away, even though both students received the same **Below grade level** classification.

Third, educators can use the percentile ranges associated with the **Far below grade level**, as well as school and individual student contexts, to identify appropriate cutoffs for Tier III interventions within a Response-to-Intervention/Multi-Tiered System of Supports (RTI/MTSS) framework.

Additional information associated with the relationship between the item response theory theta scores used to determine the student classification and percentile ranks can be found in Appendix B.

Table 9. Classifications and percentile ranks by grade and window

Classification	K	1	2	3	4	5	6	7	8
BOY									
Far Below	*NA*	1	1-4	1-16	1-32	1-57	1-74	1-76	1-81
Below	1-3	2-9	5-26	17-55	33-59	58-81	75-86	77-85	82-87
On/Above	4-99	10-99	27-99	56-99	60-99	82-99	87-99	86-99	88-99
MOY									
Far Below	*NA*	1-2	1-7	1-19	1-33	1-56	1-73	1-73	1-76
Below	1-4	3-15	8-42	20-51	34-57	57-74	74-83	74-83	77-83
On/Above	5-99	16-99	43-99	52-99	58-99	75-99	84-99	84-99	84-99
EOY									
Far Below	1	1-4	1-15	1-28	1-42	1-56	1-69	1-78	1-74
Below	2-10	5-29	16-55	29-47	43-67	57-70	70-76	79-84	75-81
On/Above	11-99	30-99	56-99	48-99	68-99	71-99	77-99	85-99	82-99

RELIABILITY & PRECISION

Background

Test reliability refers to the consistency of scores obtained from an assessment. All assessments are subject to measurement error, and reliability coefficients attempt to quantify the amount of error. A test that is free from error would have a reliability coefficient of 1.0; however, that is unlikely to occur in practice, and coefficients around 0.80 or higher are more plausible and desirable. Often, the reliability of a traditional, fixed-form assessment is evaluated via internal consistency measures such as Cronbach's alpha or McDonald's omega. However, in a CAT assessment context, reliability is typically assessed in terms of consistency between measurement events or the amount of measurement error along the student ability continuum.

Internal Consistency

Although traditional measures of reliability are not estimable for adaptive assessments, marginal reliability provides a method that closely approximates the traditional measures of internal consistency when the ability distribution and item parameters are known (Dimitrov, 2003; Samejima, 1977, 1994). Marginal reliability is the ratio of observed score variance to true score variance and is shown as

$$\widehat{\rho} = \frac{var(\widehat{\theta})}{var(\widehat{\theta}) + SE(\widehat{\theta})^2},$$

where $\widehat{\rho}$ is the marginal reliability coefficient, $var(\widehat{\theta})$ is the total variance of the student ability estimates in the population, and $SE(\widehat{\theta})^2$ is the mean squared conditional standard error of the ability estimates across the population. Table 4 shows the predicted marginal reliabilities across all grades (see Appendix C for reliability by grade and strand). .

Table 4. Marginal reliability by grade

Grade	*N*	Reliability
Kindergarten	63,918	0.86
Grade 1	90,929	0.88
Grade 2	110,192	0.88
Grade 3	128,384	0.90
Grade 4	124,513	0.90
Grade 5	127,115	0.90
Grade 6	129,359	0.92
Grade 7	126,248	0.91
Grade 8	120,939	0.92
Total	*1,021,597*	*0.92*

Test-Retest Reliability

Test-retest reliability is the extent to which test results are consistent over time. A test resulting in similar scores across multiple administrations has high test-retest reliability. If students take a test on Monday and then again on Friday, we may not expect identical scores, but we would expect to see similar scores. The more similar a test's scores are within a narrow timeframe, the greater its test-retest reliability.

Alternate-form reliability is the extent to which two forms of a test are equivalent. Traditionally, in a test-retest reliability analysis, the same test is administered multiple times. The time at which a test is administered a second time is important because enough time needs to pass to control for testing effects, which occur when merely taking a test influences subsequent scores upon repeated administrations (Shadish et al., 2002). Furthermore, allowing too much time to pass will introduce more confounding factors (e.g., differential growth). In an alternate-form reliability analysis, both forms of a test could potentially be administered simultaneously. Because a CAT delivers a different set of items to a student on different testing events, test-retest reliability is conceptually similar to alternate-forms reliability.

To estimate test-retest reliability, scores from an initial testing administration are correlated with scores from a second testing administration. For our analysis, we paired beginning-of-year (BOY) with middle-of-year (MOY) and middle-of-year (MOY) with end-of-year (EOY) testing windows and calculated an intraclass correlation coefficient (ICC). Any ICC value less than 0.40 indicates low

reliability, values between 0.40 and 0.59 indicate fair reliability, values between 0.60 and 0.74 indicate good reliability, and values above 0.75 indicate excellent reliability (Cicchetti, 1994).

Table 5 reports these test-retest reliability coefficients. Most grade-level ICCs were above 0.75, which indicates excellent test-rest reliability. All other grade-level ICCs were above 0.6, which indicates good test-retest reliability. The overall ICCs were both above 0.75, indicating excellent test-retest reliability.

Table 5. Test-retest reliability between delivery windows by grade

	BOY to MOY			MOY to EOY		
Grade	***N***	**ICC**	**95% CI**	***N***	**ICC**	**95% CI**
Kindergarten	7,578	0.616	[0.581, 0.647]	7,119	0.623	[0.609, 0.637]
Grade 1	10,617	0.741	[0.593, 0.821]	9,380	0.743	[0.733, 0.751]
Grade 2	12,463	0.834	[0.823, 0.844]	10,138	0.739	[0.731, 0.748]
Grade 3	14,263	0.840	[0.807, 0.865]	11,012	0.788	[0.781, 0.795]
Grade 4	13,046	0.871	[0.853, 0.886]	10,493	0.801	[0.794, 0.808]
Grade 5	12,699	0.862	[0.795, 0.901]	9,711	0.789	[0.782, 0.797]
Grade 6	11,190	0.874	[0.844, 0.896]	9,132	0.765	[0.756, 0.773]
Grade 7	10,299	0.861	[0.825, 0.888]	7,966	0.755	[0.745, 0.764]
Grade 8	10,045	0.855	[0.806, 0.888]	7,614	0.742	[0.731, 0.752]
Total	*102,200*	*0.826*	*[0.824, 0.828]*	*82,565*	*0.835*	*[0.833, 0.837]*

Standard Error of Measurement

A student's standard error of measurement (SEM) is an indicator of the precision of their score and describes the range in which a score may vary upon repeated testing that is due to chance. SEM is a function of the interaction between the ability of a student, the difficulty of the items, and the number of items on a test. A lower SEM indicates less error and more precision around a score. Because the CAT algorithm selects items based on a student's estimated ability level, it is able to target a student more accurately, and significantly decrease the SEM with fewer items than a traditional fixed-form assessment.

It derives the SEM by calculating the standard deviation of the posterior distribution of the EAP estimate by integrating over all possible values of ability given a response pattern (Bock & Mislevy, 1982). The following formula shows how the IXL Universal Math Screener calculates SEM:

$$SEM = \sqrt{\frac{\sum_{i=1}^{I} (X_i - \hat{\theta})^2 L(X_i) W(X_i)}{\sum_{i=1}^{I} L(X_i) W(X_i)}}$$

where *i* is the *i*th item in a student's response pattern, *I* is the total number of items the student completed, *X* is an incorrect (i.e., 0) or a correct (i.e., 1) response, $\hat{\theta}$ is the interim ability estimate, *L* is the likelihood of the response pattern conditioned on the interim ability estimate, and *W* is the weight associated with that quadrature of the prior distribution.

Although the CAT algorithm targets students more accurately according to their ability, it generally performs better for students whose ability and performance are closer to the typical ability of students at or near their grade level. Figure 6 shows the distribution of standard errors across the student ability spectrum, while Table 6 shows the mean and median standard errors by grade with corresponding 95% confidence intervals.

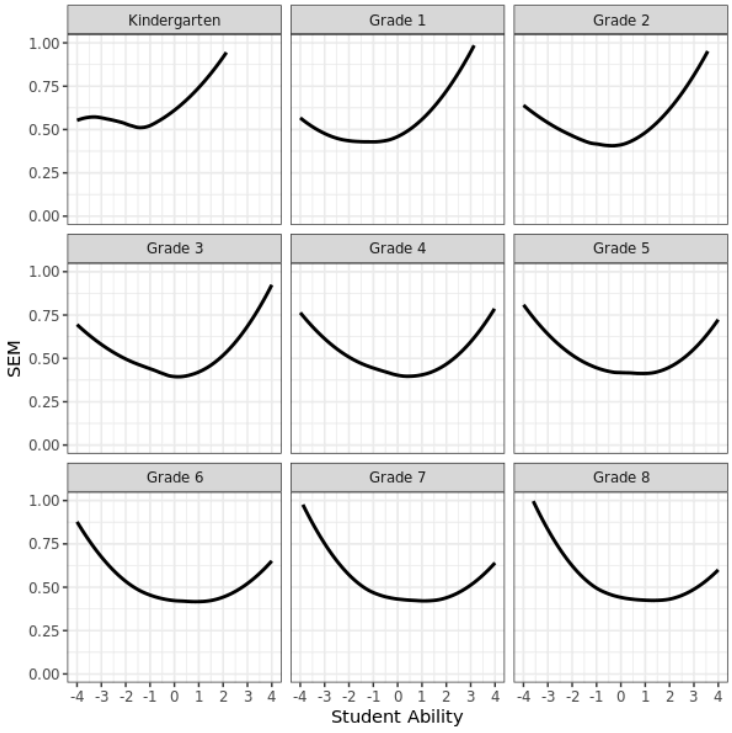


Figure 6. Standard Error of Measurement across the ability spectrum by grade

Table 6. Mean and Median standard errors by grade

Grade	*N*	Median	Median Lower CI	Median Upper CI	Mean	Mean Lower CI	Mean Upper CI
Kindergarten	63,822	0.554	0.553	0.556	0.583	0.580	0.586
Grade 1	97,333	0.453	0.453	0.454	0.476	0.474	0.478
Grade 2	122,892	0.438	0.438	0.438	0.463	0.462	0.465
Grade 3	153,882	0.433	0.433	0.433	0.455	0.454	0.456
Grade 4	154,804	0.426	0.426	0.426	0.445	0.444	0.446
Grade 5	165,838	0.427	0.427	0.427	0.448	0.447	0.449
Grade 6	171,192	0.435	0.435	0.435	0.464	0.463	0.465
Grade 7	160,038	0.441	0.441	0.442	0.476	0.474	0.477
Grade 8	148,994	0.447	0.447	0.448	0.486	0.485	0.488

