Istation’s Indicators of Progress (ISIP)™
Technical Report 2021

ISIP Math

Technical Report

by Leanne Ketterlin-Geller

Computer adaptive testing system for continuous progress monitoring of math growth for students prekindergarten through grade 8.

Istation's Indicators of Progress (ISIP)™ Math Technical Report 2021

by Leanne Ketterlin-Geller

Istation

Imagination Station – Istation

2000 Campbell Center II
8150 North Central Expressway
Dallas, Texas 75206
(214) 237-9300

Table of Contents

Figures iii

Tables iii

Acknowledgements .... **vii**

Chapter 1: Introduction **9**

The Need to Link Math Assessment to Instructional Planning 10

Continuous Progress Monitoring 11

Computer-Adaptive Testing 12

ISIP Math Domains 14

Teacher Friendly 16

Student Friendly 17

ISIP Math and Instructional Planning 17

Goals of the ISIP Math Update 19

Chapter 2: Item Writing and Item Properties **20**

ISIP Math Items 20

Item Writing Procedures 21

ISIP Math DIF Analysis for Prekindergarten through Grade 8 22

Chapter 3: IRT Calibration and the CAT Algorithm **25**

The IRT Model 25

Grades Prekindergarten–1 27

Grades 2-8 29

CAT Algorithm 33

Chapter 4: Math Domains **35**

Identifying Domains 35

Chapter 5: Vertical Scaling **43**

Introduction to Vertical Scaling 43

ISIP Math Vertical Scaling Data Collection Design 45

Vertical Scaling Study Methods 52

Appendix 5.1: Item Calibration Using Ancillary Information *64*

Appendix 5.2: Adjusting for Measurement Error *66*

Appendix 5.3: Transforming Logistic Regression Standard Errors to IRT Standard Errors Using the Delta Method *69*

Appendix 5.4: Deriving linking constants *72*

Chapter 6: Norming 74

Introduction to Norming *74*

Sampling Procedures *75*

Norming Analysis *84*

Chapter 7: Reliability and Validity 94

Internal Consistency *94*

Validity Evidence *95*

Concurrent Validity *97*

ISIP Math Predictive Validity *101*

Student Growth Expectations for Regular Education, Special Education, and Students with Disabilities *107*

Conclusion *110*

Chapter 8: Growth 111

Introduction *111*

Expected Growth *112*

References 126

Figures

Figure 1.1. Process used in a Computer Adaptive Test 14

Figure 4.1. Composition of Domain Scores and the Overall Score 42

Figure 5.1. Schematic of Common-Item Nonequivalent Groups Design For Linking Together Fixed-Forms of a Test 46

Figure 5.2. Detail of Common-Item Nonequivalent Groups Design for Linking Together the Item Pools of a Computer-Adaptive Test 47

Figure 5.3. A Simplified Schematic of the Original Data Collection Design for ISIP Math Vertical Scaling 48

Figure 5.4. Final data collection design for ISIP Math vertical scaling 51

Figure 5.4. Schematic of Pairwise and Anchor-Grade Linkages of Item Pools in the ISIP Math Vertical Scaling 58

Figure 5.5. ISIP Math final reporting growth curves: Scale score means and confidence bands by grade level (Source: ISIP Math January 2020 test administration data)...... 62

Figure 6.1. Example: ISIP Math Grade 1 Normative Sample: Boxplots by Time of Year 89

Figure 6.2. Example ISIP Math Grade 1 Normative Sample: Histograms of Scaled Scores by Time of Year with Normal Distribution Overlay 90

Tables

Table 2.1. *Differential Item Functioning (DIF factors) for Gender* 24

Table 2.2. *Differential item functioning (DIF Factors) for Race/Ethnicity* 24

Table 3.1. *Demographics of Students in IRT study Grades Pre-K to 1* 28

Table 3.2. *Item Pools by Grade* 29

Table 3.3. *Demographics of Students in IRT study Grades 2 to 8* 31

Table 3.4. *Sub Contents Measured for Each Grade in the ISIP Math IRT Study* 32

Table 4.1. *Math Domains for ISIP Math Grades Pre-K to 1* 38

Table 4.2. *Math Domains by Grade for ISIP Math Grades 2 to 5* 39

Table 4.3. *Math Domains by Grade for ISIP Math Grades 6 to 8* 40

Table 5.1. *Mitigation Strategies Used to Deal with Data Collection Issues in the ISIP Math Vertical Scaling Study* 50

Table 5.2. *Summary of the Numbers of Item Sets, Items, and Students Used in the Vertical Scaling Study by Grade Band* 52

Table 5.3. *Initial and Final Numbers of Items Used to Create Equating Constants After Item and Robust Z Flagging by Grade Band* 57

Table 5.4*. Linking Constants (Slopes and Intercepts) From Starting to Anchor Grade* 60

Table 5.5. *ISIP Math Vertical scale grade-to-grade growth, variability, and distribution separation (Source: ISIP Math January 2020 test administration data)* . 63

Table 6.1. *Information from NCES Enrollment Data Used in Construction of the Composite Index* 78

Table 6.2. *Results from the Regression Model for Constructing the School Stratification Index* 79

Table 6.3. *Means of Midyear Original ISIP Math Scores by Grade and School Index Octile Levels* 80

Table 6.4. *Percent of Public School Students by School Index Octile and of Private/Parochial School Students in Overall- and Domain-Score Samples*............... 83

Table 6.5. *ISIP Math Norms Developed by Grade and Domain* 84

Table 6.7. *ISIP Math Grade 1 Normative Sample: Overall Scale Score Descriptive Statistics by Time of Year* 88

Table 6.8. *Comparison of Ohio AIR Proficiency Rates, and the Percent of Students Who Reached Tier 1 in the Previous and Updated Norms and Met the Proficient Level or Higher* 93

Table 7.1. *Comparison of Demographics for Race/Ethnicity for the State, National, and Recruited Sample* 96

Table 7.2. *Comparison of Demographics for Free/Reduced Priced Lunch Status for the State, National, and Recruited Sample* 97

Table 7.3. *Comparison of Demographics for Gender for the State, National, and Recruited Sample.* 97

Table 7.5. *Demographic Characteristics of the ACT Aspire Study* 102

Table 7.6. *Pearson Product-Moment Correlation Coefficients between ISIP and ACT Aspire* 102

Table 7.7. *Probabilities of reaching Ready or Higher on the ACT Aspire at Key Percentiles for 3rd Grade EOY* 103

Table 7.8. *Demographic Characteristics of the Ohio AIR Study* 104

Table 7.7. *Correlations Between ISIP Math and Ohio AIR* 105

Table 7.8. *Probability of Reaching Proficient or Higher on the Ohio AIR at Key Percentiles of ISIP Math* 106

Table 7.9. *Mean and Gain Scores for Students in General and Special Education, 2018-2019 School Year* 108

Table 7.10. *Mean and Gain Scores for Students in Special Education by Type of Disability, 2018-2019 School Year* 109

Table 8.1: Normative Growth for ISIP Math Overall for Grades Prekindergarten to 3, by Grade and Decile at the Beginning of the Year (September to April) 113

Table 8.2: Normative Growth for ISIP Math Overall for Grades Prekindergarten to 3, by Grade and Decile at the Beginning of the Year (September to May) 113

Table 8.3: Normative Growth for ISIP Math Overall for Grades 4 to 8, by Grade and Decile at the Beginning of the Year (September to April) 114

Table 8.4: Normative Growth for ISIP Math Overall for Grades 4 to 8, by Grade and Decile at the Beginning of the Year (September to May) 114

Table 8.5: Pre-K Change in Performance Categories BOY-EOY for ISIP Math Overall by Decile Category 117

Table 8.6: Kindergarten Change in Performance Categories BOY-EOY for ISIP Math Overall 117

Table 8.7: First Grade Change in Performance Categories BOY-EOY for ISIP Math Overall 118

Table 8.8: Second Grade Change in Performance Categories BOY-EOY for ISIP Math Overall 118

Table 8.9: Third Grade Change in Performance Categories BOY-EOY for ISIP Math Overall 119

Table 8.10: Fourth Grade Change in Performance Categories BOY-EOY for ISIP Math Overall 119

Table 8.11: Fifth Grade Change in Performance Categories BOY-EOY for ISIP Math Overall 120

Table 8.15: Six through Eighth Grade Combined Change in Performance Categories BOY-EOY for ISIP Math Overall 120

Table 8.13: Expected Growth Pathways (Gains) for ISIP Math Overall BOY-EOY by ISIP Instructional Levels 124

Acknowledgements

The Istation Indicators of Progress-Math (ISIP Math) assessment was first developed under the guidance of our author, Dr. Leanne Ketterlin-Geller of Southern Methodist University. Dr. Ketterlin-Geller and her team put together a formative assessment that is based on the Curriculum Focal Points published by the National Council of Teachers of Mathematics (NCTM). ISIP Math incorporates the content and cognitive engagement dimensions of assessment. The content dimension refers to the knowledge and skills that form the basis of mathematics instruction. The cognitive engagement dimension refers to the level of cognitive processing through which students are expected to engage with the content. Incorporating this framework makes the ISIP Math unique, and for that we thank Dr. Ketterlin-Geller and her team in the Simmons School of Education and Human Development for an innovative framework for assessing mathematics. We especially note the contributions of Diane Gifford, Deni Basaraba, Lindsay Perry, Savannah Hill, Pooja Shivraj, and Cassandra Hatfield.

In this normative update, we had several goals including expanding the conceptual framework by providing domain scores, composing a vertical scale, and offering new, more rigorous norms.

Math specialists Morgan Hamilton, Stephanie Noland, Melinda Smith, Jennifer Newell, Darey Steacy-Spigener, Sonia Blanks, and Kelly Prunty established the domains by reviewing the literature and aligning them with the skills and expectations from NCTM. They reviewed all items and assigned them to the domains. Dr. Bill Fahle, PhD, worked on the computer adaptive testing algorithm to provide students with sufficient items for calculating domain scores, and changed the algorithm to provide both overall and subtest scores.

To compose the vertical scale, we worked under the consultation and guidance of Dr. Michael Young, PhD, of Michael J. Young Consulting, LLC. He advised us on the research plan for the scaling items, which were selected by Kelly Prunty and Sonia Blanks. Bill Fahle worked on the implementation of the scaling and linking items into the live assessments. The critical task of composing the new vertical scale was completed by Michael Young, whose decades of experience and expertise in creating vertical scales yielded a single, common scale for ISIP Math across grades Prekindergarten to eighth.

The updated norms were also composed by Michael Young, using the stratified samples created by the research team including psychometrician Dr. Chalie Patarapichayatham, PhD, and statistical analyst, Sean Lewis. The new norms yielded results that were representative, and were more rigorous than the previous version. Our thanks go to Michael, whose work and expertise were critical in ensuring that the new scale and norms provided accurate norms that will better identify students' math abilities. He also wrote the chapter in this manual on the vertical scaling, and also wrote the sections on the normative update, detailing the norming procedures.

In addition to the technical aspect of composing new norms and scales, there are dozens of people working within Istation whose diligence and expertise enhanced this effort. Our systems teams had the task of accommodating the new features. Led by Bill Lowrey, Bill Fahle, and Zachary Terry, our assessments and reporting teams worked diligently to ensure that the reports ran correctly, and that the new scores were delivered accurately. David Pearson led the team of Quality Assurance analysts, who checked the reports and whose work ensured that all of the information reported was accurate and complete. When putting together this technical manual, Chalie Patarapichayatham conducted the updated analyses in the validity sections, and the work of our copyediting team of Moniqa Paullett and Ross Frazier enhanced its readability.

Finally, sincere appreciation goes to our President and Chief Operating Officer, Ossa Fisher, and to our chairman and Chief Executive Officer, Richard Collins. Ossa and Dick provide steady guidance and enthusiastically support changes and innovation in our products that help better the lives of the students and teachers we serve.

Victoria Locke, PhD
Vice President Research and Assessment
Istation

Chapter 1: Introduction

Istation's Indicators of Progress for Math (ISIP™ Math) is a sophisticated, web-delivered, computer-adaptive testing (CAT) system that provides continuous progress monitoring (CPM) in the subject area of mathematics.

ISIP 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 to an entire classroom, school, or district in a single day. Classroom and individual student results illustrating each student's past and present performance on mathematical concepts are available to teachers in real time. 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.

This technical report describes the ISIP Math assessment, and the changes made since the assessment was released in 2015. The original version of ISIP Math divided the assessment into ISIP Early Math and ISIP Math. In this revision, we will refer to the assessment only as ISIP Math.

For students in prekindergarten through first grade, ISIP Math provides a fun and engaging computer-based universal screener designed to help teachers identify students struggling to learn critical mathematics content. Beginning in "Mario's Market," students use their math skills in a real-life setting. ISIP 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 make informed decisions about each student's response to targeted mathematics instruction and intervention strategies.

For students in grades 2 through 8, the assessment provides a testing format that is familiar to most students. Each item contains a question stem and four answer choices.

ISIP Math provides links to teaching resources and targeted intervention strategies for all grade levels. 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 better 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 teachers can identify students at risk of mathematics failure and differentiate instruction, they must first have information about the specific needs of their students. To effectively link assessment with instruction, math assessments must accomplish the following:

- identify students at risk of having difficulty in math (i.e., students who 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 meet 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, gathering assessments across multiple time points means student performance is more likely to reflect actual ability. Using the computer also reduces inaccuracies related to human administration errors.

Collecting 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.

Some schools use more labor-intensive methods, such as one-on-one administration of progress monitoring. These assessments 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 were expensive for school districts to administer 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 receives and is able to review the results of the assessment, making manual assessments less than ideal for planning instruction.

Continuous Progress Monitoring

ISIP Math grew out of the model of continuous progress monitoring (CPM) called Curriculum-Based Measurement (CBM), which 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 that reflect 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 handheld 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 by teachers and administrators in 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 on growth patterns, eliminating the need for the teacher to manually create monitoring booklets or analyze results.

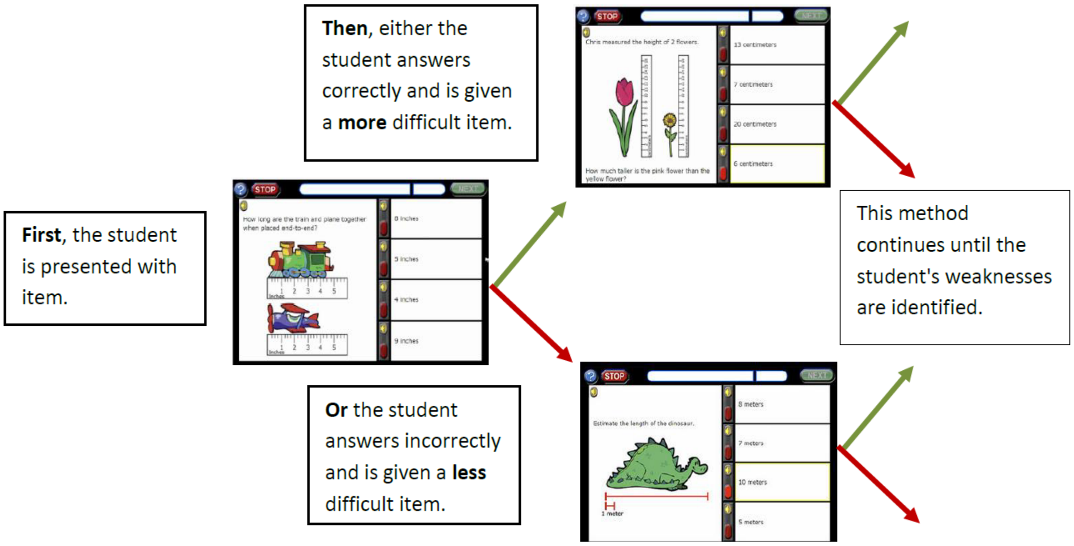
Computer-Adaptive Testing

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 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.



First, the student is presented with item.

Then, either the student answers correctly and is given a **more** difficult item.

Or the student answers incorrectly and is given a **less** difficult item.

This method continues until the student's weaknesses are identified.

How long are the train and plane together when placed end-to-end?

Chris measured the height of 2 flowers.

How much taller is the pink flower than the yellow flower?

Estimate the length of the dinosaur.

Figure 1.1. *Process used in a Computer Adaptive Test*

ISIP Math is delivered at established intervals (usually monthly) to the appropriate grade level for each student throughout a nine-month school year. This provides the 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 Domains

Designed for students in prekindergarten through 8th grade, ISIP Math provides 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 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.

This method continues until the student's weaknesses are identified. First, the student is presented with an item. Then, either the student answers correctly and is given a more difficult item, or the student answers incorrectly and is given a less difficult item.

