Istation's Indicators of Progress (ISIP)™
Math

Technical Report

Computer Adaptive Testing System for Continuous Progress Monitoring of Math Growth for Students Prekindergarten through Grade 8

Istation
Supporting Educators. Empowering Kids.
Changing Lives.
2000 Campbell Centre II
8150 North Central Expressway
Dallas, Texas 75206
866.883.7323
www.istation.com

Table of Contents

Chapter 1: Introduction 1-1

The Need to Link Math Assessment to Instructional Planning 1-2

Continuous Progress Monitoring 1-3

Computer Adaptive Testing 1-4

ISIP Math and ISIP Early Math Domains 1-5

ISIP Math and ISIP Early Math Items 1-8

The ISIP Math and ISIP Early Math Link to Instructional Planning 1-10

Chapter 2: IRT Calibration and the CAT Algorithm Grades Pre-K – 1 2-1

Data Analysis and Results 2-3

CAT Algorithm 2-5

Ability Estimation 2-6

Chapter 3: IRT Calibration and the CAT Algorithm Grades 2–8 3-1

Data Analysis and Results 3-3

CAT Algorithm 3-6

Ability Estimation 3-7

Chapter 4: Reliability and Validity of ISIP Math 4-1

Reliability 4-2

Validity Evidence 4-3

Full Validity Study 4-7

Chapter 5: Determining Norms 5-1

Sample 5-4

Computing Norms 5-5

Instructional Tier Goals 5-6

Chapter 6: References 6-1

Chapter 1: Introduction

Istation's Indicators of Progress for Math (ISIP™ Math for grades 2-8 and ISIP Early Math for prekindergarten through 1st grade) are sophisticated, web-delivered, computer-adaptive testing (CAT) systems that provide continuous progress monitoring (CPM) in the subject area of mathematics.

Assessments are computer-based, and teachers can arrange for entire classrooms to take assessments as part of scheduled computer lab time or individually as part of a workstation rotation conducted in the classroom. Each assessment period requires approximately 30 minutes. Given adequate computer resources, it would be feasible to administer ISIP Math or ISIP Early Math assessments to an entire classroom, an entire school, or even an entire district in a single day. Classroom and individual student results are available in real time to teachers, illustrating each student's past and present performance on mathematical concepts. Teachers are alerted when a particular student is not making adequate progress so that the instructional program can be modified before a pattern of failure becomes established.

ISIP Early Math is designed for students in prekindergarten through 1st grade. The ISIP Early Math assessment is a computer-based universal screener designed to help teachers identify students struggling to learn critical mathematics content. ISIP Early Math provides teachers and other school personnel with easy-to-interpret, web-based reports that detail student strengths and deficits, helping to inform teachers' instructional decision-making. Using this data allows teachers to more easily make informed decisions with regard to each student's response to targeted mathematics instruction and intervention strategies.

ISIP Math is designed in a testing format that is familiar to most students in grades 2–8. Each item contains a question stem and four answer choices. As with ISIP Early Math, ISIP Math provides teachers and other school personnel with easy-to-interpret, web-based reports that detail student strengths and deficits.

Both ISIP Early Math and ISIP Math provide links to teaching resources and targeted intervention strategies. Computer-adaptive assessments measure each student's overall proficiency and mathematical ability.

The Need to Link Math Assessment to Instructional Planning

It is well established that assessment-driven instruction is effective. Teachers who monitor their students' progress and use this data to inform instructional planning and decision-making have higher student outcomes than those who do not (Conte and Hintze 2000; Fuchs et al. 1992; Mathes et al. 1998). These teachers also have a more realistic idea of the capabilities of their students than teachers who do not regularly use student data to inform their decisions (Fuchs et al. 1984; Fuchs et al. 1991; Mathes et al. 1998).

However, before a teacher can identify students at risk of mathematics failure and differentiate instruction, that teacher must first have information about the specific needs of his or her students. To effectively link assessment with instruction, math assessments need to:

- identify students at risk of having difficulty in math (i.e., students that may need extra instruction or intensive intervention if they are to progress toward grade-level standards in math by year's end);
- monitor student progress for growth on a frequent, ongoing basis and identify students falling behind;
- provide information about students that will be helpful in planning instruction to meet their needs; and
- assess whether students have achieved grade-level mathematics standards by year's end.

In any model of instruction, for assessment data to affect instruction and student outcomes, it must be relevant, reliable, and valid.

- To be **relevant**, data must be available on a timely basis and target important skills that are influenced by instruction.
- To be **reliable**, there must be a reasonable degree of confidence in student scores.
- To be **valid**, the skills assessed must provide information that is related to future mathematical ability.

There are many reasons why a student score from a single point in time under one set of conditions may be inaccurate: confusion, shyness, illness, mood or temperament, communication or language barriers between student and examiner, scoring errors, or inconsistencies in examiner scoring. However, by gathering assessments across multiple time points, student performance is more likely to reflect actual ability. Using the computer also reduces inaccuracies related to human administration errors.

The collection of sufficient, reliable assessment data on a continuous basis can be an overwhelming and daunting task for schools and teachers. Screening and inventory tools use a benchmark or screen schema in which assessments are administered three times a year. More frequent continuous progress monitoring is recommended for all low-performing students, but administration is at the discretion of already overburdened schools and teachers.

These assessments, even in their handheld versions, require a significant amount of work to administer individually to each student. The examiners who implement these assessments must also receive extensive training in both the administration and scoring procedures to uphold the reliability of the assessments and avoid scoring errors. Because these assessments are so labor intensive, they are very expensive for school districts to implement and difficult for teachers to use for ongoing progress monitoring and validation of test results. Moreover, there is typically a delay between when an assessment is given to a student and when the teacher is able to receive and review the results of the assessment, making its utility for planning instruction less than ideal.

Continuous Progress Monitoring

ISIP Math and ISIP Early Math grow out of the model of continuous progress monitoring (CPM) called Curriculum Based Measurement (CBM). This model of CPM is an assessment methodology for obtaining measures of student achievement over time. This is done by repeatedly sampling proficiency in the school's curriculum at a student's instructional level, using parallel forms at each testing session (Deno 1985; Fuchs and Deno 1991; Fuchs et al. 1983). Parallel forms are designed to globally sample academic goals and standards reflecting end-of-grade expectations. Students are then measured in terms of movement toward those end-of-grade expectations. A major drawback to this type of assessment is that creating truly parallel forms of any assessment is virtually impossible; thus, student scores from session to session will reflect some inaccuracy as an artifact of the test itself.

Computer Application

The challenge with most CPM systems is that they have been cumbersome for teachers to implement and use (Stecker and Whinnery 1991). Teachers have to administer tests to each student individually and then graph the data by hand. The introduction of hand-held technology has allowed for organizing and displaying student results more easily, but information in this format is often not available on a timely basis. Even so, many teachers find administering such assessments onerous. The result has been that CPM has not been as widely embraced as originally hoped, especially within general education.

Computerized CPM applications, however, are a logical step toward increasing the likelihood that continuous progress monitoring occurs more frequently with monthly or even weekly

assessments. Computerized CPM applications using parallel forms have been developed and used successfully in upper grades for reading, mathematics, and spelling (Fuchs et al. 1995). Computerized applications save time and money. They eliminate burdensome test administrations and scoring errors by calculating, compiling, and reporting scores. They provide immediate access to student results that can be used to affect instruction. They provide information organized in formats that automatically group students according to risk and recommended instructional levels. Student results are instantly plotted on progress charts with trend lines projecting year-end outcomes based upon growth patterns, eliminating the need for the teacher to manually create monitoring booklets or analyze results.