Decision Accuracy & Consistency

Other methods to examine the reliability of a test, particularly ones that classify students into achievement levels, are decision accuracy and classification consistency. Decision accuracy is the degree to which students' observed classifications would agree with their "true" classifications, assuming known cut scores on a single assessment. Decision consistency is the degree to which students would be classified into the same performance categories over parallel replications of the same assessment. We used the Rudner (2001, 2005) method to calculate classification accuracy and an extension to this method (Wyse & Hao, 2012) to calculate classification consistency.

These indices are calculated using a student's standard error and their ability estimate along the latent trait scale relative to a cut score threshold. The closer a student's ability estimate is to the cut score, the greater the probability of misclassification. A probability is calculated for each student and then averaged across the sample to provide a single index. Thus, decision accuracy and consistency indices can be interpreted as the probability that an average student is correctly classified. The results in Table 7 show that across all grade levels, decision accuracy and consistency indices are on par with or better than industry standards.

Table 7. Classification accuracy and classification consistency by grade

Grade	Decision Accuracy [99% CI]	Decision Consistency [99% CI]
Kindergarten	0.924 [0.923, 0.925]	0.894 [0.892, 0.895]
Grade 1	0.889 [0.888, 0.890]	0.844 [0.842, 0.845]
Grade 2	0.883 [0.882, 0.884]	0.838 [0.836, 0.838]
Grade 3	0.898 [0.897, 0.899]	0.855 [0.854, 0.856]
Grade 4	0.916 [0.915, 0.917]	0.881 [0.880, 0.883]
Grade 5	0.944 [0.943, 0.944]	0.920 [0.919, 0.921]
Grade 6	0.957 [0.956, 0.958]	0.939 [0.938, 0.940]
Grade 7	0.960 [0.959, 0.961]	0.943 [0.942, 0.944]
Grade 8	0.964 [0.963, 0.964]	0.949 [0.948, 0.950]

VALIDITY

Background

The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) define validity as follows:

> The degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores for proposed uses that are evaluated, not the test itself. (p. 11)

Following this framework, this section describes the accumulated validity evidence, as of this writing, to support the proposed uses for the IXL Universal Math Screener. The evidence collected includes both internal and external data sources.

Evidence Based on Test Content

Test content refers to the format of items and the content balancing specifications for test delivery and administration. The section on Content Development describes the process used to ensure the content delivered to students appropriately reflects the measured construct (e.g., proper alignment to state standards) and serves as the foundation for the validity evidence based on test content.

Evidence Based on Internal Structure

Unidimensionality

As indicated in the Item & Scale Calibration section, a unidimensional Rasch model is used to calibrate the item parameters and adaptively administer the IXL Universal Math Screener. A primary assumption of this model is unidimensionality, meaning that only a single construct is the target of measurement.

Several methods exist to determine if the unidimensionality assumption is tenable. Most of these methods rely on mostly complete data (i.e., with little to no missingness) collected as responses to items on a fixed-form assessment. Given that the IXL Universal Math Screener is an adaptive assessment and produces sparse data, we first examined the structure of the data using bifactor item response theory modeling. This analytic approach models a general factor (e.g., general math

ability) alongside multiple secondary factors (e.g., algebra, geometry, and fractions) within an item response theory framework using item-level data. If the general factor meets certain statistical criteria, then we can consider the assessment to be unidimensional. To further examine the structure of the IXL Universal Math Screener, we also utilized the confirmatory factor analysis (CFA) method using strand-level scores, which are not affected by data sparseness inherent to adaptive testing. A single factor represents general math ability, with each strand score loading onto the factor. Fit indices indicate whether the specified model is a good fit to the data and if we can consider the assessment to be unidimensional.

Bifactor Analyses

We tested the unidimensionality assumption via the bifactor modeling approach (Rodriguez et al., 2016). In this approach, a general factor (i.e., general construct) accounts for the common variance shared by all items in an assessment, and multiple secondary factors (e.g., fractions and numbers & operations) account for variance beyond that accounted for by the general factor. The percentage of systematic variance accounted for by each factor estimated via the explained common variance (ECV) statistic describes the degree of (uni)dimensionality present in the response data. Values greater than 0.90 indicate that the response data are essentially unidimensional; values between 0.70 and 0.90 require additional information to determine if assuming unidimensionality is tenable; and values less than 0.70 indicate that multidimensional modeling may be required. Furthermore, stable item parameters can be estimated when the proportion of variance accounted for by the general factor is at least 0.20 (Reckase, 1979).

To confirm that the IXL Universal Math Screener is not measuring multiple secondary factors alongside a strong general factor, we examined the percentage of uncontaminated correlations (PUC). Values greater than 0.80 indicate that correlations between the items in the assessment are uncontaminated by multidimensionality. This means that whatever multidimensionality is present will not bias results. A high PUC also indicates a high degree of similarity between the general factor in the bifactor model and the single factor measured in a unidimensional model. It is worth noting that research suggests high PUC values may be more important than low ECV values in detecting potential bias when fitting a unidimensional model to multidimensional data (Bonifay et al., 2015; Reise et al., 2013). Other helpful bifactor modeling fit statistics include construct reliability (H) and factor determinacy (FD). In this context, H indicates the degree to which the construct is replicable across testing settings, and FD indicates the degree to which factor scores are reliable and representative of students' general mathematics ability. Together, these statistics indicate how well the factor is defined. H values > 0.80 and FD values > 0.90 indicate the presence of a robust latent trait (i.e., math ability).

When we examined the general factor of mathematics ability and secondary factors of ability within each strand, the general factors across all grades accounted for at least 20% of the total variance, with the secondary factors accounting for relatively minor amounts of the total variance in comparison. The general factors' ECV values across all grades were less than 0.90 but more than

0.70, indicating that additional information was necessary to determine the tenability of using a unidimensional model. The PUC values across all grades were greater than 0.70, with most approaching 0.80, suggesting that the suboptimal ECV values were less concerning. Lastly, the FD and H values of the general factors across all grades were all greater than 0.90, with most nearly equal to 1.00, indicating almost perfect reliability in the factor scores and constructs across testing administrations, respectively. The combination of these indices suggests strong evidence in support of a unidimensional Rasch model to administer the IXL Universal Math Screener (see Table 10).

Table 10. Bifactor analyses results by grade

Grade	PUC	PTV	ECV	H	FD
Kindergarten	0.713	0.372	0.825	0.992	0.998
Grade 1	0.717	0.440	0.853	0.998	0.999
Grade 2	0.731	0.457	0.863	0.999	0.999
Grade 3	0.766	0.454	0.869	0.999	0.999
Grade 4	0.777	0.409	0.835	0.999	0.999
Grade 5	0.789	0.415	0.836	0.999	0.999
Grade 6	0.792	0.389	0.808	0.999	0.999
Grade 7	0.751	0.361	0.751	0.998	0.999
Grade 8	0.781	0.366	0.781	0.998	0.999

Note: PUC = percentage of uncontaminated correlations; PTV = proportion total variance; ECV = explained common variance; H = construct reliability; and FD = factor determinacy.

Confirmatory Factor Analyses

We also performed confirmatory factor analyses (CFAs) for each grade to further support using a unidimensional scale to measure math ability. Similar to a bifactor model analysis, a confirmatory factor analysis is a statistical method for determining the appropriateness of a theoretical model specifying the relationships between observed variables (e.g., algebra, fractions, and geometry scores) and a latent, unobserved variable (e.g., general math ability; Brown, 2015; Kline, 2023).

We examined several fit indices to confirm that the unidimensional CFA model represented the data well. First, we looked at χ^2, which compares the difference between a specified model and the variances and covariances inherent to the data. Ideally, we aim for a statistically nonsignificant χ^2 value, indicating no significant difference between the theoretical model and the characteristics of the data. However, χ^2 is notoriously sensitive to sample size, where larger samples almost invariably result in a statistically significant χ^2. Next, we examined the comparative fit index (CFI),

where values less than 0.90 are considered unacceptable, values greater than 0.90 but less than 0.95 are considered acceptable, and values greater than 0.95 are considered excellent (Hu & Bentler, 1999). After CFI, we examined the root mean squared error of approximation (RMSEA), where values greater than 0.08 are considered unacceptable, values less than 0.08 but greater than 0.06 are acceptable, and values less than 0.06 are considered excellent (Hu & Bentler, 1999). Then we examined the standardized root mean square residual (SRMR), where values greater than 0.10 are unacceptable, values less than 0.10 but greater than 0.08 are considered acceptable, and values less than 0.08 are considered excellent (Hu & Bentler, 1999).