ISIP 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 before using them to solve mathematics problems or questions. 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 measurement (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.

Teacher Friendly

ISIP Math assessments 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 who may be experiencing difficulties as they complete the assessment. Teachers subsequently review student results to validate outcomes. For students whose skills may be a concern, based on performance level, teachers may easily validate student results by re-administering the entire ISIP Math assessment as an On-Demand assessment.

Student Friendly

ISIP Math is also student friendly. Each assessment session in ISIP Math for prekindergarten through first grade gives students the experience 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 assessment interfaces that allow as much hands-on/manipulative-based interaction as possible. The singular interface theme of Mario's Market minimizes student distractions and unnecessary cognitive load.

Similarly, each assessment session in ISIP Math for grades 2 through 8 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 ISIP Math and the instruction in Istation Math. Additionally, it provides motivation for students to do their best when completing the assessment.

ISIP Math and Instructional Planning

ISIP Math provides continuous assessment results that can be used in recursive assessment instructional decision loops. After students complete the assessment, the results will help teachers identify students in need of support. If the results are in question based on a student's previous achievement, 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, teachers can compare and evaluate results. When the On-Demand feature is used, the assessment will be automatically administered the next time a student logs in.

The delivery of student results facilitates the evaluation of curriculum and instructional plans. The technology behind ISIP 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 automatically alerts 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), in Response to Intervention (RTI), and in 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. Student results from Istation Math are combined with ISIP 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 results is available. Administrators, principals, and teachers may use these reports to evaluate and modify curriculum and intervention strategies and evaluate personnel performance and the effectiveness of professional development.

Goals of the ISIP Math Update

Istation had several goals for this update of ISIP Math. For one, ISIP Math gave an overall score, but there was a need to see how students were doing in the different math domains. Chapter 4 goes into depth regarding how the ISIP Math domains were constructed and the care and attention that was paid into the creation of the domains.

Further, the previous edition of ISIP Math had a separate scale for each grade, and the items were discrete across the different grades, whereas many school districts requested a longitudinal scale. Using vertical scaling constants, we now have a scale that is continuous from prekindergarten through eighth grade. These changes are described in chapter 5.

Finally, the student population has changed since ISIP Math was first released. Updated, more rigorous norms were needed to help schools identify students who were struggling in math. We updated the norms using the vertical scale, and using data from the extensive Istation database, we constructed a sample that uses post stratification methods to make the sample representative of the student population in the US. These changes are described in chapter 6.

Chapter 2: Item Writing and Item Properties

ISIP 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 et al., 2012).
    - ISIP Math assesses both proficiency in mathematical concepts and students' level of cognitive engagement.

The strands of proficiency for cognitive engagement include Strategic Competence, Adaptive Reasoning, Procedural Fluency, and Conceptual Understanding. The mathematical domains that are covered include:

1. number sense
2. operations
3. algebra
4. geometry
5. measurement
6. data analysis
7. probability and statistics
8. Ratios and Proportional Relationships

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

- 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 et al., 2012).

Item Writing Procedures

All items were written under the supervision of Leanne Ketterlin-Geller, PhD, professor at Southern Methodist University. This technical report provides a brief description of the item writing process. To access the technical reports for the Universal Screener Instrument Development for each grade level (prekindergarten through 8), refer to the external links provided at the end of this report (Hatfield et al., 2015).

All item writers had expertise and experience teaching mathematics at the grade level for which they were selected to write. Before beginning the process, item writers attended training and received a style guide. The guide provided explanations, examples, and stylistic expectations of items to support writing high-quality mathematics items. It also included information on the cognitive levels of engagement. The training covered an overview of the assessment, a review of elements of high-quality test design relating to fairness, reliability, and validity in testing, and information on the test blueprint. Each item writer was paired with a staff reviewer (Hatfield et al., 2015).

After the items passed the first review, experts reviewed the items for accuracy, precision, and appropriateness of the distractors and then rated each item as Extremely Accurate, Appropriate or Mostly Accurate, Somewhat Accurate/Appropriate, or Not at all Accurate/Appropriate. These expert reviewers made recommendations for revisions, including changes to distractors, and the corrections were then made to the items (Hatfield et al., 2015).

Next, mathematics teachers reviewed the items' appropriateness of language, mathematical vocabulary, content or concepts, distractors, and art and design. The teachers also analyzed the items for bias in language or content. Specifically, they were asked if the item required background knowledge unrelated to the concept that would differ for students with different backgrounds who might be unfamiliar with the terms or concepts in the items. The reviewers then rated each item as biased, somewhat biased, or not biased. In instances where they rated the items as biased, they were asked to provide recommendations to improve the item (Hatfield et al., 2015).

ISIP Math DIF Analysis for Prekindergarten through Grade 8

To update the bias information in this revision, we conducted a study to determine differential item functioning, or DIF, of the most commonly used items at the middle-of-the-year assessment.

DIF can be described as the difference in an item's difficulty between subgroups of examinees who have the same ability level on the trait being measured (Patarapichayatham et al., 2012). DIF occurs when an item in a test functions differently for different groups, given the same ability level. DIF is an important psychometric property to display fairness for achievement tests. There are many ways to detect DIF in a test. The logistic regression DIF detection method is applied in this study to detect uniform DIF. Swaminathan and Rogers proposed logistic regression in 1990 as an alternative to the Mantel-Haenszel test to detect DIF. Logistic regression is a generalized linear model to calculate the probability of giving a correct answer to a dichotomous item given a score and group membership.

The original data file had over 3,000 items in prekindergarten through grade 8. Two DIF factors were investigated: gender (male/female) and race/ethnicity (Non-Hispanic White/all other combined). The data were extracted from the January assessment month of the 2018-2019 school year for prekindergarten through grade 8 students who completed the ISIP Math assessment.

Because DIF analyses require complete data on the DIF factors and demographic variables are not a requirement in the Istation system, students who did not have gender and race/ethnicity in the database were excluded. Items that had less than 100 responses were also discarded. The final data file consisted of 1,682 items from 185,048 students. There were 7,856 prekindergarteners, 31,920 kindergarteners, 34,628 first graders, 34,972 second graders, 24,365 third graders, 22,794 fourth graders, 21,204 fifth graders, 3,717 sixth graders, 2,044 seventh graders, and 1,548 eighth graders.

The logistic regression DIF detection analyses by difR package were used. ISIP Math scale scores were used as matching criteria. The analysis was conducted separately for each grade and each DIF factor, totaling 20 analyses conducted in this study.

The difR obtained two DIF detection criterions: Zumbo & Thomas (ZT) and Jodoign & Gierl (JG). Both criterions had the same procedure but different cut points. There are three DIF effect sizes: A – negligible or non-significant DIF effect, B – slightly to moderate DIF effect, and C – moderate to large DIF effect. The DIF effect size under Zumbo & Thomas (ZT) is as follows: $0 < A \leq 0.13$, $0.13 < B \leq 0.26$, and $0.26 < C \leq 1$Jodoign & Gierl (JG) is much smaller: $0 < A \leq 0.035$, $0.035 < B \leq 0.07$, and $0.07 < C \leq 1$.

Results show that all items displayed as A item (negligible or non-significant DIF effect) with ZT DIF criterion. Under JG DIF criterion (see Table 1), results show that approximately 98% displayed as A item (negligible or non-significant DIF effect); 1% displayed as B item (slightly to moderate DIF effect); and less than 1% displayed as C item (moderate to large DIF effect) for both DIF factors. To be more specific, 15 items displayed as B item and three items displayed as C item with gender DIF factor. Seventeen (17) items displayed as B item and three items displayed as C item with race/ethnicity DIF factor.

Table 2.1. *Differential Item Functioning (DIF factors) for Gender*

Grade	A item	Frequency	B item	Frequency	C item	Frequency
Pre-K	136	99.27%	1	0.73%	0	
K	220	100.00%	0		0	
1	209	98.12%	4	1.88%	0	
2	264	100.00%	0		0	
3	208	99.04%	1	0.48%	1	0.48%
4	192	98.96%	1	0.52%	1	0.52%
5	172	98.85%	2	1.15%	0	
6	146	98.64%	1	0.68%	1	0.68%
7	149	96.75%	5	3.25%	0	
8	145	100.00%	0		0	

Table 2.2. *Differential item functioning (DIF Factors) for Race/Ethnicity*

Grade	A item	Frequency	B item	Frequency	C item	Frequency
Pre-K	137	100.00%	0		0	
K	217	98.64%	3	1.36%	0	
1	212	99.53%	1	0.47%	0	
2	264	100.00%	0		0	
3	208	99.04%	1	0.48%	1	0.48%
4	192	98.96%	1	0.52%	1	0.52%
5	174	100.00%	0		0	
6	142	100.00%	6	4.05%	0	
7	150	97.40%	3	1.95%	1	0.65%
8	143	98.62%	2	1.38%	0	

Chapter 3: IRT Calibration and the CAT Algorithm

The first step in any construction of a computer-adaptive test (CAT) is to collect information about the items' discrimination and difficulty. The goals of this study were to determine the appropriate item response theory (IRT) model, estimate item-level parameters, and tailor the CAT algorithms, such as the exit criteria.

The IRT Model

A two-parameter logistic IRT (Item Response Theory) model (2PL IRT) was posited. We defined the binary response data, x_{ij}, with index $i = 1, \ldots n$ for persons, and index $j = 1, \ldots j$ 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 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 $\tau\theta$ and $\tau\delta$ 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:

$\lambda \sim Gamma(a_\lambda, b_\lambda), j = 1, \ldots j.$

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).

Grades Prekindergarten–1

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

During the 2014-2015 school year, data were collected from schools across the country so that ISIP™ Math for students in prekindergarten through grade 1 would be available for schools in the 2015-2016 school year. All students in prekindergarten through first grade were invited to participate, including students with disabilities and English learners (EL). There were no specific demographic requirements 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 first 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 first 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 first grade test forms were only taken by first 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 first graders did provide such information. Gender is reported as male or female, and ethnicity is reported as African American, American Indian, Asian, Hispanic/Latino, White, or Unknown. Special education status (SPED) is divided into whether or not the student was receiving services. Students on free or reduced priced lunch (FRPL), and whether or not the student was receiving services for English as a second language (ESL). The information from these students is reported in Table 3-1.

Table 3.1. *Demographics of Students in IRT study Grades Pre-K to 1*

Demographics	Prekindergarten	Kindergarten	Grade 1
Gender: Male	500 (49.7%)	299 (53.8%)	372 (52.8%)
Gender: Female	506 (50.3%)	257 (46.2%)	333 (47.2%)
Ethnicity: African American	778 (77.3%)	107 (19.2%)	133 (18.9%)
Ethnicity: American Indian	3 (.3%)	4 (.7%)	5 (.7%)
Ethnicity: Asian	2 (.2%)	8 (1.4%)	4 (0.6%)
Ethnicity: Hispanic/Latino	12 (1.2%)	102 (18.3%)	7 (1.0%)
Ethnicity: White	172 (17.1%)	298 (53.6%)	277 (39.3%)
Ethnicity: Unknown	39 (3.9%)	37 (6.7%)	279 (39.6%)
SPED: Yes	41 (4.1%)	8 (1.4%)	10 (1.4%)
SPED: No	1 (.1%)	79 (14.2%)	175 (24.8%)
FRPL: Yes	10 (1.0%)	74 (13.3%)	106 (15.0%)
FRPL: No	1 (.1%)	79 (14.2%)	175 (24.8%)
ESL: Yes	10 (1.0%)	1 (.2%)	6 (.9%)
ESL: No	1 (.1%)	152 (27.3%)	274 (38.9%)

Regarding the content of the items, multiple sub-contents are measured for each grade. The item pools by grade are available in Table 3.2.

Table 3.2. *Item Pools by Grade*

Prekindergarten

- Counting Skills
- Number Sense
- Number and Operations
- Counting and Cardinality
- Adding to/Taking Away Skills
- Geometry
- Algebra and Functions
- Algebra
- Spatial Relations
- Measurement
- Measurement Skills
- Data Analysis
- Mathematical Reasoning
- Data collection and statistics
- Patterns and Seriation
- Patterns and Relationships

Kindergarten

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

First Grade

- Number Sense
- Operations and Algebraic Thinking
- Measurement and Data
- Patterns
- Functions
- Number and Operations
- Number and Operations in Base Ten
- Algebraic Reasoning
- Measurement and Data Analysis
- Measurement
- Data Analysis
- Personal Financial Literacy

Overall, most items were good quality in terms of item discriminations and item difficulties. For prekindergarten, 5 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 first grade, 35 items were removed and 360 calibrated item parameters remain in the item pool.

Grades 2-8

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

Tests were administered by computer to groups in a classroom or computer lab setting. There were 940 items for grade 2; 1,066 items for grade 3; 875 items for grade 4; 882 items for grade 5; 1,159 items for grade 6; 938 items for grade 7; and 616 items for grade 8. 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, meaning second grade test forms were only taken by second grade students, third grade test forms were only taken by third 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 second graders; 5,127 third graders; 5,832 fourth graders; 5,067 fifth graders; 6,347 sixth graders; 1,537 seventh graders; and 1,169 eighth graders. The information from these students is reported in Table 3.3.

Table 3.3. *Demographics of Students in IRT study Grades 2 to 8*

Demographics	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Gender: Male	1548 (39.3%)	1726 (33.7%)	2094 (35.9%)	1704 (33.6%)	2700 (42.5%)	761 (49.5%)	585 (50.0%)
Gender: Female	1336 (33.9%)	1679 (32.7%)	2049 (35.1%)	1577 (31.1%)	2617 (41.2%)	760 (49.4%)	572 (48.9%)
Ethnicity: African American	813 (20.7%)	467 (9.1%)	989 (17.0%)	612 (12.1%)	1292 (20.4%)	197 (12.8%)	203 (17.4%)
Ethnicity: American Indian	32 (0.8%)	23 (0.5%)	13 (0.2%)	20 (0.4%)	61 (1.0%)	8 (0.5%)	5 (0.4%)
Ethnicity: Asian	64 (1.6%)	53 (1.0%)	184 (3.2%)	200 (3.9%)	140 (2.2%)	13 (0.8%)	18 (1.5%)
Ethnicity: Hispanic/Latino	743 (18.9%)	117 (2.3%)	120 (2.1%)	131 (2.6%)	215 (3.4%)	111 (7.2%)	88 (7.6%)
Ethnicity: White	1137 (28.6%)	1484 (28.9%)	1750 (30.0%)	1710 (33.7%)	1830 (28.8%)	755 (49.1%)	664 (56.8%)
Ethnicity: Unknown	1148 (29.2%)	2978 (58.1%)	705 (12.1%)	2394 (47.2%)	2809 (44.3%)	453 (29.5%)	191 (16.3%)
SPED: Yes	246 (6.2%)	212 (4.1%)	289 (5.0%)	236 (4.7%)	643 (10.0%)	112 (7.3%)	109 (9.3%)
SPED: No	2401 (61.0%)	2474 (48.3%)	1754 (30.1%)	1660 (32.8%)	3767 (59.4%)	972 (63.2%)	869 (74.3%)
FRPL: Yes	1516 (38.5%)	2013 (39.3%)	74 (13.3%)	2504 (49.4%)	2641 (41.6%)	911 (59.3%)	808 (69.1%)
FRPL: No	540 (13.7%)	628 (12.2%)	79 (14.2%)	2563 (51.6%)	1242 (19.6%0	6 (0.4%)	1 (0.1%)
ESL: Yes	331 (8.4%)	1 (0.2%)	1 (0.2%)	26 (0.5%)	576 (9.1%)	23 (1.5%)	58 (4.9%)
ESL: No	2160 (54.9%)	152 (27.3%)	152 (27.3%)	2497 (49.3%)	2358 (37.2%)	1020 (66.4%)	920 (78.7%)
Has a Disability: Yes	183 (4.6%)	251 (4.9%)	305 (5.2%)	283 (5.6%)	95 (1.5%)	270 (17.6%)	252 (21.6%)
Has a Disability: No	3754 (95.4%)	4876 (95.1%)	5527 (94.8%)	4784 (94.4%)	6252 (98.5%)	1267 (82.4%)	917 (78.4%)

As with the pre-K to grade 2 IRT, multiple sub contents were measured for each grade. The item pools by grade are available in Table 3.4.