Computer Adaptive Testing

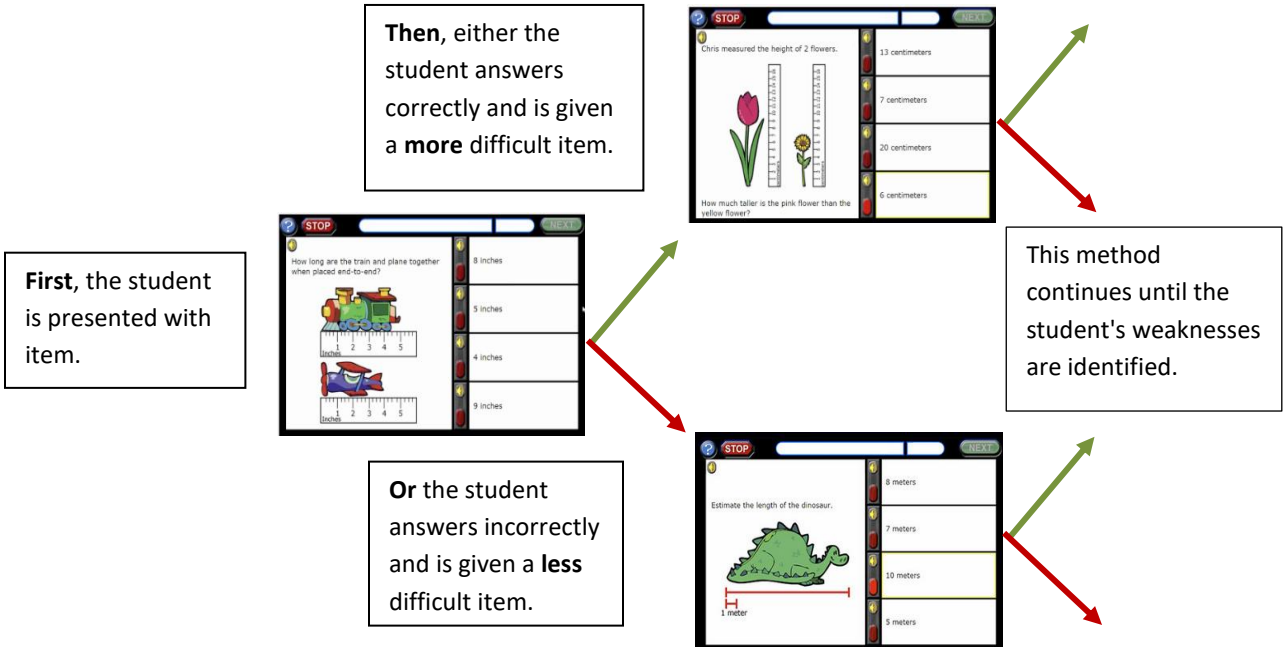
With recent advances in computer adaptive testing (CAT) and computer technology, it is now possible to create CPM assessments that adjust to the actual ability of each student. Thus, CAT replaces the need to create parallel forms. Assessments built on CAT are sometimes referred to as "tailored tests" because the computer selects items for students based on their individual performance, thus tailoring the assessment to match the performance abilities of each student.

There are many advantages to using a CAT model rather than the traditional parallel forms model, as is used in many math instruments. For instance, it is virtually impossible to create alternate forms of any truly parallel assessment. The reliability from form to form will always be somewhat compromised. However, when using a CAT model, it is not necessary that each assessment be of identical difficulty to the previous and future assessments.

In CAT models, each item within the testing battery is assessed to determine how well it discriminates ability among students and how difficult it actually is through a process called Item Response Theory (IRT). Once these parameters have been determined for each item, the CAT algorithm can be programmed. Using this sophisticated computerized algorithm, the computer adaptively selects items based on each student's performance during the assessment. Test questions range from easy to hard for each covered strand. To identify the student's overall ability and individual skill level, the difficulty of the test questions presented changes with every response.

If a student answers questions correctly on the ISIP assessment, the program will present questions that are more challenging until the student shows mastery or responds with an incorrect answer. When a student answers a question incorrectly, ISIP will present less difficult questions until the student begins answering correctly again. Through this process of selecting items based on student performance, the computer is able to generate "probes" that have higher reliability than those typically associated with alternate formats and that

better reflect each student's true ability. The ability score shows how a student is performing compared to their previous performance and to other students at the same grade level.



ISIP Math and ISIP Early Math assessments are delivered at established intervals (usually monthly) to the appropriate grade level for each student throughout a nine-month school year. This provides opportunity for teachers to identify where students fall within grade-level expectations and assists teachers in preparing for state standardized assessments which are typically delivered only at grade-level standards.

ISIP Math and ISIP Early Math Domains

Designed for students in prekindergarten through 8th grade, ISIP Early Math and ISIP Math provide teachers and other school personnel with easy-to-interpret, web-based reports that detail student strengths and deficits and provide links to additional intervention resources. Using this data allows teachers to more easily make informed decisions regarding each student's response to targeted math instruction and intervention strategies. Reports from the ISIP assessment provide teachers with the information they need to know, including:

- if students have deficits in math skills that could place them at risk for failure;
- if instruction is having the desired effect of raising students' math knowledge; and
- if students are making progress in comprehending increasingly challenging material.

ISIP Math and ISIP Early Math measures proficiency in the six primary domains of mathematical reasoning and processes — number sense, operations, algebra, geometry, measurement, and data analysis — as defined by the National Council of Teachers of Mathematics (NCTM), and it also measures personal financial literacy (PFL) as determined by the Texas Essential Knowledge and Skills (TEKS).

Number Sense

The fundamental basis of all mathematics is understanding numbers and having awareness of the relationships among numbers. Students must be taught to recognize how numbers are represented as well as number systems and counting sequences. Instruction in this essential area is the most fundamental content standard.

Operations

Comprehension of mathematical operations, concepts, and relations is critical to developing an understanding of number value and sequence. For example, what does it mean to add, subtract, multiply, or divide? How do these functions impact value? The ability to estimate and perform mental calculations as well as calculate answers on paper are both crucial components to achieving success in math.

Algebra

Students must be able to comprehend statements of relations, mathematical symbols, and rules for ordering and executing computations using them. The skills related to algebra that all students must learn include, but are not limited to:

- recognizing and comprehending numerical patterns, relationships, and functions;
- applying mathematical constructs to explain quantitative relationships;
- illustrating computational examples using algebraic symbols; and
- evaluating variance in mathematical situations.

Geometry

The ultimate goal of geometry is to arm students with foundational skills to accomplish everyday tasks such as describing shapes and angles, recognizing patterns and measurements, and even reading a map. The geometry concepts that must be taught to encourage student achievement in geometry include but are not limited to:

- calculating area and perimeter of two-dimensional geometric shapes;

- analyzing volume, surface area, and other properties of three-dimensional geometric shapes;
- constructing equations and statements to describe geometric relationships;
- characterizing spatial relationships and using coordinates to identify location; and
- applying spatial reasoning, geometric modeling, and concepts of symmetry to mathematical contexts.

Measurement

Measurement skills are unique in that students often inherently recognize their practical significance. Comprehension of measurement also provides many opportunities to practice and apply many other math skills, especially geometry and operations. Students must learn about different systems of measurements (metric vs. customary), formulae for calculating measurements (length/height, area/perimeter, weight/capacity/volume), application of appropriate tools (ruler vs. protractor), and dimensions of time and money.

Data Analysis

Beyond number recognition and operational aptitude, students must be able to form and evaluate numerical inferences and then formulate accurate mathematical conclusions. The analytical math concepts that all students should learn include, but are not limited to:

- reading, creating, and interpreting graphs and charts;
- devising and answering formulaic expressions using collected and organized data;
- analyzing data by recognizing appropriate statistical modes; and
- comprehending and executing basic probability concepts.

ISIP Math and ISIP Early Math Items

The unique item banks for ISIP Math assessments are designed to provide an accurate computer-adaptive universal screening and progress-monitoring assessment system that can support and inform teachers' instructional decisions. By administering the grade-appropriate assessments, teachers and administrators can then use the results to answer two questions:

1. Are students in the designated grade at risk of failing math?
2. What degree of instructional support will students require to be successful at math?

Because the assessments are designed to be administered at regular intervals, these decisions can be applied throughout the course of the school year (Hill, S., Ketterlin-Geller, L.R., & Gifford, D.B., 2012).

The ISIP Math and ISIP Early Math assess both proficiency in mathematical concepts and students' level of cognitive engagement.

Table 1-1. ISIP Skills and Domains.

Strands of Proficiency for Cognitive Engagement			
Strategic Competence	Adaptive Reasoning	Procedural Fluency	Conceptual Understanding
Mathematical Domains			
Number Sense	Algebra	Measurement	Probability and Statistics
Operations	Geometry	Data Analysis	Ratios and Proportional Relationships

The mathematical content (by domain) of the assessment is based on:

- the Curriculum Focal Points (developed by National Council of Teachers of Mathematics [NCTM] in 2006,
- the mathematics content standards published by the Common Core State Standards Initiative, and
- state standards from California, Florida, New York, Texas, and Virginia.

The cognitive engagement dimension refers to the level of cognitive processing at which students are expected to engage with an assessment item.

Levels of cognitive processing consists of five interdependent strands that promote mathematical proficiency:

1. conceptual understanding
2. procedural fluency
3. strategic competence
4. adaptive reasoning
5. productive disposition

The formative assessment item bank assesses student understanding of the content at varying levels of cognitive engagement. The item bank incorporates four of the five strands. Productive disposition is not assessed (Hill, S., Ketterlin-Geller, L.R., & Gifford, D.B., 2012).

To access the complete technical reports for the Universal Screener Instrument Development for pre-K through 1st grade and the Universal Screener and Inventory Instruments Interface Development for pre-K through 1st grade, refer to the external links provided at the end of this report. To access the technical reports for the Universal Screener Instrument Development for each grade level 2 through 8, refer to the external links provided at the end of this report.

Teacher Friendly

ISIP Math and ISIP Early Math are teacher friendly. Each assessment is computer based, requires little administrative effort, and requires no teacher/examiner testing or manual scoring. Teachers simply monitor student performance during assessment periods to ensure reliability and accuracy of results. In particular, teachers are alerted to observe any students identified by ISIP Math or ISIP Early Math (depending on grade level) who are experiencing difficulties as they complete the assessment. They subsequently review student results to validate outcomes. For students whose skills may be a concern, based upon performance level, teachers may easily validate student results by re-administering the entire ISIP Math or ISIP Early Math as an On-Demand assessment.