Table 11 reports the model fit indices and statistics of the CFAs. Across all grades, χ^2 was statistically significant. However, these results were not unanticipated because all sample sizes were substantially larger than 500. All CFI values were greater than 0.95, indicating an excellent fit. Except for grade 8, all RMSEA values indicated excellent fit. Lastly, all SRMR values were below 0.08, indicating an excellent fit. These CFA results, along with the results of bifactor analyses, support using a vertically scaled, unidimensional Rasch model to measure math ability. See Appendix D for the corresponding factor loadings.

Table 11. Confirmatory factor analyses results by grade

Grade	*N*	χ2	*df*	CFI	RMSEA	SRMR
Kindergarten	13,479	46.896	2	0.987	0.051	0.016
Grade 1	19,818	73.717	2	0.990	0.056	0.017
Grade 2	22,093	83.248	5	0.992	0.033	0.014
Grade 3	122,008	1116.235	5	0.977	0.051	0.020
Grade 4	117,918	1024.140	5	0.980	0.051	0.019
Grade 5	120,574	747.395	5	0.987	0.044	0.015
Grade 6	123,006	479.056	5	0.993	0.034	0.011
Grade 7	119,428	1087.584	5	0.982	0.053	0.018
Grade 8	113,403	2575.383	5	0.958	0.083	0.030

Differential Item Functioning

In addition to the requirement of unidimensionality, a further requirement of the Rasch model is that it is invariant across subgroups, meaning that an item should function the same way for any student who responds to the item, regardless of the personal or demographic characteristics of that student. Differential item functioning (DIF) analysis investigates each item for signs of interactions with sample characteristics. An item is said to exhibit DIF when equally able individuals from different groups have notably differing probabilities of answering the item

correctly. DIF detection procedures help to gather validity evidence for the proposed interpretations of test scores by ensuring that scores are free from potential bias and that individual items do not create an advantage for one group over another.

The residual-DIF method is an approach for detecting DIF by comparing the residuals of item responses between demographic groups (e.g., male and female students; Lim, Choe, & Han, 2022; Lim & Choe, 2023). Residuals are the difference between each student's actual response to an item and the expected probability of a correct response to the item conditioned on their estimated ability. The following equation defines residuals as:

$$r_i = x_i - P_i$$

where *r* is the residual, x_i is the response (i.e, 0 = incorrect; 1 = correct) to item *i*, and P_i is the probability of correctly responding to an item conditioned on estimated ability. The difference between the group means of the residuals defines the RDIF statistic and is expressed as:

$$RDIF_i = \frac{\sum_{j=1}^{J} r_{ij}}{J} - \frac{\sum_{k=1}^{K} r_{ik}}{K}$$

where $RDIF_i$ is the residual-DIF statistic of item *i*, *j* is the j^{th} person in the focal group, *J* is the total number of people in the focal group, *k* is the k^{th} person in the reference group, and *K* is the total number of people in the reference group. *RDIF* per se indicates the magnitude of DIF present in each item. Separate calculations for effect size are unnecessary. Absolute *RDIF* values less than 0.05 indicate little to no DIF, absolute *RDIF* values greater than 0.05 but less than 0.10 indicate moderate DIF, and absolute *RDIF* values greater than 0.10 indicate large DIF (Lim & Choe, 2023).

When investigating DIF based on reported sex, out of several thousand items that received sufficient exposures from the relevant groups (i.e., male vs. female), 42 items were flagged for potential DIF. Of the items flagged, 13 (31.0%) indicated possible bias in favor of male students, and 29 (69.0%) indicated possible bias in favor of female students. When investigating for DIF based on race/ethnicity (i.e., white vs. others), only 10 items were flagged for potential DIF, all of which indicated possible bias in favor of white students. However, all of the flagged items were found to be free from substantive DIF (see the following section).

Subject-Matter Expert Review

It is important to distinguish between statistical DIF and substantive DIF (Penfield & Lam, 2000; Roussos & Stout, 1996). Statistical DIF refers to the statistical identification of DIF, whereas substantive DIF refers to the identification of construct-irrelevant factors responsible for the statistical DIF (i.e., potential sources of bias). It is always important to remember that statistical DIF is simply a detection strategy and that careful scrutiny of items is always warranted to identify and

address substantive DIF. DIF detection methods may identify some items as being unbiased even though they are indeed biased, while some items identified as having bias may not actually be biased. Therefore, all items identified as potentially biased using the methods outlined above were reviewed by subject-matter experts to ensure that all items are free from substantive DIF.

Evidence Based on Relations to Other Variables

Criterion-related validity describes how well a test or measurement predicts or correlates with an external criterion or outcome. Concurrent validity is supported by high correlations between measures administered close together in time regardless of temporal order, whereas predictive validity is supported when the predictive assessment precedes the criterion measure and the two may be administered farther apart. For both predictive and concurrent validity, we examined the correlation coefficients between scores from the IXL Universal Math Screener and those from a well-established criterion math assessment, using the Pearson product-moment correlation coefficient *r*. This coefficient measures the linear relationship between two variables and ranges from -1.00 to 1.00. An *r* value of 0 indicates no correlation, whereas an *r* value greater than 0.60 indicates a strong positive relationship (Ratner, 2009). High positive correlations between the IXL Universal Math Screener and the criterion math assessment would support the validity of the IXL Universal Math Screener.

Concurrent Validity

We retrieved data from August to November 2024 and examined how well the IXL Universal Math Screener correlates with the criterion measure within a 30-day window. We found strong positive correlations between IXL Universal Math Screener scores and concurrent NWEA MAP® Growth™ (MAP) math scores in Kindergarten through 8th grade (see Table 13).

Predictive Validity

We examined how well the IXL Universal Math Screener during the early portion of the 2024 academic year (August to November) correlates with the subsequent criterion measured at least 90 days later (i.e., between February and May 2025). We found strong positive correlations between XL Universal Math Screener scores and subsequent MAP Growth math scores in Kindergarten through 8th grade (see Table 12).

Table 12. Concurrent and Predictive analysis by grade

	Concurrent			Predictive		
Grade	*N*	*r*	[95% CI]	*N*	*r*	[95% CI]
Kindergarten	223	0.67	[0.60 - 0.74]	27	0.82	[0.65 - 0.91]
Grade 1	563	0.65	[0.60 - 0.70]	267	0.69	[0.62 - 0.74]
Grade 2	322	0.70	[0.69 - 0.79]	353	0.75	[0.70 - 0.79]
Grade 3	1,091	0.77	[0.74 - 0.79]	249	0.72	[0.66 - 0.78]
Grade 4	1,229	0.84	[0.82 - 0.85]	278	0.80	[0.75 - 0.84]
Grade 5	1,267	0.82	[0.80 - 0.83]	287	0.77	[0.72 - 0.81]
Grade 6	1,746	0.81	[0.80 - 0.83]	762	0.70	[0.67 - 0.74]
Grade 7	1,476	0.82	[0.80 - 0.84]	669	0.86	[0.84 - 0.88]
Grade 8	1,165	0.78	[0.76 - 0.80]	571	0.78	[0.75 - 0.81]

Classification Accuracy

Grade-level placements represent a primary use of the IXL Universal Math Screener, translating continuous scale scores into actionable categories that inform instructional decisions and intervention planning. Evaluating the accuracy of these classifications therefore sits at the intersection of reliability and validity, measuring how consistently the assessment assigns students to categories across independent administrations, and whether those category assignments support the intended interpretations and uses of scores.

These analyses generally rely on an external criterion measure to evaluate sensitivity, specificity, and area under the curve (AUC). These statistics are defined as follows:

1. Sensitivity (true positive rate) is the probability that the IXL Universal Math Screener correctly classifies a student as proficient, provided the student reached proficiency on the external criterion measure.
2. Specificity (true negative rate) is the probability that the IXL Universal Math Screener correctly classifies a student as not proficient, provided the student did not reach proficiency on the external criterion measure.
3. AUC measures how well scores on the IXL Universal Math Screener classify students as proficient and not proficient, matching those from the criterion measure cut scores, and serves as an overall indication of classification accuracy.

Values closer to 1.0 indicate stronger accuracy, and values at 0.50 indicate the measure is no better than chance at distinguishing student proficiency.

Data from the 2024-25 and 2025-26 academic years were combined and then disaggregated by BOY, MOY, and EOY delivery windows. Cut points on the MAP Growth math assessment were

determined using the 2025 nationwide MAP Growth norms for student achievement. In line with NCII TRC guidance, students with RIT scores below the 20th percentile were considered “at risk,” while those with RIT scores at or above the 20th percentile were considered “not at risk.” See Table 13 for sensitivity, specificity, and AUC [95% CI].