Table 3.4. *Sub Contents Measured for Each Grade in the ISIP Math IRT Study*

Grades 2–5
▪ Number and Operations, Base 10 ▪ Number and Operations, Algebra ▪ Number and Operations, Fractions ▪ Measurement and Data ▪ Probability and Statistics ▪ Personal Financial Literacy Geometry
Grades 6–7
▪ Expressions, Equations and Relationships ▪ Operations and Algebraic Thinking ▪ Ratios and Proportional Relationships ▪ Probability and Statistics ▪ Personal Financial Literacy ▪ Geometry
Grade 8
▪ Expressions, Equations and Relationships ▪ Functions ▪ Proportional Relationships ▪ Number and Operations ▪ Probability and Statistics ▪ Personal Financial Literacy ▪ Geometry

Overall, most items were good quality in terms of item discriminations and item difficulties. For second grade, 44 items were removed and 896 calibrated item parameters remain in the grade 2 item pool. Under third grade, 53 items were removed and 913 calibrated item parameters remain in the grade 3 item pool. For fourth grade, 65 items were removed and 810 calibrated item parameters remain in the item pool. For fifth grade, 71 items were removed and 811 calibrated item parameters remain in the item pool. For sixth grade, 82 items were removed and 977 calibrated item parameters remain in the item pool. For seventh grade, 96 items were removed and 742 calibrated item parameters remain in the item pool. For eighth grade, 73 items were removed and 543 calibrated item parameters remain in the item pool.

CAT Algorithm

The Computerized Adaptive Test (CAT) algorithm is an iterative approach to test taking. Instead of giving a large, general pool of items to all test takers, a CAT algorithm 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. Continue until the stopping criteria is met.

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 to avoid overemphasizing possibly out-of-date information.

Item Selection and Stopping Criteria

For our item selection criteria, we simulated 12 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. In the first edition of ISIP Math, we set a 5-item minimum and 20-item maximum per subtest. Within those bounds, we ended the assessment 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.

This stopping criteria changed with the introduction of math score domains, one of the goals of this norms update. A full description of domain scores and how they are calculated is available in chapter 4. In the current operational edition of ISIP Math, each assessment consists of seven items per math domain, with four domains at each grade. Therefore, each student receives 28 items, which exceeds the former stopping criteria. The overall and domain scores are calculated with a Bayesian prior, like the original ISIP Math.

Chapter 4: Math Domains

A major goal of the renorming of the ISIP Math was to provide information on how students perform in the different math domains. This chapter describes that multifaceted process and the care and attention that went into its construction.

A unique feature of the first edition of ISIP Math was that the items were written around misconceptions that students have about math, and they were also written to better assess the different levels of cognitive engagement as described previously. Items were divided into content strands, otherwise known as domains. While there was coverage of the different content strands and alignment to state standards, we needed to update the items and identify math domains that were comparable across different grades, review the alignment with current standards, and determine which items were aligned with the domains.

Identifying Domains

To determine the domains, the team consulted the National Council of Teachers in Mathematics framework for math content strands and the expectations by grade level. The NCTM framework provides expectations by grade clusters for pre-K–grade 2, grades 3-5, grades 6-8, and grades 9-12. The team identified expectations and skills by grade cluster using this framework.

The team broke down the clusters to correspond to the Istation assessment, and these clusters are pre-K–grade 1, grades 2-5, and grades 6-8. Next, the team reviewed in depth the state standards from several states, including California, Florida, New York, Texas, Virginia, and the Common Core, and the item specification document used in the first edition of the assessment. Using the information from reviewing the standards, previous documentation, and the information from the NCTM, they composed four domains per grade to make the domains as uniform as possible and aligned them with the skills and expectations from the NCTM. Four domains were established for prekindergarten through fifth grade, and a separate set of domains was established for grades six through eight.

Computation and Algebraic Thinking is available for all grades, prekindergarten through 8. This domain involves performing operations and representing algebraic relationships. It includes recognizing and creating patterns, understanding symbols (+, −, ×, ÷), learning and applying computation strategies such as solving for an unknown, recalling basic facts, and working with expressions and equations.

Number Sense is available for prekindergarten through grade 5. This domain refers to foundational math skills, including properties of whole numbers, fractions, and decimals and the relationships between them. This includes representing numbers with visual models, understanding place value, counting, rounding, and comparing.

Number System is available for grades 6 through 8, and it extends the foundations in Number Sense to apply these skills to operations. This includes understanding the properties of positive and negative numbers, rational and irrational numbers, and integers and applying these properties to perform operations.

Measurement and Data Analysis is for prekindergarten through grade 5. It involves determining the size or amount of something. This includes length, weight, volume, area, perimeter, capacity (volume), time, and money. Both the customary and metric systems are utilized. Data Analysis includes sorting and classifying data into categories and using various types of graphs and tables to represent the data. It also involves interpreting and explaining patterns and drawing conclusions to solve problems about the data. Graphs include picture graphs, bar graphs, line/dot plots, tables, and more.

Statistics and Data Analysis for grades 6 through 8 involves answering statistical questions and drawing conclusions based on information about populations. This involves organizing data, measuring it quantitatively, and making inferences based on patterns and distribution. It also includes measures of central tendency including mean, median, mode, and range.

Geometry for prekindergarten through grade 5 involves understanding properties and attributes of shapes, lines, and angles. Students must sort, classify, name, describe, and create various shapes. This domain also includes graphing points on the coordinate plane.

Geometry and Measurement for grades 6 through 8 combines concepts of shape with measurement. This includes understanding and applying formulas for measuring various shapes and angles, comparing attributes of shapes, and using the coordinate plane to analyze relationships and solve problems.

In addition to these domains, Istation also offers Personal Financial Literacy, which is required by the Texas standards. It includes concepts of saving and spending, income, budgets, borrowing and lending, credit and debt, and producers and customers. It is currently available only through teacher resources, and Istation does not offer norms for this domain.

Probability is another domain that requires students to analyze chance events, organize samples, make predictions, and determine solutions to problems. It is only available at certain grade levels, which varies by state and region, and Istation does not offer norms for this domain.

The list of domains and sample skills by domain is available in Table 4.1.

Table 4.1. *Math Domains for ISIP Math Grades Pre-K to 1*

Computation and Algebraic Thinking

- Operations
- Patterns
- Problem solving with numbers
- Algebraic reasoning

Geometry

- Reason with shapes and their attributes
- Compose and decompose shapes

Number Sense

- Counting and cardinality
- Number concepts
- Place value understanding
- Estimations
- Fractional understanding

Measurement and Data Analysis

- Linear measurement
- Capacity, volume, and mass
- Time and temperature
- Measurable attributes
- Money
- Area and perimeter
- Represent and interpret data

Table 4.2. *Math Domains by Grade for ISIP Math Grades 2 to 5*

Computation and Algebraic Thinking

- Problem solving with measurement
- Problem solving/operations with fractions
- Factors and multiples
- Place value understanding
- Patterns/arithmetic patterns
- Operations

Geometry

- Angles
- Reason with shapes and their attributes
- Transformation
- Coordinate system
- Compose/decompose shapes

Number Sense

- Number concepts
- Place value understanding
- Fractional understanding
- Problem solving/operations with fractions
- Compose/decompose shapes
- Problem solving with numbers
- Decimal understanding

Measurement and Data Analysis

- Linear measurement
- Capacity, volume, and mass
- Time and temperature
- Measurable attributes
- Money
- Area and perimeter
- Operations with measurement

Probability (Select locations only)

- Predicting outcomes
- Probability models

Table 4.3. *Math Domains by Grade for ISIP Math Grades 6 to 8*

Computation and Algebraic Thinking

- Equations and inequalities
- Exponents
- Functions
- Proportional relationships
- Systems of equations

Geometry and Measurement

- Measurement conversions
- Area, volume, and surface area on 2D and 3D polygons
- Coordinate plane
- Scale factor
- Angles
- Transformations

Number System

- Unit rates and ratios
- Numerical expressions
- Absolute value
- Rational/irrational numbers
- Scientific notation

Statistics and Data Analysis

- Simple and compound events
- Independent and dependent events
- Probability models

Probability (Select locations only)

- Simple and compound events
- Independent and dependent events
- Probability models

The next step was to organize the items in the item bank by domain. They were listed as the content strands. The team reviewed the items by grade level by content strand and aligned the items with the newly created domains. Team members validated one another's work and also identified items that were not aligned to a domain. Items that were not well aligned to standards or a domain were removed from the item bank.

Stopping Criteria

With the addition of the math domains, the stopping criteria was changed in the assessment. Previously, the algorithm would converge after approximately 16-22 items had been administered. During the 2019-2020 school year, the first administration of the assessment contained 40 items with 10 items from each domain in order to establish a robust Bayesian prior for each domain. Subsequent administrations built on this, and items were administered based on current and prior performance. Items were delivered with an even distribution between the domains, and new scores were computed with the Bayesian prior and the current items. Beginning in the winter, the domains were calculated on at least 5 items per domain, and in the final administration in May, we administered another 40-item assessment with 10 items from each domain. In the current operational edition, we deliver 7 items from each domain, and the assessment is stopped after 28 items.

Relationship of Item Parameters, Domain Scores, and Overall Scores

In the operational version of the ISIP Math, the first time students take ISIP Math in a given academic year, they are given items of median difficulty for each domain. Based on the student's performance on an item, the next item in that domain is either more difficult or less difficult. Seven items are administered for each domain. After all of the test items have been administered, an ability score is calculated for each domain using the 7 items from the particular domain. The overall score calculation uses all 28 items that were administered across the 4 domains. The ability scores for the domains are not part of the calculation for the overall scores.

In subsequent administrations, a Bayesian prior is used for the overall and domain scores. The overall score is comprised of the Bayesian prior and the information from the item difficulty and discrimination factors for all 28 of the items administered. The domain scores are comprised of their Bayesian priors and the item difficulty and discrimination factors for the 7 items administered in that domain. It is important to note that the Bayesian priors for the domain, and the domain scores, are not factored into the overall score. Figure 4.1 displays how the domain scores versus the overall scores are derived for prekindergarten through fifth grade. The domain and overall scores have a similar relationship in grades 6 through 8.

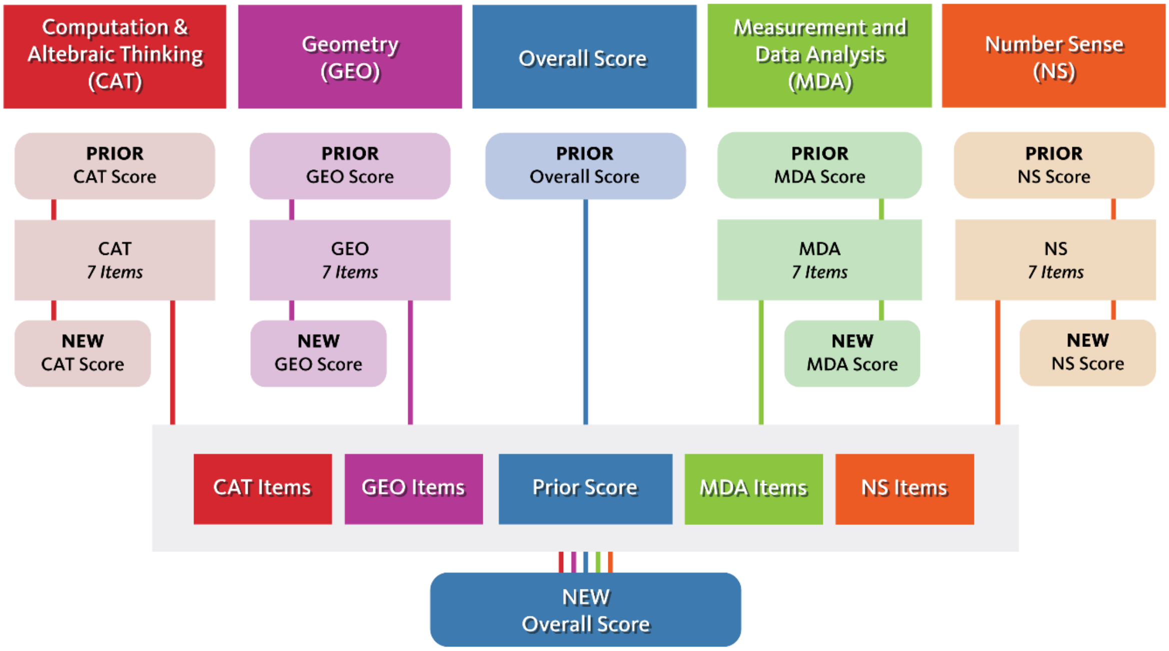


Figure 4.1. Composition of Domain Scores and the Overall Score

As shown in Figure 4.1, domain scores and overall scores are derived separately, and it is important to note that the overall score is the best estimation of a student's math ability. The domain scores are more heavily influenced by what is currently taught in the classroom. Therefore, the overall score can be used for assessing students for their math ability, and the domain scores can be used to provide relative strengths and weaknesses for a student, based on current instruction in the classroom.

Chapter 5: Vertical Scaling

Introduction to Vertical Scaling

Prior to the research described here, ISIP Math test scores were reported using a set of grade-specific scales. These distinct scales were derived by calibrating different sets of items at different grade levels. That is, the items that were appropriate for grade 2 were calibrated together using item response theory (IRT) to form the grade 2 scale, the grade 3 item pool was calibrated for the grade 3 scale, and so forth. This resulted in a set of ISIP Math scales that could be used to described student achievement only *within* their respective grades — not *across* them.[1] In order to allow for cross-grade comparisons, Istation developed an alternative to these separate grade-specific scales, namely, a *vertical scale* (Patz & Yao, 2007; Tong & Kolen, 2010; Carlson, 2011; Young & Tong, 2015)

A vertical scale — also referred to as a developmental scale — is an extended score scale that spans a series of grades and allows for the estimation of student growth along a continuum (Young & Tong, 2015). Having a vertical scale can meet the need for a common interpretive framework for test results across grades and yield important information that informs individual and classroom instruction.

[1] These scales were developed for grades 2 through 8 using data gathered during the 2013-2014 school year, and for grades pre-K through 1 with data from 2014-2015 (Istation, 2018).

For example, vertical scales can be used to:

- monitor student progress as new knowledge or skills are acquired or developed within a content area at different time periods and across the grades;
- examine growth patterns for individual students or groups of students in terms of changes in performance and variability from grade to grade;
- benchmark test items consistent with content standards or curriculum frameworks at different grade levels; and
- match items to a student's ability level for computer-adaptive testing regardless of either the student's grade level or the original grade level of the items (Young & Tong, 2015, p. 450).

The remainder of this chapter describes the processes that were used to create the ISIP Math vertical scale within the framework of item response theory. Typically, these scales are created using the same processes for equating alternate test forms[2] and include item calibrations, the calculation of linking constants, and the application of the linking constants to create the scales (Kolen & Brennan, 2004; Carlson, 2011; Young & Tong, 2015). However, in the case of the ISIP Math tests, certain adjustments needed to be made to both the statistical methods and data collection designs.

These are described first and include the original data collection design and the modifications needed to create a vertical scale for a computer-adaptive testing system. This is followed by challenges in implementing the data collection design due to the COVID-19 pandemic and the mitigation strategies that were used to address issues that arose. The vertical scaling process is described in general terms, and appendices to the chapter provide additional details on the IRT parameter estimation procedure used, the standard error calculations, the parameter estimate adjustments for measurement error, and deriving the linking constants needed to create the final vertical scale. A final section evaluates the vertical scale by applying the newly developed scale to ISIP Math data taken from the January 2020 student administration.

[2] This process is more accurately described as *calibration*, as the test forms involved are designed for different grades with different content and difficulty levels (Kolen & Brennan, 2014).

ISIP Math Vertical Scaling Data Collection Design

Original Data Collection Design

The vertical scaling study took place in March 2020 as a part of the regular monthly test administration in schools using ISIP Math. The data collection used a combination of approaches: *common-item nonequivalent groups design* and *random equivalent groups design* (see Kolen & Brennan, 2004; Kolen, 2007)*.* ISIP Math required this combination due to differences in test formats between grades pre-K through 1 versus grades 2 through 8.

In the common-item nonequivalent groups design, sets of common items (also referred to as *anchor* or *linking sets*) are present on the test forms taken by different groups of examinees. These common items can be embedded throughout the test, placed together in their own section on the test, or placed together on a section that is external to the test. The scores on the common items provide the information needed to create the statistical adjustment that puts the scores from two groups (the first group of examinees on the first form and the second group on the second form) on the same scale.

In the random equivalent groups design, a single group of examinees is randomly assigned to one of two test forms, usually via a spiraling process. The proper implementation of this process leads to two groups of examinees that are randomly equivalent with respect to their ability. The differential performance of the examinee groups on the test that they were assigned is then used to create the statistical adjustment to put the two tests on the same scale.