Student Friendly

Both the ISIP Math and ISIP Early Math are student friendly. Each assessment session in ISIP Early Math gives students the feeling of shopping in a grocery store called Mario's Market. At the beginning of the session, Mario appears onscreen and welcomes the student briefly before the assessment begins. Assessment delivery is presented in a developmentally appropriate format with respect to students' reading skills, fine/gross motor skills, and hand-eye coordination. Consideration of young students' fine motor skills informs navigation design and managing assessment interfaces that allow as much hands-on/manipulative-based interaction as possible. The singular interface theme of Mario's Market is used to minimize student distractions and unnecessary cognitive load.

Similarly, each assessment session in ISIP Math begins with an introduction from a familiar Istation Math character, the Chief. The Chief briefly explains that the student's mathematical knowledge demonstrated on the assessment will help them become a secret agent. He informs the student that once the assessment is complete, they will participate in math missions with Donnie, Stix, and Angel to defeat villains and save the world. This ties together the ISIP Math and the instruction in Istation Math. Additionally, it provides motivation for students to do their best when completing the assessment.

The ISIP Math and ISIP Early Math and Instructional Planning

ISIP Math and ISIP Early Math provide continuous assessment results that can be used in recursive assessment instructional decision loops.

First, each assessment identifies students in need of support.

Second, validation of student results and recommended instructional levels can easily be verified by re-administering assessments. If a student's results seem inconsistent with other ISIP Math data points, the teacher can use the On-Demand feature of the Istation website at www.istation.com. By assigning additional assessments to individual students, results can be compared and evaluated by the teacher. When the On-Demand feature is used, the assessment will be automatically administered the next time a student logs in.

Third, the delivery of student results facilitates the evaluation of curriculum and instructional plans. The technology behind ISIP Math and ISIP Early Math delivers real-time evaluation of results, and reports on student progress are immediately available upon assessment completion. Assessment reports automatically group students by level of support needed. Data is provided in both graphic and detailed numerical format for every test administration and for every level of a district's reporting hierarchy. Reports provide summary information for the current and prior assessment periods that can be used to evaluate curriculum, plan instruction and support, and manage resources.

At each assessment period, ISIP Math and ISIP Early Math automatically alert teachers to students in need of instructional support via the Priority Report. Students are grouped according to instructional level. Links to relevant teacher directed lessons and other instructional materials are provided for each instructional level. When student performance on assessments is below the goal for several consecutive assessments, teachers are further notified in order to raise teacher concern and signal the need to consider additional or different forms of instruction.

A complete history of Priority Report notifications, including the current year and all prior years, is maintained for each student. On the report, teachers may acknowledge that suggested interventions have been provided. A record of these interventions is maintained with the student history as an intervention audit trail. This history can be used for special education Individualized Education Plans (IEPs) and in Response to Intervention (RTI) or other models of instruction to modify a student's instructional plan.

In addition to the recommended activities, instructional coaches, intervention specialists, and teachers have access to an entire library of teacher directed lessons and support materials at **www.istation.com**. Districts and schools may also elect to enroll students in Istation's computer-based math intervention program, Istation Math. This program provides individualized instruction based on a student's results from ISIP Math or ISIP Early Math. Student results from Istation Math are combined with ISIP Math or ISIP Early Math results to provide a more accurate profile of a student's strengths and weaknesses that can help inform and enhance teacher planning.

All student information is automatically available, sorted by demographic classification and by designated subgroups of students who may need to be monitored. As students progress in the program, a year-to-year history of ISIP Math or ISIP Early Math results will be available. Administrators, principals, and teachers may use these reports to evaluate and modify curriculum, intervention strategies, the effectiveness of professional development, and personnel performance.

Chapter 2: IRT Calibration and the CAT Algorithm Grades Pre-K – 1

The goals of this study were to determine the appropriate item response theory (IRT) model, estimate item-level parameters, and tailor the computer adaptive testing (CAT) algorithms, such as the exit criteria.

During the 2014-2015 school year, data were collected from schools across the country so that ISIP™ Early Math (pre-K through 1st grade) would be available for schools in the 2015-2016 school year. All students in prekindergarten through 1st grade were invited to participate, including students with disabilities and English language learners. There were no specific demographic requirement for participants.

Tests were administered by computer to groups in a classroom or computer lab setting. There were 397 items for prekindergarten, 401 items for kindergarten, and 395 items for 1st grade. The items were divided into nine test forms per grade with linking items between forms. Each test form lasted 20-25 minutes for prekindergarten students and 30-45 minutes for kindergarteners and 1st grade students. Each grade level had its own item pool with no linking items between those pools; prekindergarten test forms were only taken by students in prekindergarten, kindergarten test forms were only taken by kindergarteners, and 1st grade test forms were only taken by 1st grade students.

Approximately 5,000 students per grade level participated in this study. The majority of students did not provide demographic information, but 1,006 prekindergartners, 556 kindergarteners, and 705 1st graders did provide such information. The information from these students is reported in Table 2-1.

Table 2-1. Student Demographics Grades Pre-K – 1.

Students	Prekindergarten Frequency (%)	Kindergarten Frequency (%)	Grade 1 Frequency (%)
Gender			
Male	500 (49.7)	299 (53.8)	372 (52.8)
Female	506 (50.3)	257 (46.2)	333 (47.2)
Ethnicity			
African American	778 (77.3)	107 (19.2)	133 (18.9)
American Indian	3 (0.3)	4 (0.7)	5 (0.7)
Asian	2 (0.2)	8 (1.4)	4 (0.6)
Hispanic	12 (1.2)	102 (18.3)	7 (1.0)
White	172 (17.1)	298 (53.6)	277 (39.3)
Unknown	39 (3.9)	37 (6.7)	279 (39.6)
Receiving Special Ed Services			
Yes	41 (4.1)	8 (1.4)	10 (1.4)
No	915 (91.0)	145 (26.1)	289 (41.0)
Receiving Free/Reduced Lunch			
Yes	10 (1.0)	74 (13.3)	106 (15.0)
No	1 (0.1)	79 (14.2)	175 (24.8)
Receiving ESL Services			
Yes	10 (1.0)	1 (0.2)	6 (0.9)
No	1 (0.1)	152 (27.3)	274 (38.9)
Disability			
Yes	—	1 (0.2)	1 (0.1)
No	—	—	—

Data Analysis and Results

A two-parameter logistic IRT (Item Response Theory) model (2PL IRT) was posited. We defined the binary response data, x_{ij}, with index i = 1, ... n for persons, and index j = 1, ... j for items. The binary variable x_{ij} = 1 was used if the response from student i to item j was correct, and the binary variable x_{ij} = 0 was used if the response was wrong. In the 2PL IRT model, the probability of a correct response from examinee i to item j was defined as:

$$P_j(\theta_i) = \frac{\exp\left[a_j(\theta_i - b_j)\right]}{1+\exp\left[a_j(\theta_i - b_j)\right]}$$

The variable θ_i is examinee i's ability parameter, b_j is item j's difficulty parameter, and a_j is item j's discrimination parameter.

While the marginal maximum likelihood estimation (MMLE) approach by Bock and Aitkin (1981) has many desirable features compared to earlier estimation procedures, such as consistent estimates and manageable computation, there are some limitations. For example, items must be eliminated if they are answered correctly by *all* of the examinees or if they are answered incorrectly by all. Also, item discrimination estimates near zero can result in very large absolute values of item difficulty estimates, which may fail the estimation process and no ability estimates can be obtained. To overcome these limitations, we employed a full Bayesian framework to fit the IRT models. More specifically, the likelihood function based on the sample data is combined with the prior distributions assumed on the set of the unknown parameters to produce the posterior distribution of the parameters; the inference is then based on the posterior distribution.

There are two roles played by the prior distribution. First, if we have information from experts or previous studies on the IRT parameters, such as a certain group of items being more challenging, we can utilize the data from the prior studies to help produce more stable estimates. On the other hand, if we know little about those parameters, we could use the non-informative prior data alongside a large variance to reflect this uncertainty. Second, in the Bayesian estimation, the primary effect of the prior distribution is to shrink the estimates toward the mean of the prior. The shrinkage towards the prior mean helps prevent deviant parameter estimates. Furthermore, with the Bayesian approach, there is no need to eliminate any data.

As for the prior specification, we assumed that the j item difficulty parameters are independent, as are the j item discrimination parameters and the n examinee ability parameters. We initially assigned the subject ability parameters and item difficulty parameters non-informative, two-stage, normal priors:

$$\theta_i \sim N(0,\tau_\theta,) \qquad i = 1, \ldots n$$

$$\delta_j \sim N(0,\tau_\delta,) \qquad j = 1, \ldots j$$

Variance parameters τ_θ and τ_δ each follow a conjugate inverse gamma prior to introduce more flexibility (where *a* and *b* are fixed values):

$$\tau_\theta \sim IG(a_\theta, b_\theta)$$

$$\tau_\theta \sim IG(a_\delta, b_\delta)$$

The hyperparameters were assigned to produce vague priors. From Berger (1985), Bayesian estimators are often robust to changes of hyperparameters when non-informative or vague priors are used. We let $a_\theta = a_\lambda = 2$ and $b_\theta = b_\delta = 1$, allowing the inverse gamma priors to have infinite variances.