Table 13. Classification accuracy by grade and delivery window

Grade	*N*	Sensitivity	Specificity	AUC	[95% CI]
BOY					
Kindergarten	389	0.74	0.76	0.71	[0.63 - 0.84]
Grade 1	552	0.85	0.93	0.76	[0.80 - 0.90]
Grade 2	849	0.72	0.74	0.86	[0.66 - 0.78]
Grade 3	1276	0.79	0.75	0.85	[0.76 - 0.84]
Grade 4	1171	0.83	0.75	0.89	[0.79 - 0.86]
Grade 5	1226	0.85	0.95	0.75	[0.83 - 0.87]
Grade 6	807	0.83	0.90	0.75	[0.80 - 0.86]
Grade 7	842	0.87	0.93	0.78	[0.83 - 0.88]
Grade 8	775	0.87	0.94	0.80	[0.84 - 0.90]
MOY					
Kindergarten	258	0.77	0.69	0.86	[0.70 - 0.84]
Grade 1	326	0.72	0.73	0.70	[0.65 - 0.78]
Grade 2	1298	0.81	0.77	0.84	[0.76 - 0.86]
Grade 3	2028	0.83	0.89	0.77	[0.80 - 0.86]
Grade 4	2027	0.83	0.89	0.78	[0.81 - 0.86]
Grade 5	1807	0.83	0.81	0.85	[0.80 - 0.86]
Grade 6	1497	0.78	0.70	0.87	[0.74 - 0.82]
Grade 7	1201	0.75	0.06	0.91	[0.71 - 0.79]
Grade 8	713	0.81	0.72	0.90	[0.76 - 0.86]
EOY					
Kindergarten	73	0.73	0.73	0.72	[0.60 - 0.86]
Grade 1	127	0.71	0.70	0.73	[0.61 - 0.82]
Grade 2	521	0.75	0.64	0.85	[0.70 - 0.80]
Grade 3	616	0.78	0.79	0.77	[0.74 - 0.82]
Grade 4	553	0.79	0.76	0.82	[0.75 - 0.84]
Grade 5	612	0.77	0.66	0.88	[0.73 - 0.81]
Grade 6	319	0.77	0.67	0.88	[0.70 - 0.84]
Grade 7	268	0.79	0.71	0.87	[0.73 - 0.86]
Grade 8	321	0.78	0.70	0.87	[0.73 - 0.84]

REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA), & National Council for Measurement in Education (NCME). (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (pp. 508–600). American Council on Education.

Bernstein, I., Samuels, E., Woo, A., & Hagge, S. L. (2013). Assessing DIF Among Small Samples with Separate Calibration t and Mantel-Haenszel x2 Statistics in the Rasch Model. *Journal of Applied Measurement*, *14*(4), 389–399.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431–444.

Bonifay, W. E., Reise, S. P., Scheines, R., & Meijer, R. R. (2015). When Are Multidimensional Data Unidimensional Enough for Structural Equation Modeling? An Evaluation of the DETECT Multidimensionality Index. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(4), 504–516.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.

Cheng, Y., & Chang, H.-H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *62*(2), 369–383.

Cicchetti, D. V. (1994). Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. *Psychological Assessment*, *6*(4), 284–290.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Dimitrov, D. M. (2003). Marginal True-Score Measures and Reliability for Binary Items as a Function of Their IRT Parameters. *Applied Psychological Measurement*, *27*(6), 440–458.

Ferrara, S., Lai, E., Reilly, A., & Nichols, P. D. (2016). Principled Approaches to Assessment Design, Development, and Implementation. In A. A. Rupp & J. P. Leighton (Eds.), *The Handbook of Cognition and Assessment* (pp. 41–74). John Wiley & Sons, Inc.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 29–48.

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Hambelton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Kluwer-Nijoff Publishing.

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for Selecting Items for Computerized Adaptive Tests. *Applied Measurement in Education*, *2*(4), 359–375.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guilford Press.

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling, and Linking*. Springer New York.

Lewis, D., & Cook, R. (2020). Embedded Standard Setting: Aligning Standard-Setting Methodology with Contemporary Assessment Design Principles. *Educational Measurement: Issues and Practice*, *39*(1), 8–21.

Lim, H., & Choe, E. M. (2023). Detecting differential item functioning in CAT using IRT residual DIF approach. *Journal of Educational Measurement*, *60*(4), 626-650.

Lim, H., Choe, E. M., & Han, K. T. (2022). A residual-based differential item functioning detection framework in item response theory. *Journal of Educational Measurement*, *59*(1), 80-104

Linacre, J. M., & Wright, B. D. (1989). Mantel-Haenszel DIF and PROX are Equivalent! *Rasch Measurement Transactions*, *3*(2), 52–53.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Erlbaum.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of the data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*(4), 719–748.

Paek, I., & Wilson, M. (2011). Formulating the Rasch Differential Item Functioning Model Under the Marginal Maximum Likelihood Estimation Context and Its Comparison With Mantel-Haenszel Procedure in Short Test and Small Sample Conditions. *Educational and Psychological Measurement*, *71*(6), 1023–1046.

Penfield, R., & Lam, T. C. (2000). Assessing Differential Item Functioning in Performance Assessment: Review and Recommendations. *Educational Measurement: Issues & Practice*, *19*(3), 5–15.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research.

Reckase, M. D. (1979). Unifactor Latent Trait Models Applied to Multifactor Tests: Results and Implications. *Journal of Educational Statistics*, *4*(3), 207–230.

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and Structural Coefficient Bias in Structural Equation Modeling: A Bifactor Perspective. *Educational and Psychological Measurement*, *73*(1), 5–26.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*(2), 137–150.

Roussos, L., & Stout, W. (1996). A Multidimensionality-based DIF Analysis Paradigm. *Applied Psychological Measurement*, *20*(4), 355–371.

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment Research & Evaluation*, *7*(14), 1–5. https://doi.org/10.7275/AN9M-2035

Rudner, L. M. (2005). Expected Classification Accuracy. *Practical Assessment Research & Evaluation*, *10*(13).

Samejima, F. (1977). A Use of the Information Function in Tailored Testing. *Applied Psychological Measurement*, *1*(2), 233–247.

Samejima, F. (1994). Estimation of Reliability Coefficients Using the Test Information Function and Its Modifications. *Applied Psychological Measurement*, *18*(3), 229–244.

Schulz, E. M., Perlman, C., Rice, W. K., & Wright, B. D. (1996). An Empirical Comparison of Rasch and Mantel-Haenszel Procedures for Assessing Differential Item Functioning. In G. Engelhard & M. Wilson (Eds.), *Objective Measurement: Theory into Practice* (Vol. 3, pp. 65–82). Ablex.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi experimental designs for generalized causal inference.* Houghton Mifflin and Company.

Smith, R. M. (1996). A Comparison of the Rasch Separate Calibration and Between-Fit Methods of Detecting Item Bias. *Educational & Psychological Measurement*, *56*(3), 403–418.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal Design Applied to Large Scale Assessments* (Synthesis Report 44). University of Minnesota, National Center on Educational Outcomes (NCEO).

Wright, B. D., & Douglas, G. A. (1975). Best Test Design and Self-Tailored Testing. *Memo No. 19 MESA Psychometric Laboratory University of Chicago*.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best Test Design*. MESA Press.

Wyse, A. E., & Hao, S. (2012). An Evaluation of Item Response Theory Classification Accuracy and Consistency Indices. *Applied Psychological Measurement*, *36*(7), 602–624.

APPENDIX A: TEST CONSTRAINTS

This appendix provides the lower and upper bounds on the number of items for each grade across strands and time of year. The lower bound represents the minimum number of items a student will see at a specific time of year, and the upper bound represents the maximum number of items a student might see. Lower bounds must always be met for all students; however, due to the adaptive nature of the IXL Universal Math Screener the upper bound is rarely met. Typically, a student's assessment event will close once the lower bounds are met, as the stopping rules for score precision are often attained prior to delivery of the minimum number of items.

Table A1. Kindergarten min and max item delivery by strand and window

	BOY		MOY		EOY	
Strand	Lower bound	Upper bound	Lower bound	Upper bound	Lower bound	Upper bound
Algebra & Algebraic Thinking	0	0	3	4	3	6
Numbers & Operations	4	10	4	10	6	12
Measurement & Data	4	8	4	8	3	7
Geometry	4	10	4	10	6	10
Fractions	0	0	0	0	0	0

Table A2. First grade min and max item delivery by strand and window

	BOY		MOY		EOY	
Strand	Lower bound	Upper bound	Lower bound	Upper bound	Lower bound	Upper bound
Algebra & Algebraic Thinking	3	6	5	9	5	8
Numbers & Operations	6	12	6	12	7	15
Measurement & Data	3	7	3	7	4	9
Geometry	6	10	4	8	5	9
Fractions	0	0	0	0	0	0

Table A3. Second grade min and max item delivery by strand and window

Strand	BOY		MOY		EOY	
	Lower bound	Upper bound	Lower bound	Upper bound	Lower bound	Upper bound
Algebra & Algebraic Thinking	5	8	5	8	5	10
Numbers & Operations	7	15	7	15	7	15
Measurement & Data	4	9	4	9	6	10
Geometry	5	9	5	9	3	4
Fractions	0	0	0	4	0	5

Table A4. Third grade min and max item delivery by strand and window

Strand	BOY		MOY		EOY	
	Lower bound	Upper bound	Lower bound	Upper bound	Lower bound	Upper bound
Algebra & Algebraic Thinking	5	10	5	10	6	10
Numbers & Operations	7	15	7	15	6	15
Measurement & Data	6	10	6	10	5	10
Geometry	3	4	3	6	3	5
Fractions	0	5	3	6	4	8