In developing the data collection for the vertical scale, it was important to take into account that ISIP Math is a *computer-adaptive* rather than a *fixed-forms* testing system. In a computer-adaptive test (CAT) each student is presented with items selected sequentially from a pool of items that have been calibrated to be on the same scale. Each item is selected to match, to the extent possible, that student's currently estimated level of achievement. Students who answer questions incorrectly will have lower estimates of achievement and thus will have less difficult items selected, while students who answer questions correctly will have items more difficult selected.[3] As students vary in their levels of achievement, this results in different students being administered different sets

[3] This a simplified explanation of the CAT algorithm and omits important details such as the use of item information, exposure control, content balancing, stopping rules, and so forth.

of items. The following schematics show how using a CAT can affect the data collection design needed to develop a vertical scale.

The first schematic (Figure 5.1) shows the fixed-forms version of a common-item nonequivalent groups design for linking the tests on two grade levels for a vertical scale. Here, all students at the lower grade take the same lower-grade-level test, and all students at the higher grade take the same higher-grade-level test. A single set of common items from the lower grade is used for linking the two tests across the grades.

However, as shown in Figure 5.2, linking across the grade levels is based on the entire pool of items at each grade level and not any particular test form. Students in the upper grade receive the set of common items from the lower grade along with the different sets of items that the CAT algorithm administers for their grade. The nature of computer-adaptive testing necessitates this linking of item pools rather than test forms. In this configuration, the common items are not used in calculating each student's score on their CAT. Instead, the common items are being used as an *external* link of the lower and upper grades.[4]

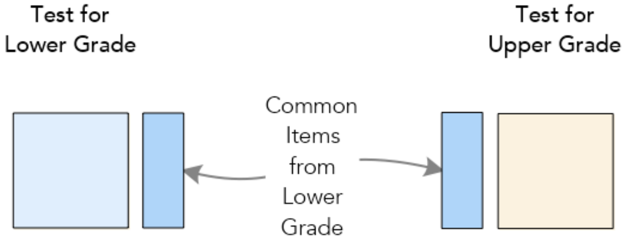


Figure 5.1. *Schematic of Common-Item Nonequivalent Groups Design For Linking Together Fixed-Forms of a Test*

[4] When the set of common items is used in calculating student test scores, it is referred to as an *internal* link (Kolen & Brennan, 2004, p. 19).

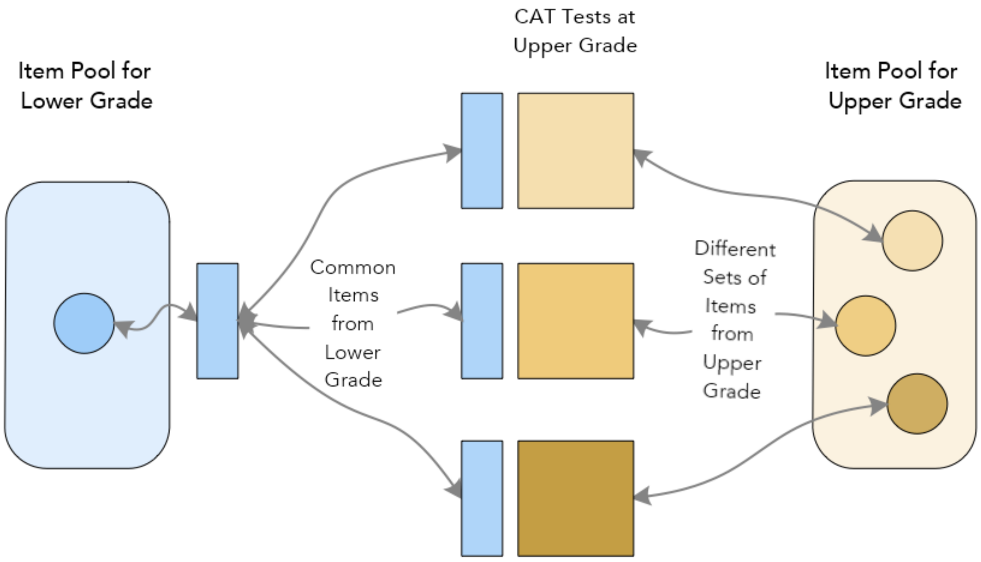

CAT Tests at Upper Grade
Item Pool for Lower Grade
Item Pool for Upper Grade
Common Items from Lower Grade
Different Sets of Items from Upper Grade

Figure 5.2. Detail of Common-Item Nonequivalent Groups Design for Linking Together the Item Pools of a Computer-Adaptive Test

Keeping this in mind, Figure 5.3 below shows a simplified view of the originally planned data collection design for the vertical scaling study. The rows of the figure denote the grades of the students in the study while the columns show the grade level of the items that they were administered. The boxes in the figure represent the groups of students taking the test content that is targeted to their level, and the different colors indicate the differences in test format between the lower grade levels (orange) and upper grade levels (blue).

The common-item nonequivalent groups design was used for most pairs of grade levels by choosing multiple sets of common items from the lower grade level and administering them to students at the higher grade level. For example, sets of items taken from the grade 4 item pool were selected and administered to students in grade 5. Using common items from only the lower grade was done to avoid possible frustration for students in the lower grade if they had been presented with items from a higher grade.

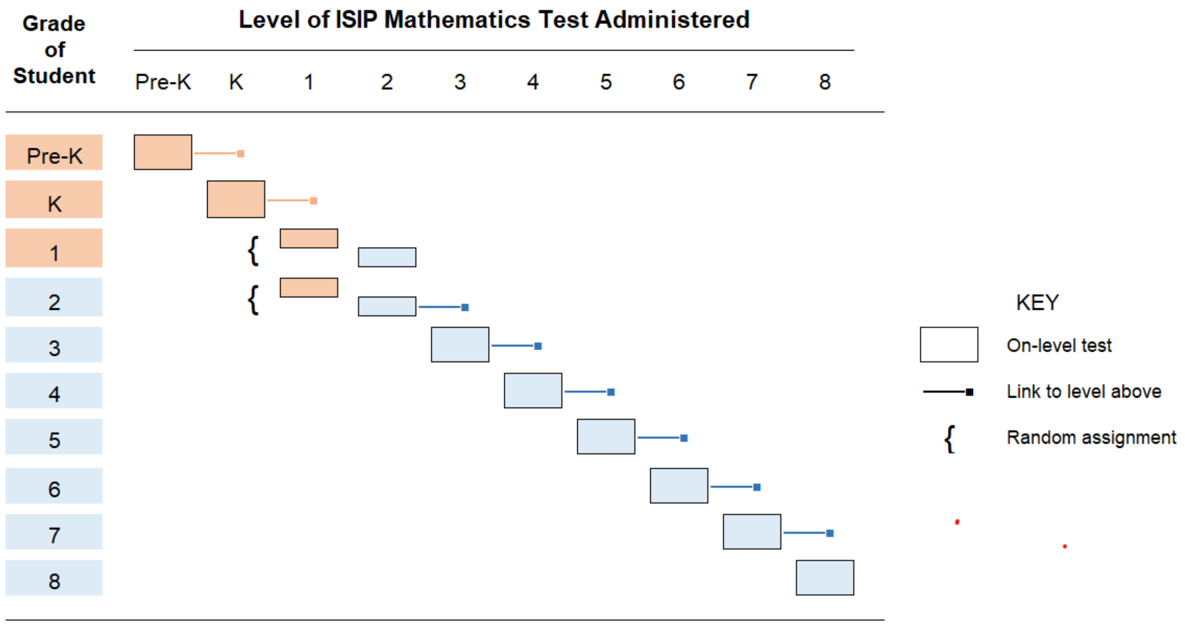


Figure 5.3. A Simplified Schematic of the Original Data Collection Design for ISIP Math Vertical Scaling

Figure 5.3 shows a combination of both the common-item nonequivalent groups design for most of the grade levels, and the random equivalent groups design for linking grades 1 and 2. Rather than having, say, a single set of 20 common items to use as the only link between a pair of grades, it was decided to reduce the amount of additional testing time for students in the study by employing multiple, five-item sets of common items. Each item set was to be taken by separate samples of students of roughly equal numbers, and the item sets were randomly administered via spiraling at the student level (Kolen & Brennan, 2004).

The common-item nonequivalent groups design was implemented at all grade levels *within* each of the grade spans pre-K–1 and 2–8. However, looking *across* grade-level spans, the format of the tests for grades 1 and 2 differed so greatly that it was not considered possible to "bridge the gap" using sets of common items, as the item features would vary too greatly. Here, it was decided to employ the equivalent groups design for the "cross-span" linkage needed for this pair of grades. Specifically, grade 1 students in the study were to be randomly administered either a grade 1 test or a grade 2 test. Similarly, the grade 2 students were also to be randomly assigned either a grade 1 or a grade 2 test. The differences in performance on the two tests at each grade level could then be taken as the difference in their difficulty (Kolen & Brennan, 2004).

The Impact of COVID-19 on the Data Collection

The planned data collection for the vertical scaling study was affected due to COVID-19-related school closures starting in mid-March 2020. The drastic shortening of the original March-through-May testing window for the data collection resulted in fewer total students being sampled than had been originally planned for the study. In addition, the spiraling mechanism for assigning students became unbalanced, resulting in both the numbers of students and their level of achievement varying widely from item set to item set due to the lack of balanced, random assignment. In order to improve the sampling, additional student data responses were collected by targeting items sets with few responses and assigning them to students taking ISIP Math on their devices at home rather than during an in-school administration.

These data collection issues had several consequences. First, the varying sample sizes meant varying degrees of precision in the item parameter estimates needed to create vertical scaling constants for linking grade-level scales. The lack of proper spiraling and the use of at-home test administrations meant that student responses may have been affected due to bias in the samples, such as the non-selection of students without home internet access, distractions during testing, increased/decreased motivation, receiving external assistance on the test, and so forth.

However, the greatest impact to the original data collection design was in the planned use of the equivalent groups design for grades 1 and 2. This part of the data collection had not yet occurred when the school closures began and was deemed too difficult to implement properly in the home setting, as it required each student to take two tests rather than a single test as for the other grades.

The general approach to dealing with these challenges was to make use of various sources of ancillary information in addition to the data that were collected, and to apply analysis approaches that made better use of the data to estimate the parameters needed to create the vertical scale. The specific strategies that were used to mitigate the data collection issues are briefly described in Table 5.1.

Table 5.1. *Mitigation Strategies Used to Deal with Data Collection Issues in the ISIP Math Vertical Scaling Study*

Issue	Mitigation Strategy
Students' testing may have been affected by the COVID-19-related data collection issues such that students are either underperforming or overperforming on their ISIP Math tests.	Identify students whose ISIP Math test scores may have been affected by changes in the data collection processes by screening for outlier scores. Use each student's prior, in-school ISIP Math CAT scores as a covariate and regress their overall CAT scores from the study on them. Examine the residuals from the fitted model to flag and potentially remove as outliers those students whose predicted scores are much less than/greater than their study test scores.
Student achievement of sampled students is unbalanced across item pools.	Adjust the student samples by post-stratifying on student performances defined by ISIP Achievement Levels to achieve more equal levels of student achievement across the item pools. Use the distribution of Achievement Levels across the entire sample as the target for creating the post-stratification weights for the item pools. Use the reweighted item-pool samples for estimating the parameters needed.
Smaller sample sizes lead to less precise parameter estimates and vertical scaling constants.	Only the estimates of the item parameters are needed to create the linking constants for the vertical scale. Use the known estimates of student ability (i.e., thetas, ISIP scale scores) from the students' CAT as ancillary information to better estimate the item parameters. Use extensive screening of the resulting parameter estimates based on their statistics prior to applying a robust procedure to calculate the vertical equating constants.
Vertical scale was created and validated while missing the equivalent groups part of the data collection for grades 1 and 2.	Approximate the vertical equating constant to be used as the link between grades 1 and 2. Use as ancillary information the historic on-grade data for ISIP Math CAT scores and the MetaMetrics Quantile Framework scale scores associated with them. Make use of the pre-existing relationship between the Quantile Framework *vertical* scale with ISIP Math's separate *grade-level* scales to inform the amount of spacing that would be represented by an approximated vertical scaling constant for grades 1 and 2. Create the full vertical scale that is based on the approximated grades 1-2 vertical equating constant and the estimated vertical scaling constants at the other grades. Transform the historic test data onto this new vertical scale and examine the student performance both within and across grade levels, smoothing appropriately.

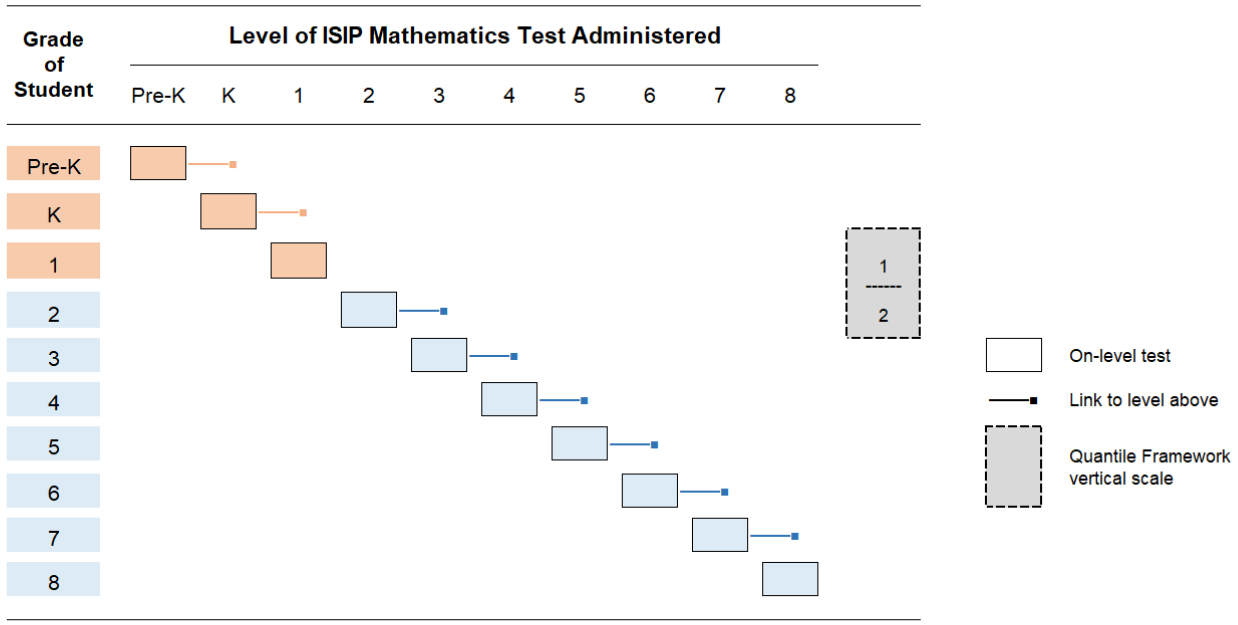


Figure 5.4. Final data collection design for ISIP Math vertical scaling

Figure 5.4 shows change to using only the common-item nonequivalent groups design and the use of the MetaMetrics Quantile Framework vertical scale to inform the linkage between grades 1 and 2. This general approach in using ancillary data was driven by the fact that students in the study were regularly administered ISIP Math tests in their schools. Typically, they had several test sessions worth of data that had been accumulated prior to, and in addition to, the data from the vertical scaling study. Different sources of ancillary information were used to:

- aid in identifying student performance outliers;
- post-stratify the data based on the item-set taken and student performance levels;
- create the covariate used to estimate item parameters; and
- inform the creation and validation of the vertical scale.

Given the issues cited above and the proposed mitigation strategies described in Table 5.1, the final data collection design that resulted is shown in Figure 5.4. The schematic shows the removal of the random equivalent groups part of original data collection design for grades 1 and 2, leaving only the common item links at the other grades. The Quantile Framework vertical scale scores — ancillary data that will be used to approximate a vertical scaling constant at grades 1 and 2 — are shown by the grey box that spans across grades.

Vertical Scaling Study Methods

Data Collected

As described earlier, in the common-item nonequivalent groups design, sets of items from the lower grade's item pool were selected and administered to students at the higher grade. These sets of common items are used to provide the information needed to create the statistical adjustment for putting the item parameters obtained at one grade level on the same scale as the item parameters from a different grade level. Table 5.2 summarizes the numbers of item sets, items, and students used in the study.