By definition, the item discrimination parameters are necessarily positive, so we assumed a gamma prior:

$$\lambda \sim \text{Gamma}(a_\lambda, b_\lambda), \qquad j = 1, \ldots j.$$

The hyper-parameters were defined as $a_\lambda = b_\lambda = 1$.

The Gibbs sampler, a Bayesian parameter estimation technique, was employed to obtain item parameter estimates by way of a BILOG program. The resulting analysis produced two parameter estimates for each item: an item difficulty parameter and an item discrimination parameter (which indicates how well an item discriminates between students with low math ability and students with high math ability). Items that did not meet Istation criteria were removed.

A huge sample size was used in this study. For prekindergarten, the responses per item ranged from 684 to 2,535. For kindergarten, the responses per item ranged from 573 to 1,888. For 1st grade, the responses per item ranged from 737 to 2,717.

Regarding the content of the items, multiple sub-contents are measured for each grade.

The prekindergarten item pool measured the following:

- Counting Skills,
- Number Sense,
- Number and Operations,
- Counting and Cardinality,
- Adding To/Taking Away Skills,
- Geometry,
- Spatial Relations,
- Measurement,
- Measurement Skills,
- Data Analysis,
- Mathematical Reasoning,
- Data Collection and Statistics,

- Algebra and Functions,
- Algebra,
- Patterns and Seriation, and
- Patterns and Relationships.

The kindergarten item pool measured the following:

- Counting and Cardinality,
- Number and Operations,
- Number and Number Sense,
- Operations and Algebraic Thinking,
- Number and Operations in Base Ten,
- Geometry,
- Geometry and Measurement,
- Measurement,
- Probability and Statistics,
- Data Analysis,
- Measurement and Data,
- Personal Financial Literacy, and
- Algebra.

The 1st grade item pool measured the following:

- Number Sense,
- Operations and Algebraic Thinking,
- Algebra,
- Measurement and Data,
- Patterns,
- Functions,
- Number and Operations,
- Number and Operations in Base Ten,
- Algebraic Reasoning,
- Geometry,
- Measurement and Data Analysis,
- Measurement,
- Data analysis, and
- Personal Financial Literacy.

Overall, most items were good quality in terms of item discriminations and item difficulties. For prekindergarten, five items were removed and 392 calibrated item parameters remain in the item pool. For kindergarten, 23 items were removed and 377 calibrated item parameters remain in the item pool. For 1st grade, 35 items were removed and 360 calibrated item parameters remain in the item pool.

CAT Algorithm

The Computerized Adaptive Testing (CAT) algorithm is an iterative approach to test taking. Instead of giving a large, general pool of items to all test takers, a CAT test repeatedly selects the optimal next item for the individual test taker, bracketing their ability estimate until some stopping criteria is met.

The algorithm is as follows:

1. Assign an initial ability estimate to the test taker.
2. Ask the question that gives the most information based on the current ability estimate.

3. Re-estimate the ability level of the test taker based on their answer to the prior question.
4. If stopping criteria is met, stop. Otherwise, return to step 2 and repeat.

This iterative approach is made possible by using IRT models. IRT models generally estimate a single, latent trait (ability) of the test taker, and this trait is assumed to account for all response behavior. These models provide response probabilities based on test taker ability and item parameters. Using these item response probabilities, we can compute the amount of information each item will yield for a given ability level. In this way, we can select the next item in a way that maximizes information gain based on student ability rather than percent correct or grade-level expectations.

Though the CAT algorithm is simple, it allows for endless variations on item selection criteria, stopping criteria, and ability estimation methods. All of these elements play into the predictive accuracy of a given implementation, and the best combination is dependent on the specific characteristics of the test and the test takers.

In developing Istation's CAT implementation, we explored many approaches. To assess the various approaches, we ran CAT simulations using each approach on a large set of real student responses to our items (1,000 students, 700 item responses each). To compute the "true" ability of each student, we used Bayes expected a posteriori (EAP) estimation on all 700 item responses for each student. We then compared the results of our CAT simulations against these "true" scores and other criteria to determine which approach was most accurate.

Ability Estimation

From the beginning, we decided to take a Bayesian approach to ability estimation, with the intent of incorporating prior knowledge about the student (from previous test sessions and grade-based averages). In particular, we initially chose Bayes EAP with good results. We briefly experimented with the maximum likelihood estimation (MLE) method as well but abandoned it because the computation required more items to converge to a reliable ability estimate.

To compute the prior integral required by EAP, we used Gauss-Hermite quadrature with 88 nodes from –7 to +7. This is certainly more than needed, but because we were able to save runtime computation by pre-computing the quadrature points, we decided to err on the side of accuracy.

For the Bayesian prior, we used a standard normal distribution centered on the student's ability score from the previous testing period (or the grade-level average for the first testing period). We decided to use a standard normal prior rather than using σ from the previous testing period in order to avoid overemphasizing possibly out-of-date information.

Item Selection

For our item selection criteria, we simulated twelve variations on maximum information gain. The difference in accuracy between the various methods was extremely slight, so we gave preference to methods that minimized the number of items required to reach a satisfactory standard error (keeping the attention span of children in mind). In the end, we settled on selecting the item with maximum Fisher information. This approach appeared to offer the best balance of high accuracy and least number of items presented.

Stopping Criteria

We set a five-item minimum and twenty-item maximum per subtest. Within those bounds, we ended ISIP Early Math when the ability score's standard error dropped below a preset threshold or when four consecutive items each reduced the standard error by less than a preset amount.

Chapter 3: IRT Calibration and the CAT Algorithm Grades 2–8

The goals of this study were to determine the appropriate item response theory (IRT) model, estimate item-level parameters, and tailor the computer adaptive testing (CAT) algorithms, such as the exit criteria.

During the 2012-2013 school year, data were collected from schools in Texas during the spring semester so that ISIP™ Math (2nd through 8th grade) would be available for schools in the 2013-2014 school year. All students in 2nd through 8th grade were invited to participate, including students with disabilities and English language learners.

Tests were administered by computer to groups in a classroom or computer lab setting. There were 940 items for 2nd grade; 1,066 items for 3rd grade; 875 items for 4th grade; 882 items for 5th grade; 1,159 items for 6th grade; 938 items for 7th grade; and 616 items for 8th grade. The items were divided into 20 test forms per grade with linking items between forms. Each test form lasted 40-55 minutes. Each grade level had its own item pool with no linking items between those pools. To be more specific, 2nd grade test forms were only taken by 2nd grade students, 3rd grade test forms were only taken by 3rd grade students, and so on.

Approximately 6,000 students per grade level participated in this study. Students had the choice to provide demographic information or not. We received data from 3,937 2nd graders; 5,127 3rd graders; 5,832 4th graders; 5,067 5th graders; 6,347 6th graders; 1,537 7th graders; and 1,169 8th graders. The information from these students is reported in Table 3-1.

Table 3-1. Student demographics grades 2–8.

Students	Grade 2 Freq. (%)	Grade 3 Freq. (%)	Grade 4 Freq. (%)	Grade 5 Freq. (%)	Grade 6 Freq. (%)	Grade 7 Freq. (%)	Grade 8 Freq. (%)
Gender							
Male	1,548 (39.3)	1,726 (33.7)	2,094 (35.9)	1,704 (33.6)	2,700 (42.5)	761 (49.5)	585 (50.0)
Female	1,336 (33.9)	1,679 (32.7)	2,049 (35.1)	1,577 (31.1)	2,617 (41.2)	760 (49.4)	572 (48.9)
Ethnicity							
African American	813 (20.7)	467 (9.1)	989 (17.0)	612 (12.1)	1,292 (20.4)	197 (12.8)	203 (17.4)
American Indian	32 (0.8)	28 (0.5)	13 (0.2)	20 (0.4)	61 (1.0)	8 (0.5)	5 (0.4)
Asian	64 (1.6)	53 (1.0)	184 (3.2)	200 (3.9)	140 (2.2)	13 (0.8)	18 (1.5)
Hispanic	743 (18.9)	117 (2.3)	120 (2.1)	131 (2.6)	215 (3.4)	111 (7.2)	88 (7.6)
White	1,137 (28.6)	1,484 (28.9)	1,750 (30.0)	1,710 (33.7)	1,830 (28.8)	755 (49.1)	664 (56.8)
Unknown	1,148 (29.2)	2,978 (58.1)	705 (12.1)	2,394 (47.2)	2,809 (44.3)	453 (29.5)	191 (16.3)
Receiving Special Ed Services							
Yes	246 (6.2)	212 (4.1)	289 (5.0)	236 (4.7)	643 (10.0)	112 (7.3)	109 (9.3)
No	2,401 (61.0)	2,474 (48.3)	1,754 (30.1)	1,660 (32.8)	3,767 (59.4)	972 (63.2)	869 (74.3)
Receiving Free/Reduced Lunch							
Yes	1,516 (38.5)	2,013 (39.3)	74 (13.3)	2,504 (49.4)	2,641 (41.6)	911 (59.3)	808 (69.1)
No	540 (13.7)	628 (12.2)	79 (14.2)	2,563 (51.6)	1,242 (19.6)	6 (0.4)	1 (0.1)
Receiving ESL Services							
Yes	331 (8.4)	1 (0.2)	1 (0.2)	26 (0.5)	576 (9.1)	23 (1.5)	58 (4.9)
No	2160 (54.9)	152 (27.3)	152 (27.3)	2,497 (49.3)	2,358 (37.2)	1,020 (66.4)	920 (78.7)
Disability							
Yes	183 (4.6)	251 (4.9)	305 (5.2)	283 (5.6)	95 (1.5)	270 (17.6)	252 (21.6)
No	3,754 (95.4)	4,876 (95.1)	5,527 (94.8)	4,784 (94.4)	6,252 (98.5)	1,267 (82.4)	917 (78.4)