Table A5. Fourth grade min and max item delivery by strand and window

Strand	BOY		MOY		EOY	
	Lower bound	Upper bound	Lower bound	Upper bound	Lower bound	Upper bound
Algebra & Algebraic Thinking	6	10	6	10	6	8
Numbers & Operations	6	15	6	15	6	15
Measurement & Data	5	10	5	10	4	8
Geometry	3	5	3	5	4	7
Fractions	4	8	4	8	4	10

Table A6. Fifth grade min and max item delivery by strand and window

Strand	BOY		MOY		EOY	
	Lower bound	Upper bound	Lower bound	Upper bound	Lower bound	Upper bound
Algebra & Algebraic Thinking	6	8	6	8	5	10
Numbers & Operations	6	15	6	15	6	12
Measurement & Data	4	8	4	8	4	8
Geometry	4	7	4	7	4	8
Fractions	4	10	4	10	5	9

Table A7. Sixth grade min and max item delivery by strand and window

Strand	BOY		MOY		EOY	
	Lower bound	Upper bound	Lower bound	Upper bound	Lower bound	Upper bound
Algebra & Algebraic Thinking	5	10	5	10	5	12
Numbers & Operations	6	12	6	12	5	10
Measurement & Data	4	8	4	8	4	9
Geometry	4	8	4	8	4	8
Fractions	5	9	5	9	5	10

Table A8. Seventh grade min and max item delivery by strand and window

Strand	BOY		MOY		EOY	
	Lower bound	Upper bound	Lower bound	Upper bound	Lower bound	Upper bound
Algebra & Algebraic Thinking	5	12	5	12	7	13
Numbers & Operations	5	10	5	10	5	10
Measurement & Data	4	9	4	9	3	8
Geometry	4	8	4	8	4	9
Fractions	5	10	5	10	3	9

Table A9. Eighth grade min and max item delivery by strand and window

Strand	BOY		MOY		EOY	
	Lower bound	Upper bound	Lower bound	Upper bound	Lower bound	Upper bound
Algebra & Algebraic Thinking	7	13	7	13	7	13
Numbers & Operations	5	10	5	10	5	10
Measurement & Data	3	8	4	8	4	8
Geometry	4	9	4	9	4	9
Fractions	3	9	4	9	4	9

APPENDIX B: NORMING TABLES

Table B1. Beginning-of-Year norming theta values

Percentile	K	1	2	3	4	5	6	7	8
1	-5.8	-5.4	-4.8	-4.5	-4.1	-3.7	-3.7	-3.5	-3.2
2	-5.3	-4.6	-4.2	-3.8	-3.5	-3.2	-3.0	-2.9	-2.9
3	-5.0	-4.4	-3.8	-3.5	-3.1	-2.9	-2.8	-2.5	-2.5
4	-4.8	-4.3	-3.6	-3.2	-2.8	-2.6	-2.6	-2.4	-2.4
5	-4.7	-4.1	-3.5	-3.0	-2.6	-2.5	-2.5	-2.2	-2.1
6	-4.6	-4.0	-3.3	-2.9	-2.4	-2.3	-2.3	-2.1	-2.0
7	-4.5	-3.8	-3.1	-2.7	-2.2	-2.2	-2.2	-1.9	-1.9
8	-4.4	-3.7	-3.0	-2.6	-2.1	-2.0	-2.1	-1.8	-1.7
9	-4.3	-3.6	-2.9	-2.4	-2.0	-1.9	-2.0	-1.8	-1.6
10	-4.2	-3.5	-2.8	-2.3	-1.9	-1.8	-1.9	-1.7	-1.6
11	-4.1	-3.4	-2.7	-2.2	-1.8	-1.7	-1.9	-1.6	-1.5
12	-4.0	-3.4	-2.6	-2.1	-1.7	-1.6	-1.8	-1.5	-1.4
13	-3.9	-3.3	-2.5	-2.1	-1.7	-1.6	-1.7	-1.4	-1.3
14	-3.8	-3.3	-2.5	-2.0	-1.6	-1.5	-1.7	-1.3	-1.2
15	-3.7	-3.3	-2.4	-1.9	-1.5	-1.4	-1.6	-1.3	-1.2
16	-3.7	-3.3	-2.4	-1.9	-1.5	-1.4	-1.5	-1.2	-1.1
17	-3.7	-3.2	-2.3	-1.8	-1.4	-1.3	-1.5	-1.1	-1.0
18	-3.7	-3.1	-2.3	-1.8	-1.4	-1.2	-1.4	-1.1	-1.0
19	-3.6	-3.1	-2.2	-1.7	-1.3	-1.2	-1.4	-1.0	-0.9
20	-3.6	-3.0	-2.1	-1.7	-1.2	-1.1	-1.3	-0.9	-0.9
21	-3.6	-3.0	-2.1	-1.7	-1.2	-1.1	-1.3	-0.9	-0.8
22	-3.5	-2.9	-2.0	-1.7	-1.1	-1.0	-1.2	-0.8	-0.8
23	-3.5	-2.9	-2.0	-1.6	-1.1	-1.0	-1.2	-0.8	-0.7
24	-3.4	-2.8	-2.0	-1.5	-1.0	-0.9	-1.2	-0.7	-0.6
25	-3.4	-2.8	-1.9	-1.5	-1.0	-0.9	-1.1	-0.7	-0.6

Table B1. Beginning-of-Year norming theta values

Percentile	K	1	2	3	4	5	6	7	8
26	-3.4	-2.8	-1.9	-1.5	-0.9	-0.8	-1.1	-0.7	-0.5
27	-3.3	-2.7	-1.8	-1.4	-0.9	-0.8	-1.0	-0.6	-0.5
28	-3.3	-2.7	-1.8	-1.4	-0.8	-0.8	-1.0	-0.6	-0.5
29	-3.3	-2.7	-1.7	-1.3	-0.8	-0.7	-0.9	-0.5	-0.4
30	-3.3	-2.6	-1.7	-1.3	-0.8	-0.7	-0.9	-0.5	-0.4
31	-3.2	-2.6	-1.7	-1.3	-0.7	-0.6	-0.9	-0.5	-0.3
32	-3.2	-2.6	-1.6	-1.2	-0.7	-0.6	-0.8	-0.4	-0.3
33	-3.2	-2.5	-1.6	-1.2	-0.6	-0.6	-0.8	-0.4	-0.2
34	-3.2	-2.5	-1.6	-1.2	-0.6	-0.5	-0.7	-0.3	-0.2
35	-3.2	-2.5	-1.6	-1.1	-0.6	-0.5	-0.7	-0.3	-0.1
36	-3.2	-2.4	-1.5	-1.1	-0.5	-0.5	-0.7	-0.2	-0.1
37	-3.1	-2.4	-1.5	-1.1	-0.5	-0.4	-0.6	-0.2	-0.1
38	-3.1	-2.4	-1.5	-1.1	-0.4	-0.4	-0.6	-0.2	0.0
39	-3.1	-2.4	-1.4	-1.0	-0.4	-0.4	-0.5	-0.1	0.0
40	-3.1	-2.3	-1.4	-1.0	-0.4	-0.3	-0.5	-0.1	0.1
41	-3.0	-2.3	-1.4	-1.0	-0.3	-0.3	-0.5	-0.1	0.1
42	-3.0	-2.3	-1.4	-0.9	-0.3	-0.3	-0.4	0.0	0.1
43	-2.8	-2.2	-1.3	-0.9	-0.3	-0.2	-0.4	0.0	0.2
44	-2.8	-2.2	-1.3	-0.9	-0.3	-0.2	-0.3	0.1	0.2
45	-2.8	-2.2	-1.3	-0.9	-0.2	-0.2	-0.3	0.1	0.3
46	-2.7	-2.2	-1.2	-0.9	-0.2	-0.1	-0.3	0.1	0.3
47	-2.7	-2.1	-1.2	-0.8	-0.2	-0.1	-0.2	0.2	0.3
48	-2.7	-2.1	-1.2	-0.8	-0.1	-0.1	-0.2	0.2	0.4
49	-2.7	-2.1	-1.1	-0.8	-0.1	0.0	-0.1	0.3	0.4
50	-2.7	-2.0	-1.1	-0.8	-0.1	0.0	-0.1	0.3	0.4
51	-2.7	-2.0	-1.1	-0.7	-0.1	0.0	-0.1	0.3	0.5
52	-2.7	-2.0	-1.0	-0.7	0.0	0.0	0.0	0.4	0.5
53	-2.6	-1.9	-1.0	-0.7	0.0	0.1	0.0	0.4	0.6
54	-2.6	-1.9	-1.0	-0.7	0.0	0.1	0.0	0.5	0.6
55	-2.6	-1.9	-1.0	-0.7	0.0	0.1	0.1	0.5	0.6
56	-2.6	-1.9	-0.9	-0.6	0.1	0.1	0.1	0.5	0.6