Table 5.2. *Summary of the Numbers of Item Sets, Items, and Students Used in the Vertical Scaling Study by Grade Band*

Grade Band	Grades[1]	Number of Item Sets	Number of Items	Number of Students	Number of Students per Item Set
Early Elementary	K–1	40	197	42,208	1,055.2
Late Elementary	3–5	35	175	56,990	1,628.3
Middle School	6–8	18	88	6,788	377.1
Overall	K–8	93	460	105,986	1,139.6

[1] *Grades indicated are those for which data were collected, not the grades of the items used. Thus, approximately 42,000 students in kindergarten were administered 197 items taken from prekindergarten in 40 different sets.*

The number of students in the study at each grade band are in line with the number of schools that currently administer the ISIP Math tests. That is, most schools use the tests at the early and late elementary levels with far fewer schools administering the tests to middle school students.

Preparation for Data Analysis

The data collected in the study were broken up into grade-specific files. Each file included fields showing the:

- test administration date and unique student identifier;
- student's ISIP Math CAT scale score, scale score standard error, normative percentile rank of the student scale score, and the MetaMetrics Quantile scale score (QSS);
- grade of the student taking the item and the grade of the item taken;
- specific item-set ID and item IDs administered to a student, the discrimination and difficulty parameters of each of those items, the student's specific response;
- student's Math CAT scale score and standard error from the test administration immediately prior to their study administration; and
- studentized residuals from regressing each student's Math CAT scale score from the study on their previous scale score.

Once the files were uploaded, the data preparation steps included:

- checking all variables for out-of-bound and missing values;
- verifying item-set ID and item ID information against test specification documents;
- independently verifying the item parameters in the file against the original item-bank source files;
- removing any duplicate student cases or cases with completely missing item response strings;
- standardizing the ISIP scale scores and standard errors to have a mean of zero and a standard deviation of one; and
- creating achievement levels based on scale score quintiles.

In order to identify students whose scores may have been affected by changes in the data collection processes, we checked their ISIP Math CAT scores for outliers. At each grade, the students' prior, in-school ISIP Math CAT score was used as a covariate, and their overall CAT scores from the study were regressed on them. The studentized residuals from the fitted models were used to flag and remove as outliers students whose residuals were less than −2.00 or greater than 2.00.

Finally, the data were post-stratified by item-set ID and achievement level quintile in order to have a better balance of student achievement across the item sets within a grade. The overall distribution of achievement level quintiles for the grade was used as the target set of percentages for the post-stratification procedure.

Item Calibration Procedures

The next step in the process after the data were prepared was to analyze the students' responses using *item response theory* (IRT). IRT is a measurement model that relates the probability of success in answering a question correctly to two components: the ability of the student and the characteristics of the item.

The *two-parameter logistic* (2PL) *model* was originally used to calibrate the ISIP Math items (Istation, 2018). In this model, the probability of a correct response (i.e., 1) is given by

$$p = P(X = 1|\theta) = \frac{\exp[a(\theta - b)]}{1 + \exp[a(\theta - b)]} = \frac{1}{1 + \exp[-a(\theta - b)]}$$

where *a* is the *item discrimination* parameter, *b* is the *item difficulty* parameter, and θ is the *person ability* parameter. In this model, all three parameters need to be estimated from the data.

To make better use of the data that had been collected in the current study, an alternative parameterization of the 2PL IRT model was used wherein

$$\text{logit}(p) = \ln\,(p/(1-p)) = A\theta + B$$

and the item discrimination and difficulty parameters are changed to a *slope* parameter *A* and an *intercept* parameter *B,* and θ is the *person ability* parameter as before. By using this *slope-intercept parameterization,* we were able to take advantage of ancillary data that had been collected during the study, namely, the ability estimates of the students that had been derived separately from their CAT. Because what interested us were the estimates of the item parameters and not the person ability parameters, we estimated the model

$$\text{logit}(p) = A\hat{\theta} + B$$

using the standard technique of *logistic regression* (LR) (Hosmer, Jr. et al., 2013; DeMaris, 1992; Gelman & Hill, 2007). We then transformed the estimated slope- and intercept-parameters (*A* and *B*) back into the IRT discrimination and difficulty parameters (*a* and *b*) needed to create the vertical scale. Additional details regarding the 2PL IRT model and logistic regression are presented in Appendix 5.1.

The transformations of the slope parameter *A* and the intercept parameter *B* back to the IRT parameterization were straightforward. However, the transformations of the standard errors of *A* and *B*, σ_A and σ_B, from the LR scale to the IRT scale were more complicated and were approximated using the *delta method* (Oehlert, 1992; Kolen & Brennan, 2004). These details can be found in Appendix 5.2.

An additional step was needed in the parameter estimation process as $\hat{\theta}$, the CAT estimate of the student's unknown ability θ, was used as a covariate rather than estimated along with the item parameters. As the effects of measurement error in covariates can bias parameter estimates in statistical models (Carroll et al., 2006, p. 1), it was important to take this into account for the parameters that would be used in the operational ISIP Math computer-adaptive tests going forward. The *simulation-extrapolation method* (SIMEX) (Cook & Stefanski, 1993; Hardin et al., 2003; Lederer & Küchenhoff, 2006; Shaw & Keogh, 2017) was used to take the item parameters that had been estimated using the logistic regression and adjust them to account for the measurement error inherent in the students' CAT ability estimates. The SIMEX method is discussed and the algorithm used to implement it is provided in Appendix 5.3.

Post-Calibration Item Checks

Three sets of checks were used to determine which items would be used to develop the equating constants needed to create the ISIP Math vertical scale. The checks were applied sequentially.

The first set of checks examined the statistical quality of the item parameters that had been estimated during the calibration process. For each grade in the study for which data had been collected, items were flagged for further inspection and possible elimination according to the following criteria:

- Discrimination parameters (*a*-parameters) were flagged as "low" when less or equal to 0.20 and "high" when greater than or equal to 2.00.
- Difficulty parameters (*b*-parameters) were flagged as "low" when less than or equal to -3.00 and "high" when greater than or equal to 3.00.
- Item fit statistics were calculated using Bock's (1972) chi-squared procedure with eight subgroups, and with the expected proportions based on the item parameter estimates the from the 2PL model, and the median of the ability estimates within

a subgroup (Stone & Zhang, 2003, p. 332). Items were flagged at both the 0.05 and the 0.01 significance levels.

The second check examined the cross-grade difficulty values of the common items. In the context of vertical scaling, one would anticipate that the difficulty of items taken from a lower grade should be the same or lower when administered to students at a higher grade, since the students at the higher grade are further along in their education and are presumably already familiar with the content. Apart from sampling and measurement error, the expectation would be that on average, the item parameter difficulties at the higher grade should be less than or equal to those at the lower grade. Items not exhibiting this kind of behavior would violate this expectation and not necessarily be appropriate for use in creating a vertical scale.

The final set of checks used the *robust z procedure* (Huynh & Meyer, 2010) to assess which of the common items were reasonably "stable" across each of the pairs of grade levels being linked together. In this procedure, the differences between item parameter estimates are compared to the overall median of the differences and then standardized by the interquartile range to create a robust version of the usual *z*-statistic. That is,

$$z_{Robust} = {(D_j - Md)}\big/{(0.74 \cdot IQR)}$$

where D_j is the difference of the j^{th} pair of parameters being examined, *Md* is the median across all of the pairs of parameter differences, and the denominator equals the interquartile range multiplied by a constant to scale it to be approximately equal to the standard deviation of a normal distribution. When the D_j are normally distributed, the z_{Robust} statistic is asymptotically normal with a mean of zero and a standard deviation of one (Huynh & Meyer, 2010, p. 2). Item pair differences with an absolute value of robust z greater than 1.96 were flagged for further review and possible administration.

Table 5.3 presents the results of applying these checks both overall and by grade band. The result of sequentially applying these item quality and stability checks was to reduce the initial pool of 460 items across the grade levels to a final pool of 200 items that were used to create the vertical scale.

Table 5.3. *Initial and Final Numbers of Items Used to Create Equating Constants After Item and Robust Z Flagging by Grade Band*

Grade Band	Grades	Initial	After Item Flagging	After Robust Z Flagging
Early Elementary	K–1	197	83	76
Late Elementary	3–5	175	65	55
Middle School	6–8	88	86	69
Overall	K–8	460	234	200

Determining the Linking Constants

The final sets of item parameters coming out of the previous step were used to develop the *linking constants* (i.e., statistical adjustments) needed to create the ISIP Math vertical scale. The schematic in Figure 5.4 shows the two kinds of linkages that needed to be considered in order to do this.

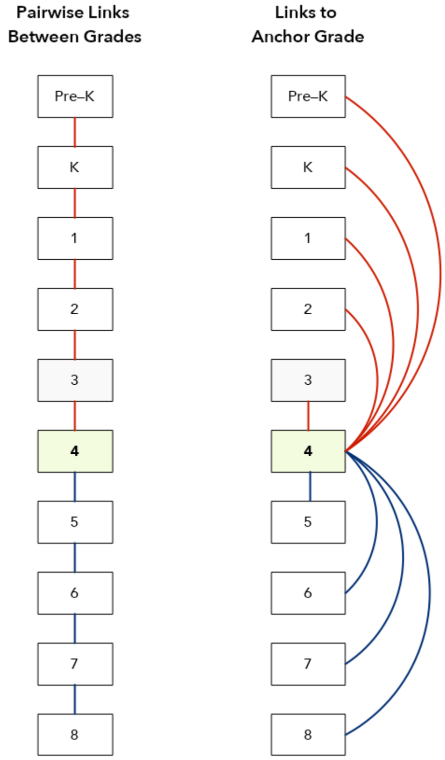


Figure 5.4. *Schematic of Pairwise and Anchor-Grade Linkages of Item Pools in the ISIP Math Vertical Scaling*

Each of the boxes in the figure represents an item pool at one of the grade levels of the ISIP Math tests with the lines depicting the linkage from one grade level pool to another. The left part of the schematic depicts *pairwise linkages* where each grade's item pool is linked to either the item pool of the grade immediately above it or below it. In this figure, the pre-K item pool is linked to the kindergarten item pool, the kindergarten item pool is linked to the grade 1 item pool, and so forth, going up the grades. Similarly, we can go down the grade levels by starting with the grade 8 item pool and linking it to the grade 7 item pool, etc.

As a result of the data collection design and the calibrations that were performed using the ISIP Math data, each pair of grades had a set of common items with *two* sets of item parameters, one for each grade. For example, consider the set of common items for grades 4 and 5. Since the common items were taken from the grade 4 item pool, each item had a set of pre-existing discrimination and difficulty parameters that were on the grade 4 scale. These same common items were administered to students at grade 5 and calibrated to have a different set of item discrimination and difficulty parameters.

Now, in item response theory, when the same set of items has been calibrated with two different groups, the resulting scales are linearly related, and so a linear function can be used to transform one scale into the other (Kolen & Brennan, 2004). That is, we can find linking constants — in this case a slope and an intercept — that can be used to transform the measures of student ability and the item parameters from one scale into another.

In our example, this means we can find a slope and an intercept such that

$$Grade\ 4\ Scale\ Values = slope \cdot (Grade\ 5\ Scale\ Values) + intercept$$

that is, transforming the grade 5 scores so that they are on the same scale as the grade 4 scores. The same kind of analysis can be performed to create linking constants for all the other pairs of grades shown in the left of Figure 5.4 and extended to create linking constants that can be used to place the scores for any grade's scale on the scale of a single chosen grade, as shown in the right of the figure.

Here, each specific grade-level scale has been linked to a single *anchor* (or target) *grade*, namely grade 4. To go to the anchor grade from a non-adjacent grade simply requires the composition of the functions for all of the pairwise links between them. For instance, to go from grade 6 to grade 4 first requires us to transform the grade 6 scores to be on the grade 5 scale, then to transform these scores to the grade 4 scale.

By using this approach of creating a "chain" from multiple links, we can take all of the pairwise sets of linking constants and use them to produce the final set of vertical linking constants that were desired. The details of this process are presented in Appendix 5.4.

Creating the Final Vertical Scale

For the most part, the linking constants were calculated for the pairwise grade level scales described above. They were then used to create the constants linking each grade's scale to the scale of the anchor grade by compositing the linear transformations of the pairwise links between each starting grade and the anchor grade.

Grade 4 was chosen to be the anchor grade as it was roughly in the middle of the grades and would thus reduce the total number of links needed to go from any other grade to it. For example, if grade 8 had been chosen as the anchor grade, then there would have been nine links needed to go from pre-K to grade 8. Using grade 4 as the anchor grade ensured that the number of links needed would never be greater than five. The final set of linking constants is shown below in Table 5.4.

Table 5.4. *Linking Constants (Slopes and Intercepts) From Starting to Anchor Grade*

Starting Grade	Anchor Grade	Slope	Intercept
Pre-K	4	1.3534	-6.3457
K	4	1.3550	-4.1587
1	4	1.1440	-2.2464
2	4	1.0119	-1.0255
3	4	0.9401	-0.3285
4	4	1.0000	0.0000
5	4	0.9552	0.2489
6	4	1.1370	0.5084
7	4	1.6292	0.5850
8	4	1.4148	0.6809

The only exception to this process occurred in creating the pairwise link between the grade 1 and grade 2 scales. Here, the median values of the slopes and intercepts from the K-to-grade 1 and the grade 2-to-3 linear transformations were used to define the grade 1-to-2 transformation. This choice of linking constants was found to produce a smooth progression across the pre-K to grade 4 span when the final linking constants in Table 5.4 were calculated.

In order to create the final vertical scale, each of the original grade-specific ISIP Math scales with a mean of 2000 and a standard deviation of 200 ($SS_{Original}$) were transformed to be on their original logit scales ($Logit_{Original}$) with a mean of zero and a standard deviation of one:

$$Logit_{Original} = (SS_{Original} - 2000)/200.$$

Next, the logit scores for each grade were transformed to a vertically scaled logit score ($Logit_{VerticalScale}$) by applying the grade-specific slope and intercept from Table 5.4:

$$Logit_{VerticalScale} = Slope_{Grade\ K} \cdot Logit_{Original} + Intercept_{Grade\ K}$$

Finally, the vertically scaled logit scores were linearly transformed to a new ISIP Math reporting vertical scale (SS_{New}):

$$SS_{New} = Slope_{SS\ New} \cdot Logit_{VerticalScale} + Intercept_{SS\ New}$$

There were two main considerations that guided the development of the final reporting scale. First, the new ISIP Math scale needed to have a range of values that was distinct from the ranges of the original scales. This was necessary to emphasize the fact that ISIP Math was now on a single vertical scale that was different from the individual grade-specific scales and to avoid confusion between the two scaling systems going forward.

The second consideration had to do with clearly specifying what the *lowest obtainable scale score* (LOSS) and the *highest obtainable scale score* (HOSS) would be on the new scale. The concern here was having an underlying vertical scale that would support the full range of student achievement on the current tests while allowing for extended levels of student achievement if high school levels were added to the ISIP Math tests in the future.

The slope and intercept of this transformation were chosen so that the vertically scaled logit scores of -9.00 and 8.00 corresponded to vertically scaled reporting scores of 100 and 900 respectively. This resulted in linear transformation coefficients of a slope of 47.058824 and an intercept of 523.529408.

Evaluation of the Vertical Scale

Kolen and Brennan (2004) provide three attributes of scales that have been used to evaluate the results of a vertical scale:

- the average grade-to-grade growth;
- grade-to-grade variability; and
- the separation of grade distributions.

These attributes were examined using ISIP Math student data taken from the January 2020 test administration. The student scores from this administration were transformed from their original grade-specific scales to the new ISIP Math reporting vertical scale. The results of applying these transformations are shown in Figure 5.5 and Table 5.5.