Data Analysis and Results

A two-parameter logistic IRT (Item Response Theory) model (2PL IRT) was posited. We defined the binary response data, x_{ij}, with index i = 1, ... n for persons, and index j = 1, ... j for items. The binary variable x_{ij} = 1 was used if the response from student i to item j was correct and the binary variable x_{ij} = 0 was used if the response was wrong. In the 2PL IRT model, the probability of a correct response from examinee i to item j was defined as:

$$P_j(\theta_i) = \frac{\exp\left[a_j(\theta_i - b_j)\right]}{1+\exp\left[a_j(\theta_i - b_j)\right]}$$

The variable θ_i is examinee i's ability parameter, b_j is item j's difficulty parameter, and a_j is item j's discrimination parameter.

While the marginal maximum likelihood estimation (MMLE) approach by Bock and Aitkin (1981) has many desirable features compared to earlier estimation procedures, such as consistent estimates and manageable computation, there are some limitations. For example, items answered correctly or incorrectly by all of the examinees must be eliminated. Also, item discrimination estimates near zero can result in very large absolute values of item difficulty estimates, which may fail the estimation process and no ability estimates can be obtained. To overcome these limitations, we employed a full Bayesian framework to fit the IRT models. More specifically, the likelihood function based on the sample data is combined with the prior distributions assumed on the set of the unknown parameters to produce the posterior distribution of the parameters; the inference is then based on the posterior distribution.

There are two roles played by the prior distribution. First, if we have information from experts or previous studies on the IRT parameters, such as a certain group of items being more challenging, we can utilize the data from the prior studies to help produce more stable estimates. On the other hand, if we know little about those parameters, we could use the non-informative prior data alongside a large variance to reflect this uncertainty. Second, in the Bayesian estimation, the primary effect of the prior distribution is to shrink the estimates towards the mean of the prior. The shrinkage towards the prior mean helps prevent deviant parameter estimates. Furthermore, with the Bayesian approach, there is no need to eliminate any data.

As for the prior specification, we assumed that the j item difficulty parameters are independent, as are the j item discrimination parameters and the n examinee ability parameters. We initially assigned the subject ability parameters and item difficulty parameters non-informative, two-stage normal priors:

$$\theta_i \sim N(0,\tau_\theta,) \qquad i = 1, \ldots n$$

$$\delta_j \sim N(0,\tau_\delta,) \qquad j = 1, \ldots j$$

Variance parameters τ_θ and τ_δ each follow a conjugate inverse gamma prior to introduce more flexibility (where a and b are fixed values):

$$\tau_\theta \sim IG(a_\theta, b_\theta)$$

$$\tau_\theta \sim IG(a_\delta, b_\delta)$$

The hyper-parameters were assigned to produce vague priors. From Berger (1985), Bayesian estimators are often robust to changes of hyper-parameters when non-informative or vague priors are used. We let $a_\theta = a_\lambda = 2$ and $b_\theta = b_\delta = 1$, allowing the inverse gamma priors to have infinite variances.

By definition, the item discrimination parameters are necessarily positive, so we assumed a gamma prior:

$$\lambda \sim \text{Gamma}(a_\lambda, b_\lambda), \qquad j = 1, \ldots j.$$

The hyper-parameters were defined as $a_\lambda = b_\lambda = 1$.

The Gibbs sampler, a Bayesian parameter estimation technique, was employed to obtain item parameter estimates by way of a BILOG program. The resulting analysis produced two parameter estimates for each item-an item difficulty parameter and an item discrimination parameter, which indicates how well an item discriminates between students with low math ability and students with high math ability. Items that did not meet Istation criteria were removed. A huge sample size was used in this study. The responses per item ranged from 984 to 1,106 for 2nd grade, 1,037 to 1,142 for 3rd grade, 858 to 975 for 4th grade, 861 to 950 for 5th grade, 458 to 566 for 6th grade, 92 to 136 for 7th grade, and 142 to 167 for 8th grade.

Regarding the content of the items, multiple sub-contents are measured for each grade.

The 2nd grade item pool measured the following:

- Number and Operations – Base 10,
- Number and Operations,
- Number and Operations – Algebra,
- Number and Operations – Fractions,
- Measurement and Data,
- Probability and Statistics,
- Personal Financial Literacy, and
- Geometry.

The 3rd grade item pool measured the following:

- Number and Operations – Base 10,
- Number and Operations,
- Number and Operations – Algebra,
- Number and Operations – Fractions,
- Measurement and Data,
- Probability and Statistics,
- Personal Financial Literacy,
- and Geometry.

The 4th grade item pool measured the following:

- Number and Operations – Base 10,
- Number and Operations – Algebra,
- Number and Operations – Fractions,
- Measurement and Data,
- Probability and Statistics,
- Personal Financial Literacy,
- and Geometry.

The 5th grade item pool measured the following:

- Number and Operations – Base 10,
- Number and Operations – Fractions,
- Measurement and Data,
- Probability and Statistics,
- Personal Financial Literacy,
- Operations and Algebraic Thinking,
- and Geometry.

The 6th grade item pool measured the following:

- Expressions, Equations, and Relationships,
- Operations and Algebraic Thinking,
- Ratios and Proportional Relationships,
- Probability and Statistics,
- Personal Financial Literacy,
- and Geometry.

The 7th grade item pool measured the following:

- Expressions, Equations, and Relationships,
- Number and Operations,
- Ratios and Proportional Relationships,
- Probability and Statistics,
- Personal Financial Literacy,
- and Geometry.

The 8th grade item pool measured the following:

- Expressions, Equations, and Relationships,
- Functions,
- Number and Operations,
- Proportional Relationships,
- Probability and Statistics,
- Personal Financial Literacy,
- and Geometry.

Overall, most items were good quality in terms of item discriminations and item difficulties. For 2nd grade, 44 items were removed and 896 calibrated item parameters remain in the grade 2 item pool. Under 3rd grade, 53 items were removed and 913 calibrated item parameters remain in the grade 3 item pool. For 4th grade, 65 items were removed and 810 calibrated item parameters remain in the item pool. For 5th grade, 71 items were removed and 811 calibrated item parameters remain in the item pool. For 6th grade, 82 items were

removed and 977 calibrated item parameters remain in the item pool. For 7th grade, 96 items were removed and 742 calibrated item parameters remain in the item pool. For 8th grade, 73 items were removed and 543 calibrated item parameters remain in the item pool.

CAT Algorithm

The Computerized Adaptive Testing (CAT) algorithm is an iterative approach to test taking. Instead of giving a large, general pool of items to all test takers, a CAT test repeatedly selects the optimal next item for the test taker, bracketing their ability estimate until some stopping criteria is met.

The algorithm is as follows:

1. Assign an initial ability estimate to the test taker.
2. Ask the question that gives the most information based on the current ability estimate.
3. Re-estimate the ability level of the test taker based on their answer to the prior question.
4. If stopping criteria is met, stop. Otherwise, return to step 2 and repeat.

This iterative approach is made possible by using IRT models. IRT models generally estimate a single latent trait (ability) of the test taker, and this trait is assumed to account for all response behavior. These models provide response probabilities based on test taker ability and item parameters. Using these item response probabilities, we can compute the amount of information each item will yield for a given ability level. In this way, we can select the next item in a way that maximizes information gain based on student ability rather than percent correct or grade level expectations.

Though the CAT algorithm is simple, it allows for endless variations on item selection criteria, stopping criteria and ability estimation methods. All of these elements play into the predictive accuracy of a given implementation and the best combination is dependent on the specific characteristics of the test and the test takers.

In developing Istation's CAT implementation, we explored many approaches. To assess the various approaches, we ran CAT simulations using each approach on a large set of real student responses to our items (1,000 students, 700 item responses each). To compute the "true" ability of each student, we used Bayes expected a posteriori (EAP) estimation on all 700 item responses for each student. We then compared the results of our CAT simulations against these "true" scores and other criteria to determine which approach was most accurate.