Table B1. Beginning-of-Year norming theta values

Percentile	K	1	2	3	4	5	6	7	8
57	-2.4	-1.8	-0.9	-0.6	0.1	0.1	0.1	0.6	0.6
58	-2.4	-1.8	-0.9	-0.6	0.1	0.2	0.2	0.6	0.7
59	-2.3	-1.8	-0.8	-0.6	0.1	0.2	0.2	0.6	0.8
60	-2.3	-1.7	-0.8	-0.5	0.2	0.2	0.3	0.7	0.8
61	-2.3	-1.6	-0.8	-0.5	0.2	0.2	0.3	0.7	0.8
62	-2.3	-1.6	-0.7	-0.5	0.2	0.3	0.3	0.8	0.9
63	-2.3	-1.6	-0.7	-0.4	0.2	0.3	0.4	0.8	0.9
64	-2.2	-1.5	-0.7	-0.4	0.3	0.3	0.4	0.9	1.0
65	-2.2	-1.5	-0.6	-0.4	0.3	0.3	0.4	0.9	1.0
66	-2.2	-1.5	-0.6	-0.4	0.3	0.4	0.5	1.0	1.0
67	-2.2	-1.5	-0.6	-0.3	0.4	0.4	0.5	1.0	1.1
68	-2.2	-1.5	-0.6	-0.3	0.4	0.4	0.6	1.1	1.1
69	-2.2	-1.5	-0.5	-0.3	0.4	0.5	0.6	1.1	1.2
70	-2.1	-1.5	-0.5	-0.3	0.4	0.5	0.7	1.1	1.2
71	-2.0	-1.4	-0.5	-0.2	0.5	0.5	0.7	1.2	1.3
72	-1.9	-1.4	-0.4	-0.2	0.5	0.5	0.7	1.2	1.3
73	-1.9	-1.4	-0.4	-0.2	0.5	0.6	0.8	1.3	1.3
74	-1.9	-1.3	-0.4	-0.1	0.6	0.6	0.8	1.3	1.4
75	-1.8	-1.3	-0.3	-0.1	0.6	0.6	0.9	1.4	1.4
76	-1.8	-1.3	-0.3	0.0	0.6	0.7	0.9	1.4	1.5
77	-1.8	-1.2	-0.3	0.0	0.6	0.7	1.0	1.5	1.5
78	-1.8	-1.2	-0.2	0.1	0.7	0.7	1.0	1.5	1.6
79	-1.8	-1.1	-0.2	0.2	0.7	0.7	1.1	1.5	1.6
80	-1.8	-1.1	-0.1	0.2	0.7	0.8	1.1	1.6	1.7
81	-1.7	-1.0	-0.1	0.3	0.8	0.8	1.2	1.6	1.7
82	-1.5	-1.0	0.0	0.3	0.8	0.9	1.2	1.7	1.8
83	-1.4	-0.9	0.0	0.3	0.8	0.9	1.3	1.7	1.8
84	-1.4	-0.9	0.0	0.4	0.9	1.0	1.3	1.8	1.9
85	-1.4	-0.8	0.0	0.5	0.9	1.0	1.4	1.8	2.0
86	-1.4	-0.7	0.1	0.5	0.9	1.1	1.4	1.9	2.0
87	-1.3	-0.6	0.1	0.6	1.0	1.1	1.5	1.9	2.1

Table B1. Beginning-of-Year norming theta values

Percentile	K	1	2	3	4	5	6	7	8
88	-1.3	-0.6	0.1	0.7	1.0	1.2	1.5	2.0	2.2
89	-1.3	-0.5	0.2	0.8	1.0	1.2	1.6	2.1	2.2
90	-1.0	-0.4	0.2	0.9	1.1	1.3	1.7	2.1	2.3
91	-0.9	-0.3	0.3	1.0	1.2	1.3	1.7	2.3	2.4
92	-0.9	-0.3	0.4	1.1	1.2	1.4	1.8	2.3	2.5
93	-0.8	-0.1	0.4	1.2	1.3	1.4	1.9	2.4	2.6
94	-0.8	0.0	0.5	1.3	1.4	1.5	2.0	2.6	2.8
95	-0.4	0.2	0.6	1.4	1.5	1.6	2.1	2.7	2.9
96	-0.2	0.3	0.8	1.6	1.6	1.8	2.2	2.9	3.0
97	-0.2	0.6	0.9	1.8	1.8	1.9	2.4	3.1	3.2
98	0.8	0.9	1.2	2.1	1.9	2.1	2.7	3.4	3.5
99	1.0	1.6	1.6	2.5	2.3	2.4	3.2	3.7	4.0

Table B2. Middle-of-Year norming theta values

Percentile	K	1	2	3	4	5	6	7	8
1	-5.4	-5.0	-4.5	-3.8	-3.8	-3.2	-5.1	-9.4	-9.3
2	-5.0	-4.6	-3.8	-3.2	-3.2	-2.7	-3.5	-3.5	-3.2
3	-4.7	-4.3	-3.4	-2.9	-2.7	-2.4	-3.1	-2.8	-2.6
4	-4.6	-4.0	-3.2	-2.7	-2.3	-2.2	-2.8	-2.5	-2.5
5	-4.4	-3.8	-3.0	-2.5	-2.1	-2.0	-2.7	-2.4	-2.2
6	-4.3	-3.6	-2.9	-2.3	-1.9	-1.9	-2.4	-2.2	-2.1
7	-4.2	-3.5	-2.7	-2.2	-1.8	-1.7	-2.3	-2.1	-1.9
8	-4.2	-3.4	-2.6	-2.1	-1.7	-1.6	-2.1	-2.0	-1.7
9	-4.1	-3.2	-2.6	-2.0	-1.5	-1.5	-2.0	-1.9	-1.6
10	-4.0	-3.1	-2.5	-1.9	-1.4	-1.4	-1.9	-1.8	-1.5
11	-4.0	-3.0	-2.5	-1.8	-1.3	-1.4	-1.9	-1.7	-1.4
12	-3.9	-3.0	-2.3	-1.7	-1.2	-1.3	-1.8	-1.6	-1.3
13	-3.8	-2.9	-2.2	-1.6	-1.2	-1.2	-1.7	-1.5	-1.2
14	-3.7	-2.8	-2.2	-1.6	-1.1	-1.2	-1.6	-1.4	-1.2
15	-3.7	-2.8	-2.1	-1.5	-1.1	-1.1	-1.5	-1.4	-1.1
16	-3.6	-2.7	-2.1	-1.4	-1.0	-1.1	-1.4	-1.3	-1.0
17	-3.6	-2.7	-2.0	-1.4	-1.0	-1.0	-1.4	-1.2	-0.9
18	-3.5	-2.6	-2.0	-1.3	-0.9	-1.0	-1.3	-1.2	-0.9
19	-3.5	-2.6	-1.9	-1.3	-0.9	-0.9	-1.3	-1.1	-0.8
20	-3.4	-2.5	-1.9	-1.2	-0.8	-0.9	-1.2	-1.1	-0.7
21	-3.4	-2.5	-1.8	-1.2	-0.8	-0.8	-1.2	-1.0	-0.7
22	-3.3	-2.5	-1.8	-1.1	-0.7	-0.8	-1.1	-1.0	-0.6
23	-3.3	-2.4	-1.8	-1.1	-0.7	-0.8	-1.1	-0.9	-0.5
24	-3.2	-2.4	-1.7	-1.1	-0.6	-0.7	-1.0	-0.9	-0.5
25	-3.2	-2.4	-1.7	-1.0	-0.6	-0.7	-1.0	-0.8	-0.4
26	-3.1	-2.3	-1.7	-1.0	-0.5	-0.6	-0.9	-0.8	-0.4
27	-3.1	-2.3	-1.6	-0.9	-0.5	-0.6	-0.9	-0.7	-0.3
28	-3.0	-2.3	-1.6	-0.9	-0.4	-0.6	-0.8	-0.7	-0.2
29	-3.0	-2.2	-1.6	-0.9	-0.4	-0.5	-0.8	-0.6	-0.2
30	-3.0	-2.2	-1.5	-0.8	-0.4	-0.5	-0.8	-0.6	-0.1

Table B2. Middle-of-Year norming theta values

Percentile	K	1	2	3	4	5	6	7	8
31	-2.9	-2.2	-1.5	-0.8	-0.3	-0.4	-0.7	-0.5	-0.1
32	-2.9	-2.1	-1.5	-0.8	-0.3	-0.4	-0.7	-0.5	0.0
33	-2.9	-2.1	-1.5	-0.7	-0.3	-0.3	-0.7	-0.4	0.0
34	-2.8	-2.1	-1.4	-0.7	-0.2	-0.3	-0.6	-0.4	0.1
35	-2.8	-2.1	-1.4	-0.7	-0.2	-0.2	-0.6	-0.4	0.1
36	-2.8	-2.0	-1.4	-0.7	-0.2	-0.2	-0.5	-0.3	0.2
37	-2.7	-2.0	-1.4	-0.6	-0.1	-0.2	-0.5	-0.3	0.2
38	-2.7	-2.0	-1.3	-0.6	-0.1	-0.1	-0.5	-0.3	0.3
39	-2.6	-1.9	-1.3	-0.6	-0.1	-0.1	-0.5	-0.2	0.3
40	-2.6	-1.9	-1.3	-0.6	-0.1	-0.1	-0.4	-0.2	0.4
41	-2.6	-1.9	-1.3	-0.5	0.0	-0.1	-0.4	-0.1	0.4
42	-2.6	-1.9	-1.3	-0.5	0.0	0.0	-0.3	-0.1	0.5
43	-2.5	-1.8	-1.2	-0.5	0.0	0.0	-0.3	0.0	0.5
44	-2.5	-1.8	-1.2	-0.4	0.1	0.0	-0.2	0.0	0.6
45	-2.5	-1.8	-1.2	-0.4	0.1	0.1	-0.2	0.1	0.6
46	-2.5	-1.8	-1.2	-0.4	0.1	0.1	-0.2	0.1	0.6
47	-2.5	-1.7	-1.1	-0.4	0.1	0.1	-0.1	0.2	0.7
48	-2.5	-1.7	-1.1	-0.3	0.1	0.1	-0.1	0.2	0.7
49	-2.5	-1.7	-1.1	-0.3	0.2	0.2	-0.1	0.2	0.8
50	-2.5	-1.7	-1.1	-0.3	0.2	0.2	0.0	0.3	0.8
51	-2.4	-1.6	-1.1	-0.3	0.2	0.2	0.0	0.4	0.9
52	-2.4	-1.6	-1.0	-0.2	0.3	0.3	0.1	0.4	0.9
53	-2.4	-1.6	-1.0	-0.2	0.3	0.3	0.1	0.4	0.9
54	-2.4	-1.6	-1.0	-0.2	0.3	0.3	0.1	0.5	0.9
55	-2.4	-1.5	-1.0	-0.2	0.3	0.3	0.2	0.5	1.0
56	-2.4	-1.5	-1.0	-0.2	0.3	0.4	0.2	0.6	1.0
57	-2.4	-1.5	-0.9	-0.1	0.4	0.4	0.3	0.6	1.1
58	-2.4	-1.5	-0.9	-0.1	0.4	0.4	0.3	0.7	1.1
59	-2.4	-1.4	-0.9	-0.1	0.4	0.5	0.3	0.7	1.1
60	-2.3	-1.4	-0.9	-0.1	0.4	0.5	0.4	0.8	1.2
61	-2.3	-1.4	-0.9	0.0	0.4	0.5	0.4	0.8	1.2