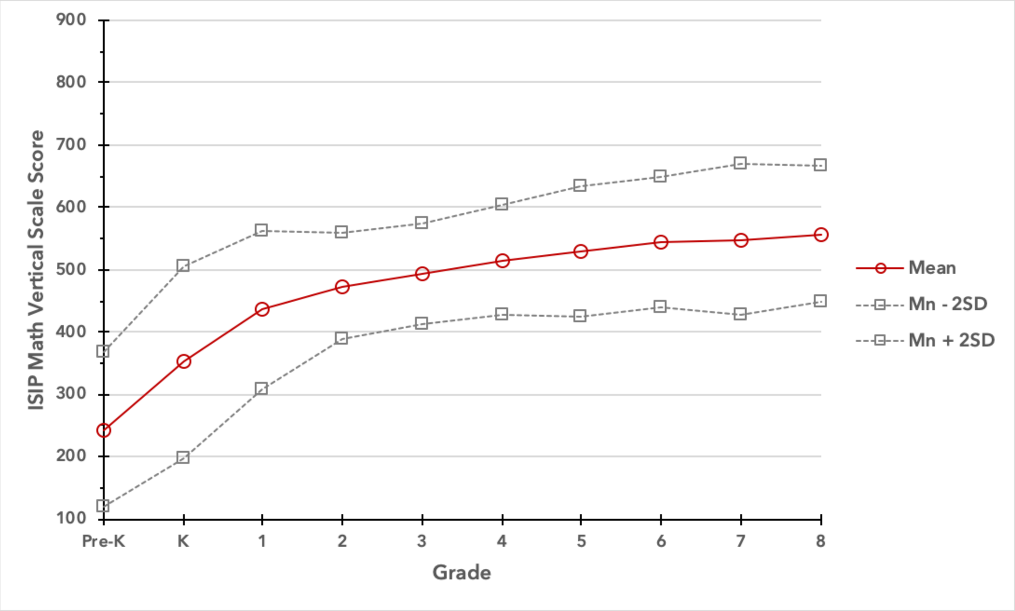


Figure 5.5. ISIP Math final reporting growth curves: Scale score means and confidence bands by grade level (Source: ISIP Math January 2020 test administration data)

Table 5.5. *ISIP Math Vertical scale grade-to-grade growth, variability, and distribution separation (Source: ISIP Math January 2020 test administration data)*

Lower Grade	Upper Grade	Difference	Pooled SD	Effect Size
Pre-K	K	109.4	74.7	1.47
K	1	83.9	70.3	1.19
1	2	37.0	54.1	0.68
2	3	20.6	41.4	0.50
3	4	21.5	42.2	0.51
4	5	13.4	48.2	0.28
5	6	15.0	51.9	0.29
6	7	4.4	55.1	0.08
7	8	9.0	58.2	0.16

The figure and the table clearly show that the grade-to-grade growth is curvilinear with the greatest growth occurring from pre-K through grade 2, lessening from grade 2 through grade 4, and leveling off above grade 4. This pattern is similar to other vertically scaled achievement tests such as the *Stanford Achievement Test* series (10th Edition) (Young & Tong, 2015).

Interestingly, the grade-to-grade variability decreases through the end of grade 4 but then begins to increase again. This may be indicative of the changes in the spread of student achievement during the transition from the elementary curriculum to that of the middle school curriculum.

Finally, the effect-size measures decrease from nearly 1.5 standard deviations at the lowest grades to around one-tenth of a standard deviation at the highest grades. This indicates a clear separation of the student achievement distributions at the lower elementary grades with the distributions becoming more and more overlapped as the students move into middle school and the curriculum changes.

Appendix 5.1: Item Calibration Using Ancillary Information

The measurement model originally used to calibrate the dichotomously scored[5] multiple-choice items of ISIP Math tests was the *two-parameter logistic* (2PL) *item response model* (IRT). In this model, the probability of a correct response (i.e., 1) is given by

$$p = P(X = 1|\theta) = \frac{\exp[a(\theta - b)]}{1 + \exp[a(\theta - b)]} = \frac{1}{1 + \exp[-a(\theta - b)]}$$

where a is the *item discrimination* parameter[6], b is the *item difficulty* parameter, and θ is the *person ability* parameter. Similarly, the probability of an incorrect response (i.e., 0) is given by

$$1 - p = P(X = 0|\theta) = \frac{1}{1 + \exp[a(\theta - b)]} = \frac{\exp[-a(\theta - b)]}{1 + \exp[-a(\theta - b)]}.$$

The parameterization shown above using a and b is the one most seen in introductions to the 2PL model and can be thought of as the *IRT parameterization*.

An alternate parameterization can be derived by re-writing the exponent in the equations above as

$$a(\theta - b) = a\theta - ab = a\theta + (-ab).$$

Then, setting $A = a$ and $B = -ab$, we see that

$$a(\theta - b) = A\theta + B.$$

This change recasts the item discrimination and difficulty parameters as a *slope* parameter A and an *intercept* parameter B respectively. If we use this *slope-intercept parameterization,* then

$$p = P(X = 1|\theta) = \frac{\exp[A(\theta + b)]}{1 + \exp[A(\theta + b)]} = \frac{1}{1 + \exp[-A(\theta + b)]}$$

and

[5] Dichotomous items such as multiple-choice items are usually scored as 1 for a correct answer and 0 for an incorrect answer.

[6] For notational convenience, we will assume that the item discrimination parameter already includes the usual scaling factor of $D = 1.7$ to make the logistic item response curve similar to that of the normal ogive curve.

$$1 - p = P(X = 0|\theta) = \frac{1}{1 + \exp[A(\theta + B)]} = \frac{\exp[-A(\theta + B)]}{1 + \exp[-A(\theta + B)]}.$$

Using this parameterization and looking at the odds-ratio of the probably of a correct response vs. an incorrect response, we get (after some algebraic simplification):

$$p/(1 - p) = \exp(A\theta + B).$$

Finally, taking the natural log of both sides of this equation then yields an alternative formulation of the 2PL IRT model as

$$\text{logit}(p) = \ln\,(p/(1 - p)) = A\theta + B.$$

In this equation, the *logit* or *log-odds* of obtaining a correct response for an item is a linear function of a student's ability θ with slope and intercept parameters A and B respectively (Baker & Kim, 2004; Engec, 1998).

It is very important to note in either one of these formulations that both the item parameters – A and B for the logit version or a and b for the IRT version – and the person parameter θ, are unknown and need to be estimated. However, if we were to use the person parameter for each student that is estimated by their ISIP Math CAT as a covariate, then we could model the logits as

$$\text{logit}(p) = A\hat{\theta} + B$$

using the standard technique of *logistic regression* (LR) (Hosmer, Jr. et al., 2013; DeMaris, 1992; Gelman & Hill, 2007). This would provide us with the opportunity of estimating the A and B parameters, and consequently, transforming them into the IRT a and b parameters we need to estimate for creating our vertical scaling constants.

Appendix 5.2: Adjusting for Measurement Error

As stated elsewhere in this technical report, the logistic regression model being fitted to estimate the item parameters is of the form

$$\text{logit}(p) = A\hat{\theta} + B$$

where $\hat{\theta}$ is the CAT estimate of a student's unknown, ability θ.

The effects of measurement error in covariates include bias in parameter estimation for statistical models (Carroll et al., 2006, p. 1). In the case of using logistic regression to estimate IRT item parameters, this is of major concern as these estimates will be used in the operational ISIP Math computer-adaptive tests going forward.

The amount of measurement error in the ISIP Math computer-adaptive tests can be estimated looking at the error associated with $\hat{\theta}$, namely its standard error $SE(\hat{\theta})$. This information can be used with the *simulation-extrapolation method* (SIMEX) (Cook & Stefanski, 1993; Hardin et al., 2003; Lederer & Küchenhoff, 2006; Shaw & Keogh, 2017) to provide item parameter estimates that have been adjusted to account for measurement error. In this approach, "error of increasing amounts is added artificially to data, and a relationship between the size of the error and regression coefficients is estimated" (Shaw & Keogh, 2017, p. 3). The regression coefficients are then extrapolated for the case when there is zero measurement error.

SIMEX has been used in the educational measurement context by Shang and his colleagues (Shang, 2012; Shang et al., 2015) to correct for measurement error in covariates for when estimating quantile regressions and student growth percentiles (SGP). They found that SIMEX was effective in reducing bias for both individual and aggregate student growth percentiles.

The following description of the SIMEX algorithm is based on those of Hardin et al (2003) and Shaw and Keogh (2017), and has been particularized for ISIP Math analyses. The R package `simex` (Lenhard & Seibold, 2019) was used for the analyses.

The setup:

- We start by assuming additive error in our estimate of student ability, the covariate $\hat{\theta} = \theta + u$, where $u \sim Normal(0, \sigma^2)$.
- We use a regression model for an outcome Ythat includes $\theta.$ Since we are using logistic regression, model is of the form $\text{logit}(p) = A\theta + B + \epsilon$. However, due to the measurement error, we are observing $\hat{\theta}$ rather than θ.
- To obtain an estimate of the measurement error associated with $\hat{\theta}$, we used the mean of the error variances of the estimated abilities across persons and items within an item set p, $Var(\hat{\theta}_p) = [SE(\hat{\theta}_p)]^2$.

The simulation step of the SIMEX procedure:

- For each item in the *p*-th item set, we run Kbootstrap iterations, based on our estimate of the error variance $Var(\hat{\theta}_p)$, and on the different values of a scaling factor $\lambda > 0$, where $\lambda \in \{0.5, 1.0, 1.5, 2.0\}$. The changes in the scaling factor are used to model the effect of changes in the covariate on the amount of measurement error.
- For the *k*-th bootstrap, generate $\theta_{k,i} = \hat{\theta}_i + \sqrt{\lambda}\, SE(\hat{\theta}_p)\, Z_{k,i}$ for each value of λ where $Z_{k,i} \sim Normal(0,1)$ and $\hat{\theta}_i$ is the ability estimate for the *i*-th student on an item. This step just adds in the measurement error.
- Since $\hat{\theta}_p$ and $Z_{k,i}$ are independent, the total error variance in $\theta_{k,i}$ will be

$$Var(\theta_{k,i}) = Var(\hat{\theta}_i + \sqrt{\lambda}\, SE(\hat{\theta}_p)\, Z_{k,i}) = Var(\hat{\theta}_i) + Var(\sqrt{\lambda}\, SE(\hat{\theta}_p)\, Z_{k,i})$$

$$= Var(\hat{\theta}_p) + \left[\sqrt{\lambda}\, SE(\hat{\theta}_p)\right]^2 Var(\, Z_{k,i}) \;= Var(\hat{\theta}_p) + \lambda\, Var(\hat{\theta}_p)$$

$$= (1 + \lambda)\, Var(\hat{\theta}_p).$$

- Next, we fit the logistic regression model of interest with $\theta_{k,i}$ in place of $\hat{\theta}_i$. That is, we fit $\text{logit}(p) = A_{k,\lambda}\theta_{b,i} + B_{k,\lambda}$.
- Obtain overall estimated parameters $\hat{A}_\lambda$ and $\hat{B}_\lambda$ for each λ as the mean of the K bootstrap estimates of $A_{b,\lambda}\ and\ B_{b,\lambda}$.

The extrapolation step of the SIMEX procedure:

- Now, we fit curves to the pairs of generated parameters $(\lambda, \hat{A}_\lambda)$ and $(\lambda, \hat{B}_\lambda)$. That is, we will fit two functions, $\hat{A}_\lambda = f_A(\lambda)$ and $\hat{B}_\lambda = f_B(\lambda)$, one for each of the parameters. These functions will be used to describe the change in parameter estimates as a function of the amount of measurement error.
- Although a variety of functions could be used to model these relationships, a quadratic function such as $f(\lambda) = \delta_0 + \delta_1\lambda + \delta_2\lambda^2$ has been shown to work well in practice (Lederer & Küchenhoff, 2006; Shaw & Keogh, 2017).
- Finally, the SIMEX estimators for the logistic regression parameters are the fitted values $\hat{A}_\lambda = f_A(-1)$ and $\hat{B}_\lambda = f_B(-1)$. This is because at the value $\lambda = -1$, the error variance goes to zero since $Var(\theta_{k,i}) = (1 + (-1)) \cdot Var(\hat{\theta}_p) = 0 \cdot Var(\hat{\theta}_p) = 0$.

The jackknife procedure developed by Stefanski and Cook (1995) was used to estimate the variances of the SIMEX-adjusted item parameters. Finally, as these estimates were in terms of the slope-intercept parameterization used by the logistic regression model, they were transformed back to the IRT parameterization described in Appendix 5.1.

Appendix 5.3: Transforming Logistic Regression Standard Errors to IRT Standard Errors Using the Delta Method

Using logistic regression allowed for the ability of each student from their computer-adaptive test to be employed as a covariate in estimating item response parameters, and the student ability is subsequently adjusted for measurement error using the SIMEX procedure. However, the LR-estimated parameters and their standard errors needed to be transformed back to the IRT parameterization that Istation needed for the operational administration of their testing system.

The transformations of the slope parameter A and the intercept parameter B back to the IRT parameterization is straightforward and given by $a = A$ and $b = -B/A$. However, the transformations of the standard errors of A and B, σ_A and σ_B, from the LR scale to the IRT scale are more complicated.

Fortunately, they can be well approximated using the *delta method* (Oehlert, 1992; Kolen & Brennan, 2004). This method approximates the transformations of random variables using a first-order Taylor approximation and then taking its variance.

If G is a transformation function where $\mathbf{X}$ is a p-dimensional random vector with a mean $\mathbf{m}$ and $G(\,)$ is differentiable**,** then the transformation of $\mathbf{X}$ to a q-dimensional random vector $\mathbf{Y}$ is

$$\mathbf{Y} = G(\mathbf{X}) \approx G(\mathbf{m}) + G'(\mathbf{m}) \cdot (\mathbf{X} - \mathbf{m}).$$

Calculating the expected value and the variance of Y based on this approximation yields

$$E(\mathbf{Y}) \approx G\big(E(\mathbf{X})\big) = G(\mathbf{m})$$

and

$$Var(\mathbf{Y}) \approx G^{\prime}\ \big(\mathrm{E}(\mathbf{X})\big)\ Var(\mathbf{X})\ \ G^{\prime}\ \big(\mathrm{E}(\mathbf{X})\big)^{\mathrm{t}} = G^{\prime}\ (\mathbf{m})\ \ Var(\mathbf{X})\ \ G^{\prime}\ (\mathbf{m})^{\mathrm{t}}$$

where G'denotes $q\ \times p$-matrix of partial derivatives, $Var(\,)$ is the variance-covariance matrix, and the subscript "t" denotes the transpose of the matrix.

In our case, our vector $\mathbf{X}$ is composed of the parameters of the logistic regression, and as they are fixed values, the expected value of $\mathbf{X}$ is

$$E(\mathbf{X}) \;=\; \mathbf{m} \;=\binom{A}{B}.$$

The transformation function takes us from the LR parameters A and B to the IRT parameters a and b, and is given by

$$E(\mathbf{Y}) = \binom{a}{b} \approx G\big(E(\mathbf{X})\big) = G(\mathbf{m}) = G\binom{A}{B} = \binom{A}{-B/A}$$

as was described above. The variance-covariance matrix $Var(\mathbf{X})$ is simply that of the logistic regression and is denoted by

$$Var(\mathbf{X}) = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix}.$$

Finally, the matrix of partial derivatives is given by

$$G'(\mathbf{m}) = \begin{pmatrix} \frac{\partial G(A)}{\partial A} & \frac{\partial G(A)}{\partial B} \\ \frac{\partial G(B)}{\partial A} & \frac{\partial G(B)}{\partial B} \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial A}(A) & \frac{\partial}{\partial B}(A) \\ \frac{\partial}{\partial A}(-B/A) & \frac{\partial}{\partial B}(-B/A) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ B/A^2 & -1/A \end{pmatrix}.$$

Therefore, the variance-covariance matrix associated with the transformation of logistic regression parameters to the IRT scale, is

$$Var(\mathbf{Y}) = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix} \approx G^{\prime}\,(\mathbf{m})\;\; Var(\mathbf{X})\;\; G^{\prime}\,(\mathbf{m})^{\mathrm{t}}$$

$$= \begin{pmatrix} 1 & 0 \\ B/A^2 & -1/A \end{pmatrix} \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix} \begin{pmatrix} 1 & B/A^2 \\ 0 & -1/A \end{pmatrix}$$

with the standard errors of the discrimination parameter a and difficulty parameter b being given by the diagonal elements $SE(\mathbf{Y}) = \sqrt{Var(\mathbf{Y})}$.