Ability Estimation

From the beginning, we decided to take a Bayesian approach to ability estimation, with the intent of incorporating prior knowledge about the student (from previous test sessions and grade-based averages). In particular, we initially chose Bayes EAP with good results. We briefly experimented with the maximum likelihood estimation (MLE) method as well but abandoned it because the computation required more items to converge to a reliable ability estimate.

To compute the prior integral required by EAP, we used Gauss-Hermite quadrature with 88 nodes from –7 to +7. This is certainly more than needed, but because we were able to save runtime computation by pre-computing the quadrature points, we decided to err on the side of accuracy.

For the Bayesian prior, we used a standard normal distribution centered on the student's ability score from the previous testing period (or the grade-level average for the first testing period). We decided to use a standard normal prior rather than using σ from the previous testing period in order to avoid overemphasizing possibly out-of-date information.

Item Selection

For our item selection criteria, we simulated twelve variations on maximum information gain. The difference in accuracy between the various methods was extremely slight, so we gave preference to methods that minimized the number of items required to reach a satisfactory standard error (keeping the attention span of children in mind). In the end, we settled on selecting the item with maximum Fisher information. This approach appeared to offer the best balance of high accuracy and least number of items presented.

Stopping Criteria

We set a five-item minimum and twenty-item maximum per subtest. Within those bounds, we ended ISIP Math when the ability score's standard error dropped below a preset threshold or when four consecutive items each reduced the standard error by less than a preset amount.

Chapter 4: Reliability and Validity of ISIP™ Math

The primary objective of this study was to establish the technical adequacy of the Computer Adaptive Testing (CAT)-based ISIP Math and ISIP Early Math for students in kindergarten through 8th grade. This consisted of conducting test-retest reliability and concurrent and predictive validity work. We compared ISIP Math scores to scores from norm-referenced measures with good psychometric properties of similar constructs.

To establish reliability and validity evidence, data were collected during the 2015–2016 school year at three school districts in Texas. Demographics of the study's participants are found in Table 4-1.

Table 4-1. Student demographics grades K – 8.

Students	Sample Distribution (%)
By Race/Ethnicity	
African American	13.76
Hispanic	36.05
Caucasian	42.88
American Indian/Alaskan Native	0.42
Asian	4.83
Native Hawaiian/Other or Pacific Islander	0.42
Two or More Races	2.23
By Gender	
Male	51.29
Female	48.71
Free/Reduced Lunch	
Yes	47.86
No	52.14

Reliability

Assessments used to obtain criterion-related evidence of validity for the ISIP Math included STAR Math™, Test of Early Mathematics Ability – Third Edition (TEMA-3), Pearson's Stanford Achievement Test – Tenth Edition (SAT10), and the State of Texas Assessments for Academic Readiness (STAAR).

Internal Consistency

STAR Math assesses a similar construct as ISIP Math and ISIP Early Math and has a similar purpose. Therefore, it was selected to provide criterion-related evidence for ISIP Math. However, STAR Math was not used as a criterion assessment or benchmark.

- Internal consistency reliabilities ranged from .90–.95 across grades, with the test-retest coefficient ranging from .76–.84. Predictive and concurrent correlations ranged from moderate to strong, with predictive correlations ranging from r = .63-. 80, and concurrent correlations ranging from r = .57-.68.

This study selected the Test of Early Mathematics Ability – Third Edition (TEMA-3), which seeks to identify students significantly behind or ahead of peers in mathematical skills. It was used as a criterion assessment for kindergarten through 2nd grade students.

- The TEMA-3 is available in two parallel forms, Form A and Form B. Research indicates that internal consistency reliabilities for both forms are above .92. Test-retest estimates are .82 for Form A and .93 for Form B. Ginsburg and Baroody (2003) also found that items in Form A contained bias. Given these findings, Form B was selected for this study. Criterion validity coefficients ranged from r = .36-.71, with the majority of coefficients in the r = .50-.60 range.

The SAT10 online math assessment, with its web-based multiple-choice format, was selected for this study as a criterion assessment for students in grades 3 through 8.

- Internal consistencies range from .80-.87. Convergent validity coefficients range from r = .70-.80 across grade levels.

The STAAR is Texas's current testing program, with the mathematics STAAR being a mandatory EOY state assessment for students in grades 3 through 8. The format of the STAAR is multiple-choice items. It was also used as a criterion assessment to support inferences made from ISIP Math for grades 3 through 8.

- Internal consistency reliabilities for STAAR range from .81 -.93 across grade levels.

Validity Evidence

Technical adequacy data were collected to document the utility of ISIP Math in making screening decisions for students in kindergarten through 8th grade. The criteria used within this study were identified by the National Center on Response to Intervention (NCRTI) in 2010 and include:

- generalizability of the sample;
- classification accuracy of the performance level;
- reliability (of either the data or administrations of the assessment over time);
- evidence for validity; and
- evidence for reliability and validity disaggregated by relevant subgroup.

Furthermore, the items were calibrated under a two-parameter logistic item response theory (2PL-IRT) model. Item parameters were examined, and those items with unacceptable fit statistics with regards to the subtest which they measured were removed from the pool. Based on the combined processes used to establish content validity, the items in the operational pool grouped by subtest are believed to be accurate representations of the domain that they intend to measure.

Generalizability

Generalizability was analyzed as a way to illustrate the extent to which the analytic sample for the study was comparable to the state and national population.

Table 4-2. Comparison of demographics for the state, national, and recruited sample.

Students	Statewide Distribution[a] (%)	National Distribution[bcd] (%)	Sample Distribution (%)
By Race/Ethnicity			
African American	12.61	15.60	13.76
Hispanic	52.22	24.88	36.05
Caucasian	28.55	50.28	42.88
American Indian/Alaskan Native	0.39	1.05	0.42
Asian	4.03	5.18	4.83
Native Hawaiian/Other or Pacific Islander	0.14	—	0.42
Two or More Races	2.05	3.02	2.23
By Gender			

Students	Statewide Distribution[a] (%)	National Distribution[bcd] (%)	Sample Distribution (%)
Male	51.30	51.40	51.29
Female	48.70	48.60	48.71
Free/Reduced Lunch			
Yes	50.10	48.10	47.86
No	49.90	51.90	52.14

[a]Texas Education Agency (2015).

[b]U.S. Department of Education, National Center for Education Statistics, and Common Core of Data (2012).

[c]U.S. Department of Education, National Center for Education Statistics, and Common Core of Data (2016).

[d]U.S. Census Bureau (2014).

Concurrent Validity

Concurrent-related evidence for validity examines the relationship between performance on the screener and a criterion assessment with similar content that is administered at the same point in time. Concurrent-related evidence for validity at each administration of ISIP Math was calculated by determining the correlation between the scaled scores of ISIP Math for that administration and the scaled scores of the STAR Math for the same administration by grade level. Concurrent-related evidence for validity at the EOY administration of the ISIP Math was also calculated by determining the correlation — individually by grade level — between the scaled scores of the EOY ISIP Math and the scaled scores of the TEMA-3, SAT10 (and its two subtests), and the STAAR.

Table 4-3. Concurrent-related evidence for validity.

Assessment	Grade	n	Coefficient
STAR Math (BOY)	1	208	.66
	2	185	.76
	3	170	.71
	4	81	.64
	5	224	.55
	6	174	.74
	7	222	.61
	8	165	.61
STAR Math (MOY)	1	212	.77
	2	183	.81

Assessment	Grade	n	Coefficient
	3	169	.75
	4	69	.67
	5	198	.71
	6	173	.77
	7	199	.60
	8	167	.59
STAR Math (EOY)	1	213	.72
	2	181	.75
	3	167	.74
	4	81	.78
	5	235	.76
	6	162	.80
	7	211	.76
	8	145	.61
STAR Math (EOY)	K	152	.49
	1	210	.66
	2	195	.69
SAT10	3	196	.82
	4	131	.82
	5	250	.82
	6	197	.83
	7	146	.57
	8	152	.67
SAT10 PS	3	196	.82
	4	131	.82
	5	250	.75
	6	197	.83
	7	146	.45
	8	152	.65
SAT10 P	3	196	.69
	4	131	.71
	5	250	.78
	6	197	.74

Assessment	Grade	n	Coefficient
	7	146	.58
	8	152	.54
STAAR	3	190	.81
	4	129	.80
	5	241	.81
	6	234	.85
	7	192	.70
	8	130	.68

Discussion

Reliability and validity are two important qualities of measurement data. Reliability can be thought of as consistency, either consistency over items within a testing instance or over scores from multiple testing instances. Validity can be thought of as accuracy, either accuracy of the content of the items or of the constructs being measured. In this study, both qualities were examined using ISIP Math data collected from kindergarten through 8th grade students at three school districts in Texas during the 2015-2016 school year.

The sensitivity of ISIP Math for kindergarten through 2nd grade using TEMA-3 as the criterion assessment was between .80 and .92. In other words, between 80% and 92% of the students who were classified as “at-risk” on the TEMA-3 were also classified as “at-risk” on the EOY ISIP Math.