Table B2. Middle-of-Year norming theta values

Percentile	K	1	2	3	4	5	6	7	8
62	-2.3	-1.4	-0.8	0.0	0.5	0.6	0.5	0.9	1.3
63	-2.3	-1.3	-0.8	0.0	0.5	0.6	0.5	0.9	1.3
64	-2.2	-1.3	-0.8	0.0	0.5	0.6	0.5	1.0	1.3
65	-2.2	-1.3	-0.8	0.0	0.5	0.7	0.6	1.0	1.4
66	-2.2	-1.3	-0.8	0.1	0.6	0.7	0.6	1.1	1.4
67	-2.2	-1.2	-0.7	0.1	0.6	0.7	0.7	1.1	1.5
68	-2.1	-1.2	-0.7	0.1	0.6	0.8	0.7	1.2	1.5
69	-2.1	-1.2	-0.7	0.1	0.6	0.8	0.8	1.2	1.6
70	-2.1	-1.1	-0.7	0.2	0.6	0.8	0.8	1.3	1.6
71	-2.1	-1.1	-0.7	0.2	0.7	0.9	0.9	1.3	1.7
72	-2.0	-1.1	-0.6	0.2	0.7	0.9	0.9	1.4	1.7
73	-2.0	-1.0	-0.6	0.2	0.7	0.9	1.0	1.5	1.7
74	-2.0	-1.0	-0.6	0.3	0.7	1.0	1.0	1.5	1.8
75	-2.0	-0.9	-0.6	0.3	0.7	1.0	1.1	1.5	1.8
76	-1.9	-0.8	-0.5	0.3	0.8	1.1	1.1	1.6	1.9
77	-1.9	-0.8	-0.5	0.3	0.8	1.1	1.2	1.6	1.9
78	-1.8	-0.7	-0.5	0.4	0.8	1.2	1.2	1.7	2.0
79	-1.8	-0.7	-0.4	0.4	0.8	1.2	1.3	1.7	2.0
80	-1.8	-0.7	-0.4	0.4	0.9	1.2	1.3	1.8	2.1
81	-1.7	-0.7	-0.4	0.5	0.9	1.3	1.4	1.8	2.2
82	-1.7	-0.6	-0.4	0.5	1.0	1.3	1.4	1.8	2.2
83	-1.7	-0.6	-0.3	0.5	1.0	1.4	1.5	1.9	2.3
84	-1.7	-0.6	-0.3	0.5	1.1	1.4	1.5	1.9	2.3
85	-1.6	-0.5	-0.2	0.6	1.1	1.5	1.6	2.0	2.4
86	-1.6	-0.5	-0.2	0.6	1.2	1.5	1.6	2.0	2.5
87	-1.5	-0.5	-0.2	0.7	1.2	1.6	1.7	2.1	2.5
88	-1.4	-0.4	-0.1	0.7	1.2	1.6	1.7	2.1	2.6
89	-1.4	-0.4	0.0	0.7	1.3	1.7	1.8	2.2	2.7
90	-1.3	-0.3	0.1	0.8	1.3	1.7	1.8	2.3	2.7
91	-1.3	-0.3	0.1	0.8	1.4	1.8	1.9	2.4	2.8
92	-1.2	-0.2	0.2	0.9	1.4	1.9	2.0	2.5	2.9

Table B2. Middle-of-Year norming theta values

Percentile	K	1	2	3	4	5	6	7	8
93	-1.1	-0.2	0.3	1.0	1.5	1.9	2.1	2.6	3.1
94	-1.0	-0.1	0.4	1.0	1.6	2.0	2.2	2.7	3.2
95	-0.9	0.0	0.5	1.1	1.7	2.1	2.2	2.8	3.3
96	-0.7	0.1	0.7	1.2	1.8	2.2	2.4	2.9	3.5
97	-0.4	0.2	0.9	1.3	1.9	2.5	2.6	3.1	3.8
98	-0.2	0.4	1.2	1.6	2.1	2.8	2.8	3.2	4.0
99	0.4	0.7	1.6	2.0	2.5	3.5	3.2	3.6	4.7

Table B3. End-of-Year norming theta values

Percentile	K	1	2	3	4	5	6	7	8
1	-5.2	-4.8	-4.5	-4.3	-3.6	-3.4	-4.4	-9.3	-9.2
2	-4.6	-4.2	-3.9	-3.7	-3.1	-2.9	-3.4	-3.9	-3.7
3	-4.4	-4.0	-3.5	-3.3	-2.7	-2.6	-2.9	-3.1	-2.7
4	-4.3	-3.7	-3.2	-3.0	-2.4	-2.3	-2.8	-3.0	-2.4
5	-4.2	-3.6	-3.0	-2.7	-2.2	-2.1	-2.5	-2.7	-2.1
6	-4.1	-3.4	-2.9	-2.5	-2.0	-2.0	-2.4	-2.5	-1.9
7	-4.0	-3.4	-2.7	-2.4	-1.8	-1.9	-2.2	-2.4	-1.7
8	-3.9	-3.1	-2.5	-2.2	-1.6	-1.8	-2.1	-2.4	-1.6
9	-3.7	-3.0	-2.4	-2.0	-1.5	-1.6	-2.0	-2.2	-1.5
10	-3.7	-2.9	-2.3	-1.9	-1.4	-1.5	-1.9	-2.1	-1.3
11	-3.6	-2.8	-2.3	-1.8	-1.3	-1.5	-1.8	-2.0	-1.2
12	-3.5	-2.7	-2.1	-1.7	-1.3	-1.4	-1.7	-1.9	-1.1
13	-3.4	-2.6	-2.1	-1.6	-1.2	-1.3	-1.6	-1.8	-1.0
14	-3.4	-2.5	-2.0	-1.5	-1.1	-1.2	-1.5	-1.7	-1.0
15	-3.3	-2.5	-1.9	-1.5	-1.1	-1.2	-1.5	-1.7	-0.9
16	-3.3	-2.4	-1.8	-1.4	-1.0	-1.1	-1.4	-1.6	-0.8
17	-3.3	-2.4	-1.8	-1.3	-1.0	-1.1	-1.3	-1.5	-0.8
18	-3.3	-2.3	-1.8	-1.2	-0.9	-1.0	-1.2	-1.4	-0.7
19	-3.3	-2.3	-1.7	-1.2	-0.8	-1.0	-1.1	-1.4	-0.7
20	-3.2	-2.2	-1.7	-1.1	-0.8	-0.9	-1.1	-1.3	-0.6
21	-3.1	-2.2	-1.7	-1.0	-0.7	-0.8	-1.0	-1.3	-0.5
22	-3.1	-2.2	-1.7	-1.0	-0.7	-0.8	-0.9	-1.2	-0.5
23	-3.0	-2.1	-1.6	-0.9	-0.6	-0.7	-0.9	-1.1	-0.4
24	-3.0	-2.0	-1.5	-0.9	-0.6	-0.7	-0.8	-1.1	-0.3
25	-2.9	-2.0	-1.5	-0.8	-0.6	-0.6	-0.8	-1.0	-0.3
26	-2.9	-2.0	-1.4	-0.8	-0.5	-0.6	-0.8	-0.9	-0.2
27	-2.8	-1.9	-1.4	-0.7	-0.5	-0.5	-0.7	-0.9	-0.2
28	-2.8	-1.9	-1.4	-0.7	-0.4	-0.5	-0.7	-0.8	-0.1
29	-2.8	-1.9	-1.3	-0.6	-0.4	-0.4	-0.6	-0.8	-0.1
30	-2.8	-1.8	-1.3	-0.6	-0.4	-0.4	-0.6	-0.7	0.0