For example, if the parameters estimated by a logistic regression are $A = 0.7784$ and $B = 0.5779$, then

$$E(\mathbf{Y}) = \binom{a}{b} \approx G\big(E(\mathbf{X})\big) = G(\mathbf{m}) = G\binom{0.7784}{0.5779} = \binom{0.7784}{-0.7784/0.5779}$$

and thus $a = 0.7784$ and $b = -0.7424$. If the variance-covariance matrix for these parameters is

$$Var(\mathbf{X}) = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix} = \begin{pmatrix} 0.0077 & 0.0023 \\ 0.0023 & 0.0063 \end{pmatrix}$$

then, the matrix of partial derivatives is

$$G'(\mathbf{m}) = \begin{pmatrix} 1 & 0 \\ B/A^2 & -1/A \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0.9538 & -1.2847 \end{pmatrix}$$

and the variance-covariance matrix for the IRT parameterization is

$$Var(\mathbf{Y}) = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix} \approx G'(\mathbf{m}) \; Var(\mathbf{X}) \; G'(\mathbf{m})^{\mathrm{t}}$$

$$= \begin{pmatrix} 1 & 0 \\ 0.9538 & -1.2847 \end{pmatrix} \begin{pmatrix} 0.0077 & 0.0023 \\ 0.0023 & 0.0063 \end{pmatrix} \begin{pmatrix} 1 & 0.9538 \\ 0 & -1.2847 \end{pmatrix}$$

$$= \begin{pmatrix} 0.0077 & 0.0044 \\ 0.0044 & 0.0118 \end{pmatrix}.$$

Finally, the standard errors for a and b are given by $\sigma_a = \sqrt{0.0077} = 0.0877$ and

$\sigma_b = \sqrt{0.0118} = 0.1086.$

Appendix 5.4: Deriving linking constants

Pairwise linking constants

When the same set of items has been calibrated using item response theory with two different groups of examinees, the resulting scales are linearly related (Kolen & Brennan, 2004). That is, we can find *linking constants* (i.e., a slope and an intercept) that can transform the measures of student ability and the item parameters from one scale into another. If we let S denote the grade of scale that one is starting with and T as the target grade of the transformation, then transformation from the grade S scale to the grade T scale is given by

$$\theta_{Ti} = \gamma_{ST}\theta_{Si} + \delta_{ST}$$

where for an individual i, θ_{Si} is the grade S ability scale, θ_{Ti} is the grade T ability scale, and γ_{ST} and δ_{ST} are the slope and intercept of the linear function. The transformations of parameters from the grade S scale to the grade T scale for the j-th item, $j = 1, \cdots, J$, are given by

$$a_{Tj} = {}^{a_{Sj}}\!/_{\gamma_{ST}}$$

and

$$b_{Ti} = \gamma_{ST} b_{Si} + \delta_{ST}$$

for the item discrimination and item difficulty parameters respectively. Finally, if we look at the entire set of J linking items, then the slope and intercept of the linear transformation can be estimated using the sample means and variances of the item parameters

$$\gamma_{ST} = {}^{\mu(a_S)}\!/_{\mu(a_T)}$$

and

$$\delta_{ST} = \mu(b_T) - \gamma_{ST}\,\mu(b_S).$$

Compositing linking constants

The linking constants described above were calculated for the pairwise grade levels. Now each of these linking constants are the slopes and the intercepts of a linear transformation. Thus, we can denote the linear function that transforms the scale from grade S to the scale of grade T as $f: \theta_s \rightarrow \theta_T$ where $\theta_T = f(\theta_S) = \gamma_{ST}\theta_S + \delta_{ST}$.

Now, suppose we have a second linear transformation from grade R scale to grade S scale given by $g: \theta_R \rightarrow \theta_S$, where $\theta_S = g(\theta_R) = \gamma_{RS}\theta_R + \delta_{RS}$. Then the transformation from the grade R scale to the grade T scale is given by the composition of the functions g and f or

$$\begin{aligned}\theta_T = f(\theta_S) = f\big(g(\theta_R)\big) = f(\gamma_{RS}\theta_R + \delta_{RS}) &= \gamma_{ST}(\gamma_{RS}\theta_R + \delta_{RS}) + \delta_{ST} \\ &= \gamma_{ST} \cdot \gamma_{RS}\theta_R + \gamma_{ST} \cdot \delta_{RS} + \delta_{ST}\end{aligned}$$

If we let $\gamma_{RT} = \gamma_{ST} \cdot \gamma_{RS}$ and $\delta_{RT} = \gamma_{ST} \cdot \delta_{RS} + \delta_{ST}$ then we have that $g \circ f: g: \theta_R \rightarrow \theta_T$ is given by $\theta_T = \ f\big(g(\theta_R)\big) = \gamma_{RT}\theta_R + \delta_{RT}$.

We can use this process to start at any grade level and step by step, taking the composites of the transformation functions using the pairwise linking constants, derive the linking constants needed to place each of the grades on the scale of a given anchor grade to create a vertical scale.

Chapter 6: Norming

Introduction to Norming

A *norm-referenced interpretive framework* is used when inferences regarding a student's test score are made by comparing their score to the distribution of scores in a relevant group (Kolen, 2006; Nitko & Brookhart, 2011). Istation has used such an interpretive framework for its tests have since their inception. As explained in the original technical report for ISIP Math:

> … (W)e 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. (Istation, 2018, pp. 5–1)

Three kinds of norm-referenced scores are reported for ISIP Math, namely, *percentile ranks* (PRs), *levels*, and *instructional tier goals*. The percentile rank shows the percentage of students in the norm group that were lower than a given scale score for a given test grade level and time of year. The percentiles are used in turn to define five broad levels of student performance based on the quintiles of the distribution. That is, the cut scores at the 20th, 40th, 60th, and 80th percentiles are used to define Level 1 through Level 5, which denote increasingly higher student performance. The instructional tier goals are a three-level grouping based on cut scores that are used to help teachers determine the level of instruction for each student. Students whose test scores are below the 20th percentile are said to be in Tier 3 and are at significant risk of not meeting grade-level expectations. Students whose test scores are in Tier 2 (between the 20th and 40th percentiles) are said to be at some risk of not meeting grade-level expectations. Finally, students with test scores above the 60th percentile (Tier 1) are said to be on track to meet grade-level expectations.

The data for developing the original norms were collected from ISIP Math test users from grades prekindergarten through 8 during the 2011–2015 school years (Istation, 2018). The norms were based on students' IRT-based scale scores, and separate sets of norms were developed by time of year (beginning, middle, or end of the school year) for each grade.

However, since the initial ISIP Math norming, there have been changes in both the population of students taking the tests and the tests themselves. More students from more states are now taking ISIP Math than when the assessment was first introduced. In addition, as described in chapter 4, the mathematics item pool has been updated to identify math domains that were comparable across different grades, review the alignment of these items with current standards, and determine which items were aligned with the domains.

In light of these changes, and in order to maintain the relevance of the interpretative framework for ISIP Math test scores, Istation decided to develop new norms. The remainder of this chapter describes the processes that were used to do this.

First, the procedures used in developing the samples of student data that were needed to create the norms are described. These procedures include sampling ISIP Math student test data and using a post-stratification index to construct a nationally representative sample that characterizes public school students in the US. The development of sampling targets, the selection of the samples of students needed, and a brief discussion of the post-stratification results are also presented.

This is followed by the procedures used to develop the overall- and domain-score norms. The specific norm sets that were developed for ISIP Math are shown along with the steps that were used in cleaning the norming sample. This is followed by an outline of the main stages of the continuous norming process that was used: examining the empirical scale score distributions, selecting a family of statistical distributions to model the empirical distributions, modeling the distributional parameters as a function of time of year, and how the results were reviewed.

Sampling Procedures

One of the major goals in this revision was to construct a sample that closely represents students in public schools. To conduct the norms update, we derived a sample from the extensive Istation user database. ISIP Math is administered to students across the US, but these students may not be a representative sample of students across the US based on the districts and schools that subscribe to ISIP Math. Therefore, we constructed a nationally representative sample using *post-stratification* methods.

Post stratification can reduce bias in sampling as long as the stratification variables have a relationship with the population and its characteristics (Jagers, 1986). Therefore, the post stratification variables used to compose the normative sample must have a relationship with student achievement. Decades of research in education show that socioeconomic status at either the student or the school level predicts substantial variance in student achievement as measured by test scores. Istation does not require that our users provide student-level demographic information regarding gender, race/ethnicity, or economic status. Approximately half of the students have missing data in these areas. Because of incomplete data at the student level, we relied on the school-level characteristics to conduct post stratification.

The post stratification of the sample was completed in a series of steps. First, we created a post-stratification index to simplify the process. Second, we created population targets based on enrollment information from the National Center for Education Statistics. Third, we selected eligible student observations based on patterns of missing or non-missing data by month. Fourth, we randomly selected from the eligible students within each stratum to construct the final sample.

Post-Stratification Index

To construct the post-stratification index, we relied on research regarding the *school challenge index*, designed by researchers at the Northwest Evaluation Association (NWEA) and based on the *similar schools index* in California (Thum & Hauser, 2015). The original school challenge index was created to encompass the known sociodemographic characteristics that contribute to differences in school-level student achievement. We modified the index to include variables that helped to account for unique characteristics of the Istation database.

The school-level variables we used in the index were selected because of their importance and established relationship with student achievement. We also included the region of the country as this may impact student achievement based on local funding and policy. Istation also has several schools that are under the Bureau of Indian Education (BIE), so we also included this variable in the index.

We constructed the index from school-level information available in the National Center for Education Statistics' *Common Core of Data in the Public School Universe* (NCES-CCD-PSU). We created continuous variables based on population rates for socioeconomic disadvantage and race/ethnicity. These variables are explained in Table 6.1.

The first step in this process was to collect the data from NCES. We used the most recent year available, which was the 2017–2018 school year (U.S. Department of Education, 2018). We obtained the data for public schools, public charter schools, and BIE schools. We compared the NCES list of schools with schools in the Istation database. There were a few public schools that were in Istation but not in NCES. We added these schools to the NCES list with any known information such as region and school district.

Since the NCES information is based on administrative data, there were observations with missing values in addition to the schools that were present in Istation but missing from NCES. To account for missing data, we imputed the values using predicted means matching in *R* statistical software. Predicted means matching is a regression-based method that imputes a value based on known values in the data set. We imputed five different values for the missing values. The final data set had 99,786 unique schools.

Table 6.1. *Information from NCES Enrollment Data Used in Construction of the Composite Index*

Variable	Description
Free or Reduced-Priced Lunch (FRPL)	**Percentage of students eligible for free or reduced-priced lunch**
Percent White	Percentage of students who are non-Hispanic White
Percent Black or African American	Percentage of students who are non-Hispanic Black or African American
Percent Hispanic	Percent of students who are of Hispanic origin of any race
Teacher	Total number of full-time teachers
Teacher-Pupil Ratio	Ratio of teachers per student
Locale of school	Whether the school is located in a rural, urban, or suburban area. Towns were divided between suburban and rural areas.
Bureau of Indian Education School	School is a BIE school or a tribally controlled school.
Magnet	School is a magnet school.
Charter	School is a charter school.
School Level	School is an elementary, middle, high, or multi-grade school.
Type of School	School is a regular, special education, or vocational school.
Region of the Country	Which census region the school is located in (Northeast, South, West, Midwest)
Title I Eligibility	Whether or not the school is eligible for Title I funds
Title I Type of Program	If eligible, the type of program the school implements, partial or school-wide

The next step consisted of transforming the continuous variables. Since sociodemographic rates are not normally distributed, we transformed them into logit units. The categorical variables were transformed into dummy variables. These variables were then put into a regression model with the percentage of students receiving free or reduced-priced lunch (FRPL) as the dependent variable. The results from the regression model are in Table 6.2.

Table 6.2. *Results from the Regression Model for Constructing the School Stratification Index*

Variable (N = 99,786)	*Coefficient*	*SE*	*Beta*	*t*	*p*
Intercept	-3.55	0.03		-104.22	0.00
Percent White	-0.23	0.00	-0.17	-45.68	0.00
Percent Black/African American	0.02	0.00	0.01	4.23	0.00
Percent Hispanic/Latino	0.10	0.00	0.11	31.88	0.00
FTE Teachers	0.15	0.00	0.14	42.73	0.00
Teacher-Student Ratio	0.21	0.00	0.14	47.13	0.00
Locale (REF = Urban)	-	-	-	-	-
Suburban	0.05	0.01	0.01	3.05	0.00
Rural	0.51	0.02	0.13	29.64	0.00
Type of School (REF = Elementary)	-	-	-	-	-
Middle Schools	-0.02	0.02	0.00	-1.33	0.18
High Schools	-0.12	0.02	-0.02	-7.52	0.00
Multi Grade Schools	-0.05	0.02	-0.01	-2.93	0.00
Region (REF = South)	-	-	-	-	-
Northeast	-0.44	0.02	-0.08	-24.89	0.00
Midwest	0.06	0.02	0.01	3.93	0.00
West	0.04	0.02	0.01	2.42	0.02
Title I Eligibility	0.93	0.02	0.21	61.63	0.00
Title I Program (REF = Partial)	0.12	0.02	0.03	8.27	0.00
BIE School	0.59	0.14	0.01	4.34	0.00
Charter	-0.50	0.02	-0.07	-22.37	0.00
Magnet	0.16	0.03	0.02	5.30	0.00
Regular	0.36	0.02	0.05	15.18	0.00

Using the predicted value for the outcome measure, we rescaled them to create a normal curve equivalent.

$$School\ Index = 50 + 21.06[(predicted\ value - mean)/SD]$$

Next, the school index (SI) was divided into eighths of the distribution, or octiles. A low value indicated schools with greater challenges due to sociodemographics and locale, and a high value indicated schools with fewer challenges. Since the regression model showed that the middle school level was not significantly different than elementary schools, we did not run separate regression models based on type of school. We divided the indexes into eight equal parts, with 12,473 schools represented in each octile. The mean ISIP Math score increased by each octile with a few exceptions, indicating that the variable worked to account for achievement at the school level. Table 6.3 shows the mean ISIP Math score, using the original ISIP Math scale for the middle-of-the-year scores, across grades and SI octile levels.

Table 6.3. *Means of Midyear Original ISIP Math Scores by Grade and School Index Octile Levels*

Grade	SI1	SI 2	SI 3	SI 4	SI 5	SI 6	SI 7	SI 8
Pre-K	2037	2064	2083	2077	2077	2122	2053	2079
K	2060	2079	2100	2123	2127	2168	2192	2166
1	2034	2070	2096	2106	2131	2171	2191	2155
2	1974	2001	2011	2019	2035	2064	2077	2103
3	1928	1944	1955	1958	1968	2007	2021	2039
4	1967	1988	1972	1999	2015	2063	2041	2080
5	1954	1957	1983	1987	2007	2063	2076	2099
6	2001	1997	1961	2008	2043	1990	2028	2058
7	1989	1971	2016	1982	1998	1959	1991	2134
8	2025	1968	1995	1949	2017	1928	2017	2086

Sample Targets and Selection

We then computed sampling targets using the enrollment data for elementary and middle schools. We created the targets based on type of school because we were relying on administrative data, which can fluctuate by year, and thus using an aggregate across grades is preferred. We computed the targets separately by school level, as there are fewer middle schools than there are elementary schools. Sixth grade was considered to be in middle school. We compared the targets to the enrollment in the Istation Math program and determined that our database tended to skew to the lower octiles, indicating the need for post stratification.

Selecting eligible student observations

We used two separate norming years for the overall score and the domain scores. The 2019-2020 school year was disrupted by the COVID-19 pandemic, so we used the 2018–2019 school year as the base norming year for the overall score. This school year was not selected for the domain scores due to changes implemented in the assessment. Prior to the 2019–2020 school year, students received approximately 20 items per assessment, and the selection was only based on the difficulty of the item, not the particular domain. In 2019–2020, we began administering an equal number of items per domain to collect data for norming domain scores. This was the only year data was available for domain scores. We calibrated the performance of the two years, which is described below.

Overall Score Sample. Since ISIP Math can be used for progress monitoring and benchmarking, selecting only those observations with complete data can bias the sample. In addition, the benchmarking month can vary by school district or state. Istation divides instructional months depending on the first day of school, and the first month is considered Period 0. Subsequent months are numbered sequentially. We evaluated the monthly number of ISIP Math scores and noticed that the majority of students assessed monthly, with some peaks in Periods 1, 5, 8, and 9. Benchmarking patterns were observed in Periods 0, 5, and 9; Periods 1, 5, and 8; and Periods 1, 5, and 9. Student observations were selected as eligible for norming if they had scores in the benchmark months or if they had scores in five or more periods. Student observations that did not have all eligible benchmark months or less than five periods were determined as ineligible for selection. This reduced the number of eligible observations.

Next, after creating the sample targets for each SI, we conducted sampling by grade. Sampling was conducted without replacement for the lower grades (pre-K–3) and with replacement in grades 4-8. While the targets were set for public school enrollment, which is the majority of Istation users, we also included some observations from private or parochial schools, approximately 1%. The exception was prekindergarten, where 15% of the sample came from private or parochial schools.

Domain Score Sample. The same process was used for the sample for the domain scores; however, we used data from the 2019–2020 school year that had scores available in the first two benchmark months and scores in March before the schools closed. The 2019–2020 school year was the only year for which data were available to compute the domain scores. Most students received approximately 20 items, or 5 items per domain. Some students, however, received fewer items. These students were included in the data set because the CAT algorithm had converged with fewer items, and eliminating these students may have biased the sample.

Sample calibration. Since the domain scores and the overall scores come from two different norming years, it was important to calibrate the sample so that they were equivalent on ability. We used the middle-of-the-year (MOY) overall score from the domain score file, and when the sample for the overall score was selected, the selection was weighted so that the samples were equivalent on their math ability for MOY. This was done using code in *R* statistical software that was developed internally for sample selection.