The specificity of ISIP Math for kindergarten through 2nd grade using TEMA-3 as the criterion assessment was lower at between .61 and .79. In other words, between 61% and 79% of the students who were classified as “not-at-risk” on the TEMA-3 were also classified as “not-at-risk” on the EOY ISIP Math. This also indicates that between 21% and 39% of students classified as “at-risk” on the ISIP Math were classified as “not-at-risk” on the TEMA-3.

The positive predictive value (PPV), or precision of classification, ranges from .90 to .97 across grades. This indicates that 90-97% of the students who were truly “at-risk” were classified as “at-risk” on both the ISIP Math and the TEMA-3. The negative predictive value (NPV) ranges from .29-.70 across grades, indicating that 29-70% of students who were truly “not-at-risk” were classified as “not-at-risk” on both the ISIP Math and the TEMA-3. The NPV value coincides with the large proportion of students who were classified as “at-risk” on the EOY ISIP Math and were classified as “not-at-risk” on the TEMA-3.

The accuracy of identification ranges from .80 to .89, indicating that the number of students correctly classified on the EOY ISIP Math with respect to the TEMA-3 was between 80% and 89% across all grades. The Area Under the Curve (AUC) indices range from .74 to .84 across grades. Using the guidelines suggested by Kettler et al. (2014), the AUC indices are moderate to high. Using the guidelines set by the NCRTI (2010), kindergarten and 2nd grade ISIP Math provide partially convincing evidence for classification accuracy based on TEMA-3, while 1st grade ISIP Math provides unconvincing evidence for classification accuracy based on TEMA-3.

One possible explanation for over-classification of "at-risk" students is that the cut score used for classification of "at-risk" and "not-at-risk" students on the TEMA-3 is the 20th percentile, while the cut score used for ISIP Math is the 25th percentile.

Taken together, the evidence supports the claim that ISIP Math produces reliable and valid data for measuring key areas of math skills development including number sense, operations, algebra, geometry, measurement, and data analysis.

Full Validity Study

To review the complete validity study — *Imagination Station (Istation): Istation's Indicators of Progress (ISIP) Math Validity Studies - Overview of Results* — visit the following link and click the link found under the **Technical Reports** heading.

http://www.smu.edu/Simmons/Research/RME/Explore/Publications

Chapter 5: Determining Norms

Norm-referenced tests are designed so that test administrators have a way of comparing the results of a given test taker to the hypothetical "average" test taker to determine whether they meet expectations. In the case of the Computerized Adaptive Testing (CAT)-based ISIP™ Math test, we are interested in comparing students to a national sample of students who have taken the ISIP Math test. We are also interested in knowing what the expected growth of a given student is over time, and in administering our test regularly to students to determine how they are performing relative to this expected growth.

By determining and publishing these norms, called Instructional Tier Goals, we enable teachers, parents, and students to see how students' scores compare with a representative sample of children in their particular grade for the particular period (month) in which the test is administered. The norming samples were obtained as part of Istation's ongoing research in assessing reading ability. The samples were drawn from enrolled ISIP Math users during the 2011–2015 school years in grades pre-K – 8. The state distributions for the sample are found in Table 5-1a and Table 5-1b.

Table 5-1a. State distributions and demographics for ISIP Math pre-K – 3 norming sample.

	Grade				
	Pre-K	**K**	**1st**	**2nd**	**3rd**
	Frequency (%)	**Frequency (%)**	**Frequency (%)**	**Frequency (%)**	**Frequency (%)**
Gender					
Female	1,988 (21.9)	9,168 (18.5)	10,994 (18.3)	18,325 (23.4)	1,703 (29.1)
Male	2,025 (22.4)	9,772 (19.7)	11,676 (19.4)	19,957 (25.5)	1,798 (30.8)
Special Education					
No	3,543 (67.5)	14,535 (29.3)	16,540 (27.6)	28,629 (36.6)	2,519 (43.2)
Yes	252 (2.7)	905 (1.8)	1,315 (2.3)	2,346 (3.0)	229 (3.9)
State					
Alaska	—	2 (0.004)	12 (0.02)	—	—
Alabama	21 (0.2)	170 (0.3)	410 (0.7)	1,726 (2.2)	452 (7.7)
Arizona	22 (0.2)	50 (0.1)	59 (0.1)	172 (0.2)	—
Arkansas	—	—	—	9 (0.1)	—
California	58 (0.6)	481 (0.9)	1,069 (1.8)	460 (0.6)	—
Colorado	—	356 (0.7)	457 (0.8)	307 (0.4)	—
Connecticut	—	—	4 (0.007)	—	—
Delaware	—	—	99 (0.2)	58 (0.1)	—

	Grade				
	Pre-K	K	1st	2nd	3rd
District of Columbia	—	—	—	1 (0.001)	—
Florida	210 (2.3)	4,949 (7.5)	5,938 (9.9)	4,486 (5.8)	15 (0.3)
Georgia	74 (0.8)	2,889 (5.8)	3,305 (5.5)	2,465 (3.2)	465 (7.9)
Hawaii	—	52 (0.1)	51 (0.1)	—	—
Idaho	—	22 (0.04)	16 (0.03)	18 (0.02)	—
Illinois	1 (0.01)	299 (0.6)	309 (0.5)	474 (0.6)	343 (5.8)
Indiana	—	326 (0.7)	314 (0.5)	252 (0.3)	—
Kansas	35 (0.4)	406 (0.8)	535 (0.9)	44 (0.1)	—
Kentucky	—	79 (0.2)	76 (0.1)	172 (0.2)	—
Louisiana	66 (0.7)	213 (0.4)	310 (0.5)	169 (0.2)	—
Maryland	158 (1.7)	571 (1.2)	717 (1.2)	555 (0.7)	26 (0.4)
Michigan	1 (0.01)	279 (0.6)	263 (0.4)	—	—
Minnesota	—	26 (0.05)	42 (0.07)	59 (0.1)	—
Missouri	—	36 (0.07)	44 (0.07)	84 (0.1)	11 (0.2)
Mississippi	29 (0.3)	102 (0.2)	135 (0.2)	54 (0.1)	—
Montana	36 (0.4)	304 (0.6)	310 (0.5)	453 (0.6)	36 (0.4)
North Carolina	1 (0.01)	746 (1.5)	846 (1.4)	994 (1.2)	422 (7.2)
North Dakota	22 (0.2)	49 (0.1)	66 (0.1)	86 (0.1)	40 (0.7)
Nevada	—	—	—	426 (0.5)	—
New Jersey	70 (0.7)	214 (0.4)	372 (0.6)	579 (0.7)	30 (0.5)
New York	58 (0.6)	119 (0.2)	111 (0.2)	122 (0.2)	170 (0.1)
Ohio	—	96 (0.2)	109 (0.2)	51 (0.1)	—
Oklahoma	—	1 (0.002)	20 (0.03)	16 (0.02)	—
Oregon	—	9 (0.02)	11 (0.02)	7 (0.01)	—
Pennsylvania	25 (0.3)	277 (0.6)	321 (0.5)	168 (0.2)	1 (0.01)
South Carolina	9 (0.1)	162 (0.3)	176 (0.3)	659 (0.8)	132 (2.3)
South Dakota	—	—	—	30 (0.04)	—
Tennessee	221 (2.4)	729 (1.5)	634 (1.1)	1,201 (1.5)	7,016 (3.9)
Texas	7,898 (87.4)	34,711 (70.0)	41,655 (69.4)	57,917 (74.2)	3,220 (55.2)
Virginia	14 (0.2)	459 (0.9)	733 (1.2)	3,647 (4.7)	670 (11.5)
Washington	—	13 (0.02)	3 (0.01)	—	—
West Virginia	—	—	—	52 (0.1)	—

	Grade				
	Pre-K	K	1st	2nd	3rd
Wisconsin	—	2 (0.004)	2 (0.01)	—	—
Wyoming	—	46 (0.1)	27 (0.04)	—	—

Table 5-1b. State distributions and demographics for ISIP Math grades 4 – 8 norming sample.