Table B3. End-of-Year norming theta values

Percentile	K	1	2	3	4	5	6	7	8
31	-2.7	-1.8	-1.2	-0.5	-0.3	-0.3	-0.5	-0.7	0.0
32	-2.7	-1.7	-1.2	-0.5	-0.3	-0.3	-0.5	-0.6	0.1
33	-2.7	-1.7	-1.2	-0.5	-0.3	-0.2	-0.4	-0.6	0.1
34	-2.6	-1.7	-1.2	-0.4	-0.2	-0.2	-0.4	-0.5	0.2
35	-2.6	-1.6	-1.1	-0.4	-0.2	-0.1	-0.4	-0.5	0.2
36	-2.6	-1.6	-1.1	-0.3	-0.1	-0.1	-0.3	-0.4	0.3
37	-2.6	-1.6	-1.1	-0.3	-0.1	0.0	-0.3	-0.4	0.3
38	-2.5	-1.6	-1.1	-0.2	-0.1	0.0	-0.3	-0.3	0.4
39	-2.5	-1.5	-1.0	-0.2	0.0	0.1	-0.2	-0.3	0.4
40	-2.5	-1.5	-1.0	-0.1	0.0	0.1	-0.2	-0.2	0.5
41	-2.5	-1.5	-1.0	-0.1	0.0	0.2	-0.1	-0.2	0.5
42	-2.4	-1.5	-1.0	-0.1	0.1	0.2	-0.1	-0.1	0.6
43	-2.4	-1.4	-0.9	-0.1	0.1	0.3	0.0	0.0	0.6
44	-2.4	-1.4	-0.9	0.0	0.1	0.3	0.0	0.0	0.7
45	-2.3	-1.4	-0.9	0.0	0.1	0.3	0.0	0.1	0.7
46	-2.3	-1.4	-0.9	0.0	0.2	0.4	0.1	0.1	0.8
47	-2.3	-1.3	-0.8	0.1	0.2	0.4	0.1	0.2	0.8
48	-2.2	-1.3	-0.8	0.1	0.2	0.5	0.2	0.2	0.9
49	-2.2	-1.3	-0.8	0.1	0.3	0.5	0.2	0.2	0.9
50	-2.2	-1.3	-0.8	0.2	0.3	0.6	0.3	0.3	0.9
51	-2.2	-1.2	-0.7	0.2	0.3	0.6	0.3	0.3	1.0
52	-2.1	-1.2	-0.7	0.2	0.3	0.6	0.4	0.4	1.0
53	-2.1	-1.2	-0.7	0.3	0.4	0.7	0.4	0.4	1.1
54	-2.1	-1.2	-0.7	0.3	0.4	0.7	0.5	0.5	1.1
55	-2.1	-1.1	-0.7	0.3	0.4	0.7	0.5	0.5	1.2
56	-2.0	-1.1	-0.6	0.3	0.5	0.8	0.6	0.6	1.2
57	-2.0	-1.1	-0.6	0.4	0.5	0.8	0.6	0.6	1.3
58	-2.0	-1.0	-0.6	0.4	0.5	0.8	0.7	0.6	1.3
59	-1.9	-1.0	-0.6	0.4	0.6	0.9	0.7	0.7	1.4
60	-1.9	-1.0	-0.5	0.5	0.6	0.9	0.8	0.8	1.4
61	-1.9	-1.0	-0.5	0.5	0.6	1.0	0.9	0.8	1.4

Table B3. End-of-Year norming theta values

Percentile	K	1	2	3	4	5	6	7	8
62	-1.8	-0.9	-0.5	0.5	0.6	1.0	0.9	0.8	1.5
63	-1.8	-0.9	-0.4	0.5	0.7	1.0	1.0	0.9	1.5
64	-1.8	-0.9	-0.4	0.6	0.7	1.1	1.0	0.9	1.6
65	-1.7	-0.8	-0.4	0.6	0.7	1.1	1.1	1.0	1.6
66	-1.6	-0.8	-0.4	0.6	0.7	1.1	1.1	1.0	1.7
67	-1.6	-0.8	-0.3	0.7	0.8	1.2	1.2	1.1	1.7
68	-1.5	-0.7	-0.3	0.7	0.8	1.2	1.2	1.1	1.8
69	-1.5	-0.7	-0.3	0.7	0.8	1.2	1.3	1.2	1.8
70	-1.5	-0.7	-0.2	0.7	0.9	1.3	1.3	1.2	1.9
71	-1.5	-0.6	-0.2	0.8	0.9	1.3	1.4	1.3	1.9
72	-1.5	-0.6	-0.2	0.8	1.0	1.4	1.5	1.3	2.0
73	-1.5	-0.6	-0.1	0.8	1.0	1.4	1.5	1.4	2.0
74	-1.4	-0.6	-0.1	0.8	1.1	1.4	1.6	1.4	2.1
75	-1.4	-0.5	0.0	0.9	1.1	1.5	1.6	1.5	2.1
76	-1.4	-0.5	0.0	0.9	1.1	1.5	1.7	1.5	2.2
77	-1.3	-0.4	0.1	0.9	1.2	1.6	1.8	1.6	2.2
78	-1.3	-0.4	0.1	1.0	1.2	1.6	1.8	1.7	2.3
79	-1.3	-0.4	0.2	1.0	1.2	1.6	1.9	1.7	2.3
80	-1.2	-0.4	0.2	1.0	1.3	1.7	1.9	1.8	2.4
81	-1.2	-0.3	0.3	1.1	1.3	1.7	2.0	1.9	2.5
82	-1.1	-0.3	0.3	1.1	1.3	1.8	2.0	1.9	2.5
83	-1.1	-0.2	0.4	1.1	1.4	1.8	2.1	2.0	2.6
84	-1.0	-0.2	0.4	1.2	1.4	1.9	2.2	2.1	2.7
85	-0.9	-0.2	0.5	1.2	1.4	2.0	2.3	2.1	2.8
86	-0.9	-0.1	0.6	1.3	1.5	2.0	2.4	2.2	2.9
87	-0.8	0.0	0.6	1.3	1.5	2.1	2.5	2.3	2.9
88	-0.7	0.0	0.7	1.4	1.6	2.1	2.6	2.4	3.0
89	-0.6	0.0	0.8	1.5	1.7	2.2	2.7	2.5	3.1
90	-0.6	0.1	0.9	1.5	1.7	2.3	2.8	2.6	3.3
91	-0.4	0.1	1.0	1.6	1.8	2.4	2.9	2.7	3.4
92	-0.3	0.2	1.1	1.7	1.9	2.5	3.0	2.8	3.5

Table B3. End-of-Year norming theta values

Percentile	K	1	2	3	4	5	6	7	8
93	-0.2	0.2	1.2	1.8	1.9	2.7	3.2	2.9	3.6
94	-0.1	0.3	1.3	1.9	2.0	2.8	3.3	3.0	3.8
95	0.0	0.4	1.5	2.0	2.1	3.0	3.5	3.1	3.9
96	0.2	0.5	1.6	2.1	2.2	3.2	3.8	3.3	4.0
97	0.3	0.7	1.9	2.3	2.5	3.5	4.2	3.5	4.3
98	0.6	1.0	2.1	2.6	2.8	3.8	4.6	3.8	4.8
99	0.8	1.3	2.7	3.1	3.3	4.5	4.9	4.3	5.1

APPENDIX C: RELIABILITY BY GRADE AND STRAND

This appendix reports marginal reliability coefficients by grade and strand, supplemental to the section on Internal Consistency. Despite being based on a small subset of items students answered, all values were at least 0.70, indicating at least acceptable internal consistency; many were at least 0.80, indicating robust reliability; and several were greater than 0.90, indicating excellent reliability.

Table C1. Marginal reliability by grade and strand

Strand	K	G1	G2	G3	G4	G5	G6	G7	G8
Algebra and algebraic thinking	0.82	0.83	0.80	0.83	0.84	0.83	0.83	0.84	0.86
Data and measurement	0.82	0.81	0.81	0.82	0.83	0.83	0.83	0.83	0.83
Geometry	0.83	0.81	0.79	0.81	0.82	0.82	0.82	0.83	0.83
Fractions	-	-	-	0.81	0.80	0.81	0.83	0.85	0.83
Number and Operations	0.83	0.85	0.83	0.83	0.80	0.81	0.84	0.84	0.85
Overall	*0.83*	*0.88*	*0.88*	*0.90*	*0.90*	*0.90*	*0.92*	*0.91*	*0.92*

APPENDIX D: FACTOR LOADINGS

This appendix reports the factor loadings from the Confirmatory Factor Analyses subsection. Across grades, the factor loadings tended to trend upward, suggesting that the general math ability factor became more robust as students' general math knowledge improved. Given that the purpose of IXL Universal Math Screener is to classify students according to their ability as quickly as possible, it is unsurprising that some of the factor loadings were somewhat low. However, the purpose of these CFAs was to gather additional evidence supporting the use of a unidimensional Rasch model to measure math ability, not to modify the model until we found the best statistical fit.

Table D1. Factor loadings by grade and strand

Grade	Algebra and algebraic thinking	Data and measurement	Fractions	Geometry	Numbers and operations
Kindergarten	0.670	0.448	-	0.326	0.646
Grade 1	0.648	0.434	-	0.394	0.809
Grade 2	0.628	0.580	-	0.410	0.759
Grade 3	0.685	0.561	0.176	0.406	0.629
Grade 4	0.762	0.562	0.307	0.472	0.543
Grade 5	0.716	0.561	0.437	0.480	0.601
Grade 6	0.662	0.599	0.557	0.456	0.663
Grade 7	0.659	0.597	0.513	0.466	0.596
Grade 8	0.719	0.509	0.470	0.568	0.598