	Grade				
	4th	5th	6th	7th	8th
	Frequency (%)	Frequency (%)	Frequency (%)	Frequency (%)	Frequency (%)
Gender					
Female	4,896 (41.2)	5,399 (41.6)	321 (42.3)	1,207 (37.8)	866 (39.9)
Male	5,070 (42.6)	5,510 (42.5)	351 (46.3)	1,315 (41.3)	891 (41.0)
Special Education					
No	5,136 (43.2)	6,466 (49.8)	78 (10.2)	2,031 (63.8)	1,442 (66.4)
Yes	672 (5.6)	880 (6.7)	22 (2.9)	237 (7.4)	193 (8.8)
State					
Alaska	—	—	—	—	—
Alabama	—	—	22 (2.9)	1,048 (32.9)	693 (31.9)
California	163 (1.4)	141 (1.1)	—	—	—
Colorado	118 (1.0)	126 (1.0)	—	—	—
Florida	106 (0.9)	94 (0.7)	—	—	—
Georgia	175 (1.5)	211 (1.6)	—	148 (4.6)	1 (0.05)
Illinois	204 (1.7)	163 (1.3)	51 (6.7)	181 (5.7)	116 (5.3)
Kentucky	75 (0.6)	87 (0.7)	243 (32.0)	—	—
Louisiana	81 (0.7)	213 (0.4)	310 (0.5)	169 (0.2)	—
Missouri	—	—	—	4 (0.1)	2 (0.1)
Montana	—	—	—	226 (7.1)	280 (12.9)
North Carolina	431 (3.6)	496 (3.8)	—	—	—
North Dakota	—	—	21 (2.8)	30 (0.9)	15 (0.7)
Nevada	—	23 (0.2)	—	—	—
New Jersey	94 (0.8)	79 (0.6)	—	—	—
New York	162 (1.4)	230 (1.8)	—	—	—
Ohio	32 (0.3)	27 (0.2)	—	—	—

	Grade				
	4th	5th	6th	7th	8th
Oklahoma	29 (0.2)	2 (0.01)	—	—	—
South Carolina	728 (6.1)	693 (5.3)	—	—	—
Tennessee	666 (5.6)	533 (4.1)	—	—	—
Texas	7,821 (65.8)	9,049 (69.8)	407 (53.7)	1,543 (48.4)	1,064 (48.9)
Utah	27 (0.2)	1 (0.01)	—	—	—
Virginia	647 (5.4)	596 (4.6)	13 (1.7)	—	—
West Virginia	121 (1.0)	121 (0.9)	—	—	—

Sample

We updated the ISIP Math Instructional Tier Goals from August 2012 through August 2015 as various grades came online. Since that time, there has been substantial growth in the number of students using the ISIP Math assessment. Due to this growth in population, it will soon be necessary to establish a new norming sample in order to derive updated expected growth and goals that represent the current population of students using ISIP Math.

Students completing three assessments during the school year — in September (BOY), January (MOY), and May (EOY) — starting in 2011 were sampled from the total population to establish the norming sample. The total population by grade (N) and the sample size (n) by grade are found in Table 5-2a and Table 5-2b. In total, the ISIP Math scores from 44,847 students were considered to establish norms. This sample was used in establishing the Instructional Tier Goals for the ISIP Math ability score.

Table 5-2a. ISIP Math population and norm sample grades pre-K – 3.

ISIP	Size	Pre-K	K	1st	2nd	3rd
ISIP_BOY	N	1,140	12,817	20,564	47,468	60,865
	n	445	8,080	12,774	7,478	12,245
ISIP_MOY	N	3,916	27,307	33,554	31,842	39,527
	n	445	8,080	12,774	7,478	12,245
ISIP_EOY	N	5,493	31,956	37,917	24,772	31,760
	n	445	8,080	12,774	7,478	12,245

Table 5-2b. ISIP Math population and norm sample grades 4 – 8.

ISIP	Size	4th	5thK	6th	7th	8th
ISIP_BOY	N	8,064	8,712	333	1,930	982
	n	1,406	1,565	269	393	192
ISIP_MOY	N	6,241	6,938	427	2,013	1,575
	n	1,406	1,565	269	393	192
ISIP_EOY	N	4,172	4,021	587	816	512
	n	1,406	1,565	269	393	192

Computing Norms

Istation's norms are time-referenced to account for expected growth of students over the course of a semester. The ISIP Math test consists of an overall score for each grade. Each of these is normed separately so that interested parties can determine performance in various areas independently.

To determine these norms, percentiles were computed from the three assessment points collected and then interpolated for the months in between. Because of the test design, retakes of the test will result in different test items for a given student, so it is expected that improved scores on the test reflect actual growth over time. Norms were computed for each time period, so that over time a student's score on ISIP Math is expected to go up with student growth. Norming tables for Overall Math scores can be found at Istation's website, and these represent the results of norming the overall score across all test-taking periods. For each period, these scores were averaged and a standard deviation was computed. Then, to determine expected Tier 2 and Tier 3 scores, the 20th and 40th percentiles on a true normal bell curve were computed, and these numbers are given as norms for those tier groups.

Instructional Tier Goals

Consistent with other assessments, Istation has defined a three-tier normative grouping, based on scores associated with the 20th and 40th percentiles. Students with a score above the 40th percentile for their grade are placed in Tier 1. Students with a score at or below the 20th percentile are placed in Tier 3.

These tiers are used to guide educators in determining the level of instruction for each student. Students classified as:

- Tier 1 are on track to meet grade-level expectations;
- Tier 2 are at some risk of not meeting grade-level expectations; and
- Tier 3 are at significant risk of not meeting grade-level expectations.

References

Online

Quantile Framework Overview: This brief video on the basic concepts and use of The Quantile Framework for Mathematics is the perfect resource for educators and parents who are new to Quantile measures.
https://youtu.be/Ne5a9KtbAU8

Math Skills Database: Search the Math Skills Database for Quantile Skills and Concepts (QSCs) using state standards or Common Core Standards. The database contains targeted, free resources appropriately matched to students by Quantile measures and content.
https://quantiles.com/tools/math-skills-database/

Publications

Cao, J. & Stokes, S. L. (2006). Bayesian IRT guessing models for partial guessing behaviors, manuscript submitted for publication.

Conte, K. L., & Hintz, J. M. (2000). The effect of performance feedback and goal setting on oral reading fluency with CBM. *Diagnostique*, *25*, 85-98.

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

Deno, S. L. (1985). Curriculum based measurement: The emerging alternative. *Exceptional Children, 52,* 219-232.

Espin, C., Deno, S., Maruyama, G. & Cohen, C. (1989). The basic academic skills samples (BASS): An instrument for the screening and identification of children at-risk for failure in regular education classrooms. Paper presented at the annual American Educational Research Association Conference, San Francisco, CA.

Foorman, B. R., Santi, K., & Berger, L. (in press). Scaling assessment-driven instruction using the Internet and handheld computers. In B. Schneider & S. McDonald (Eds.), *Scale-up in education, vol. 1: Practice*. Lanham, MD: Rowan & Littlefield Publishers, Inc.

Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinction between instructionally relevant measurement models. *Exceptional Children*, *57*, 488-500.

Fuchs, L. S., Deno, S. L., & Marston, D. (1983). Improving the reliability of curriculum-based measures of academic skills for psycho education decision making. *Diagnostique*, *8*, 135-149.

Fuchs, L. S., Deno, S. L., & Mirkin, P. K. (1984). The effects of frequent curriculum-based measurement and evaluation on pedagogy, student achievement, and student awareness of learning. *American Educational Research Journal*, *21*, 449-460.

Fuchs, D., & Fuchs, L. S. (1990). Making educational research more important. *Exceptional Children*, *57*, 102-108.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Ferguson, C. (1992). Effects of expert system consultation within curriculum-based measurement using a reading maze task. *Exceptional Children*, 58, 436-450.

Fuchs, L. S., Hamlett, C., & Fuchs, D. (1995). *Monitoring basic skills progress: Basic reading – version 2* [Computer program]. Austin, TX: PRO-ED.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Stecker, P. M. (1991). Effects of curriculum-based measurement and consultation on teacher planning and student achievement in mathematics operations. *American Educational Research Journal*, *28*, 617-641.

Gilks, W. R., & Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics, 41*, 337-348.

Hatfield, C., Perry, L., Basaraba, D., & Ketterlin-Geller, L. R. (2015). Imagination Station (Istation): Universal screener instrument development for grades PK-1 (Tech. Rep. No. 15-01) Dallas, TX: Southern Methodist University, Research in Mathematics Education.

Hatfield, C., Perry, L., Basaraba, D., Miller, S. J., Simon, E., Ketterlin-Geller, L. R. (2014). Imagination Station (Istation): Universal screener and inventory instruments interface development for grades PK-1 (Tech. Rep. No. 14-01) Dallas, TX: Southern Methodist University, Research in Mathematics Education.

Jenkins, J. R., Pious, C. G., & Jewell, M. (1990). Special education and the regular education initiative: Basic assumptions. *Exceptional Children*, *56*, 479-491.

Marston, D. B. (1989). *A curriculum-based measurement approach to assessing academic performance. What is it and why do it?* New York. Guillford Press.

Mathes, P. G., Fuchs, D., Roberts, P. H., & Fuchs, L. S. (1998). Preparing students with special needs for reintegration: Curriculum-based measurement's impact on transenvironmental programming. *Journal of Learning Disabilities*, 31(6), 615-624.

National Council of Teachers of Mathematics. (2006). *Curriculum focal points for prekindergarten through grade 8 mathematics: A quest for coherence*. Reston, VA: The National Council of Teachers of Mathematics, Inc.

Stecker, P. M., & Fuchs, L. S. (2000). Effecting superior achievement using curriculum-based measurement: The importance of individual progress monitoring. *Learning Disabilities Research and Practice*, *15*, 128-134.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361-370.