Final Sample Description

Table 6.4 shows the percent of observations from each SI and the sample targets based on the NCES enrollment in each type of school.

Table 6.4. *Percent of Public School Students by School Index Octile and of Private/Parochial School Students in Overall- and Domain-Score Samples*

Grade Span	Targets and Actuals	SI 1	SI 2	SI 3	SI 4	SI 5	SI 6	SI 7	SI 8	Private/ Parochial School
Pre-K to 5	NCES Target	19.0%	16.7%	14.5%	11.7%	10.5%	10.6%	12.0%	5.2%	n/a
Pre-K to 5	Overall Score Sample (N = 61,500)	19.0%	16.6%	14.4%	11.6%	10.2%	10.5%	11.8%	4.6%	1.3%
Pre-K to 5	Domain Score Sample (N = 61,500)	19.0%	16.6%	14.4%	11.6%	10.2%	10.3%	11.6%	4.9%	1.4%
6 — 8	NCES Target	17.9%	14.2%	12.7%	10.8%	10.6%	13.9%	14.6%	5.3%	n/a
6 — 8	Overall Score Sample (N = 10,000)	18.4%	15.5%	13.6%	11.2%	10.4%	12.0%	13.0%	5.1%	1.0%
6 — 8	Domain Score Sample (N = 10,106)	17.5%	14.0%	12.5%	10.5%	10.4%	13.6%	14.3%	5.2%	2.0%

Source for NCES Targets: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, Common Core of Data 2017-2018.

The total final sample for the overall and domain score samples consisted of 71,500 and 71,506 students respectively, of which about 10.8% had been resampled. The NCES targets show the distribution by SI, which is for public schools. The sample closely matches the NCES targets; however, it varies slightly as we added some observations from private and parochial schools.

Norming Analysis

Initial Considerations and Data Preparation

As described in chapter 4, ISIP Math produces a scale score that describes overall student performance in mathematics. In addition, the assessment is also designed to produce domain scores in four areas for grades prekindergarten through 5 and a separate set of domain scores in four areas for grades 6 through 8. In all, 50 sets of norms needed to be developed for the assessment, and these are summarized by grade and domain in Table 6.5.

Table 6.5. *ISIP Math Norms Developed by Grade and Domain*

Grade	Overall	Comp. & Alg. Think.	Num. Sense	Num. System	Meas. & Data Analysis	Stats. & Data Analysis	Geom.	Geom. & Meas.
Pre-K	Yes	Yes	Yes	No	Yes	No	Yes	No
K	Yes	Yes	Yes	No	Yes	No	Yes	No
1	Yes	Yes	Yes	No	Yes	No	Yes	No
2	Yes	Yes	Yes	No	Yes	No	Yes	No
3	Yes	Yes	Yes	No	Yes	No	Yes	No
4	Yes	Yes	Yes	No	Yes	No	Yes	No
5	Yes	Yes	Yes	No	Yes	No	Yes	No
6	Yes	Yes	No	Yes	No	Yes	No	Yes
7	Yes	Yes	No	Yes	No	Yes	No	Yes
8	Yes	Yes	No	Yes	No	Yes	No	Yes

The data collected in the study were broken up into grade-specific files based on overall- and domain-score samples respectively. The norms for the overall scale scores were developed first on a grade-by-grade basis and then were followed by analyses for each of the mathematics domains.

The files for the overall-score samples included fields showing the following:

- unique student identifier
- grade
- unique identifier for the student's school
- octile of the school index used to post-stratify schools
- student's ISIP Math overall CAT scale score
- scale score standard error
- overall performance level

The final three fields were repeated for each test administration period for which a given student had ISIP Math data. The administration periods were denoted using Table 6.6, so a field denoted *score_overall_5,* for example, indicates the overall CAT score for a student taking the test in the fifth month of the school year. Istation sets Period 0 as the initial month of school followed by Periods 1 through 9 sequentially. Depending on the school start date, Period 1 is usually September, Period 2 is October, and so on. Schools that start in July would have a different calendar month for their administration period. The most common test administration periods are below in Table 6.6. However, the sample did not constrain the calendar month and period month; therefore, the data for Period 1 may have some assessments from August or October, and Period 5 may have some assessments from December, depending on the day school started in a given district or school.

Table 6.6. ISIP Math Test Administration Periods

Period	Time of Year
0	August
1	September
2	October
3	November
4	December
5	January
6	February
7	March
8	April
9	May

The domain-score sample files included similar fields but were much more extensive, as the set of three fields used to denote student achievement needed to be repeated for each domain and test administration period taken by the student. Once the files were uploaded, the data preparation steps included checking all variables for out-of-bound and missing values and removing any duplicate student cases or cases with completely missing item response strings.

This was followed by calculating summary measures to describe the distribution of student scale scores for each combination of grade, period, and norm set needed in the analysis. These statistics included the following:

- the number of cases, mean, standard deviation, skewness, and kurtosis;
- the minimum and maximum scale scores observed; and
- the values of key percentiles, including the 1st, 5th, 10th, 20th, 25th, 40th, 50th, 60th, 75th, 80th, 90th, 95th, and 99th percentiles.

The percentiles were selected to provide information regarding the tails of the distribution, the median, the interquartile range, and the cut scores that are used by ISIP Math to report students' performance levels and instructional tier goals.

Norming Approach

ISIP Math requires normative information to be developed for each period of the school year. In traditional approaches to norming, demographic variables such as the time of year that a student took a test would be treated as a discrete variable. The students sampled at each of the times would then represent different subgroups that required separate norms be estimated.

However, this approach ignores the fact that variables such as time of year and age are actually continuous in nature. By using models that treat time period as a continuous variable to predict test scores, one can use information taken from across the entire year to estimate norms. This approach, called *parametric continuous norming,* was used to develop the new norms for ISIP Math (Zhu & Chen, 2011; Voncken et al., 2019; Lenhard et al., 2018).

As described by Lenhard et al. (2019) in the context of raw-score-based age norms, "(k)nown parametric functions are used to model the raw score distributions at specific age levels. The function parameters are subsequently modeled as a function of explanatory variables such as age" (p. 5, Fig. 2).

In developing the ISIP Math norms, this approach was modified to substitute overall/domain scale scales for raw scores and within-grade, time-of-year periods for ages. For each grade and set of overall or domain scale scores, the process consisted of the stages outlined in the next section.

Examination of Empirical Scale Score Distributions

The empirical scale score distributions were examined across the time-of-year periods. We used descriptive statistics and boxplots, histograms, and scatterplots to assess the family of statistical distributions that would most reasonably model the observed scale score distributions.

Although some of the score distributions appeared to be approximately normal, in most cases, the distributions showed a great deal of skewness — which often changed throughout the school year. An example of the skewed nature of the scaled score distributions can be seen for the first-grade overall scale scores. Table 6.7 shows the descriptive statistics for these data, while Figures 6.1 and 6.2 show the boxplots and histograms.

The descriptive statistics clearly show the change in the skewness across the year, as does the change in the placement of the median line with respect to the lower and upper quartiles in the boxplots. The boxplots also show that most of the outliers for periods at the beginning of the year are for high scale score values (positive skew), more evenly divided towards the middle of the year (approaching symmetry), and low scale score values at the end of year (negative skew). The change in the skewness across the year can be best seen in histograms. Each histogram has been fitted with a normal distribution based on the mean and standard deviation of the scale score distribution for that period and highlights the departure of the empirical distribution from normality.

Due to the skewness of most of the scale score distributions, the more flexible *beta distribution* was used as the parametric function to model the distributional shapes rather than the normal distribution. The beta distribution is given by the probability density function

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

where $\alpha > 0$ and $\beta > 0$ are parameters governing the shape of the density function, $\Gamma(\cdot)$ is the gamma function, and $0 \leq x \leq 1$.

Table 6.7. *ISIP Math Grade 1 Normative Sample: Overall Scale Score Descriptive Statistics by Time of Year*

Period	N	Mean	SD	Skewness	Kurtosis
0	2,425	403.6	58.1	0.59	0.10
1	9,909	408.5	60.2	0.46	0.17
2	9,915	421.8	60.3	0.29	0.11
3	9,919	436.6	61.9	0.21	0.15
4	9,939	447.2	62.8	0.14	0.11
5	9,951	454.9	64.9	0.10	0.19
6	9,959	464.5	66.5	0.04	0.13
7	9,945	476.7	67.8	-0.06	0.09
8	4,370	506.5	65.9	-0.19	0.02
9	3,284	511.6	68.7	-0.28	-0.09

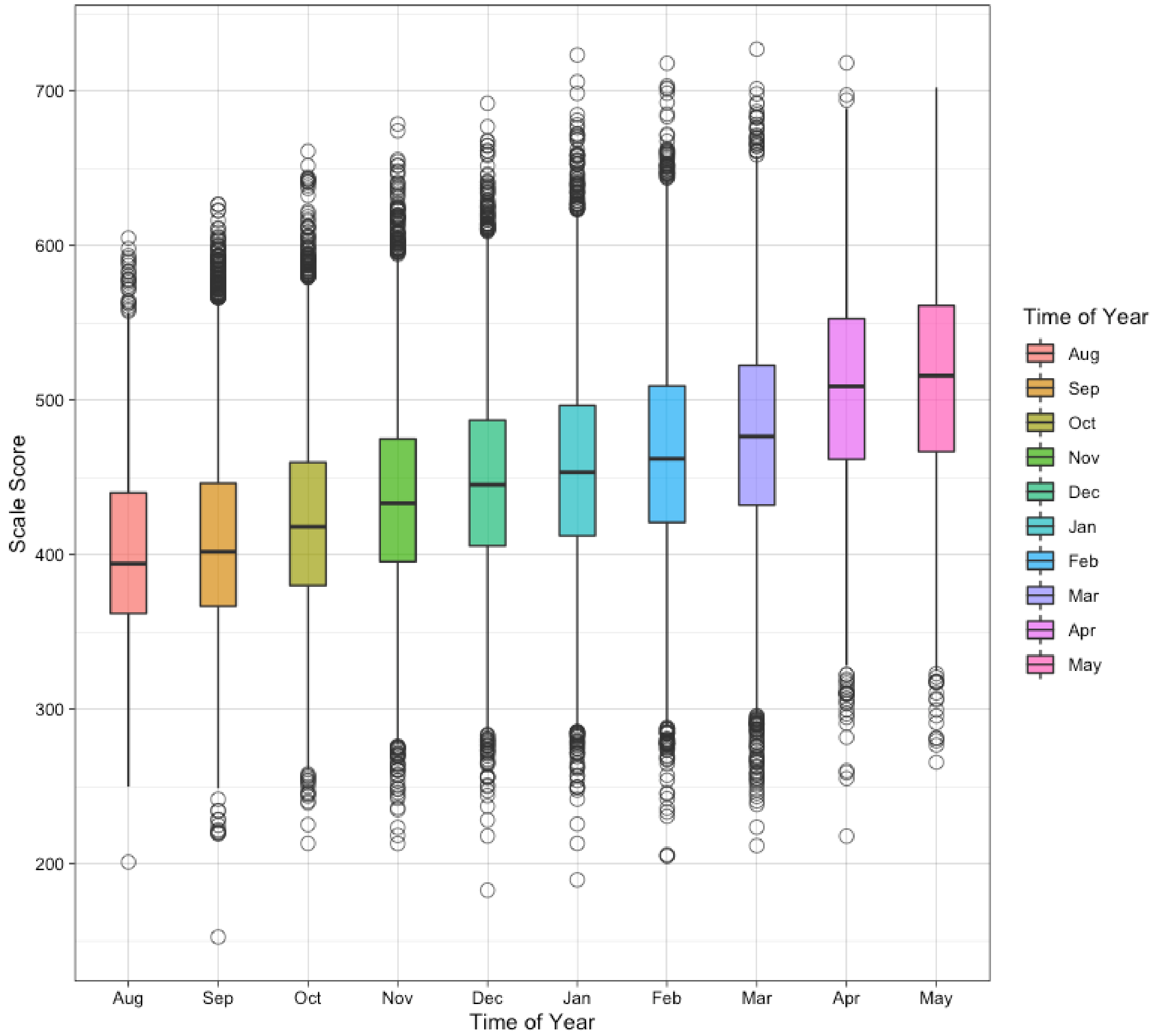


Figure 6.1. Example: ISIP Math Grade 1 Normative Sample: Boxplots by Time of Year

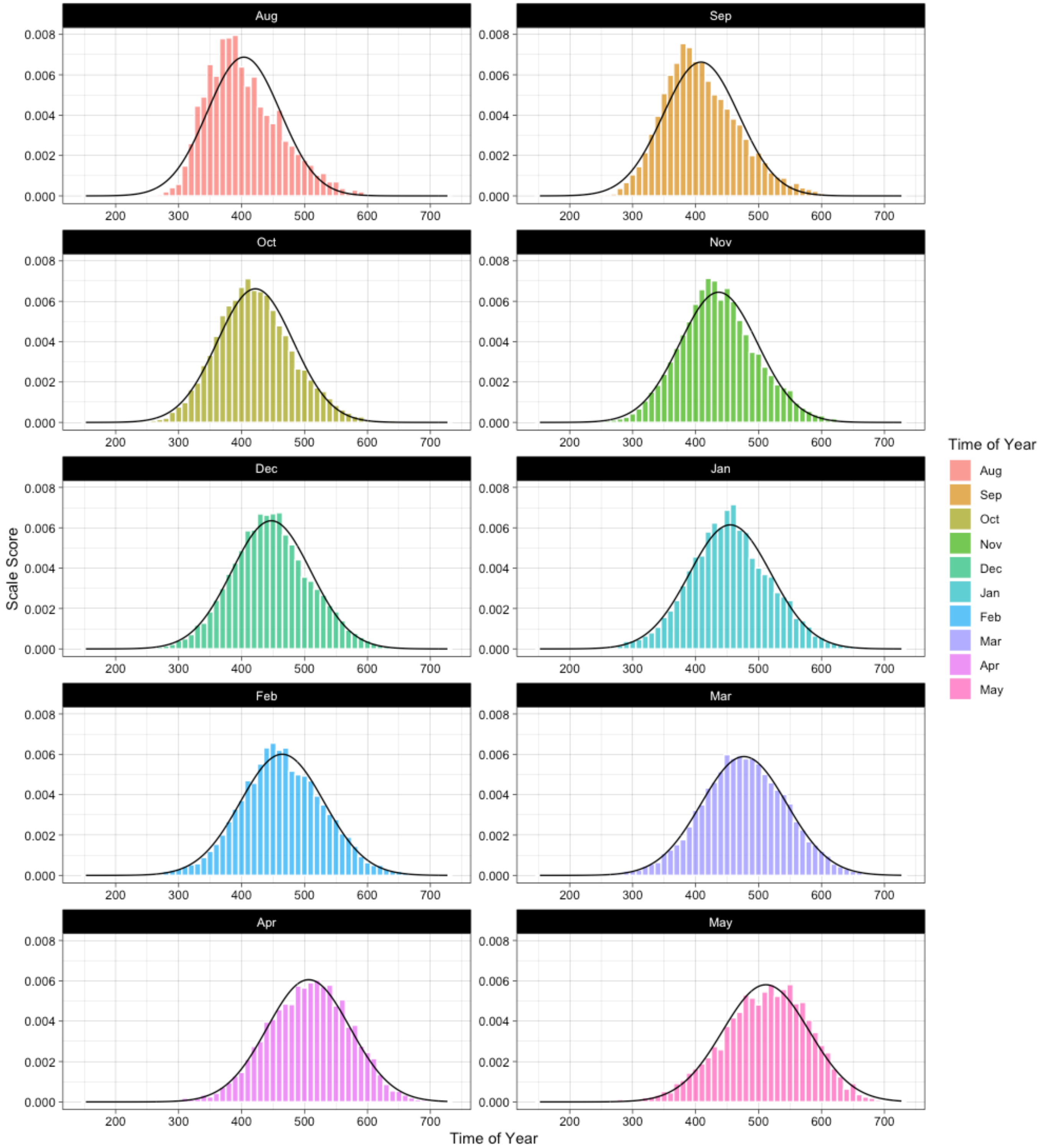


Figure 6.2. *Example ISIP Math Grade 1 Normative Sample: Histograms of Scaled Scores by Time of Year with Normal Distribution Overlay*

Modeling the Distributional Parameters

After choosing to use the beta distribution to model the shape of the empirical scale score distributions, the next step involved modeling its shape parameters as a function of the time of year.

First, polynomial regression was used to fit selected percentiles from the empirical scale score distribution as a function of the time of year. These smoothed percentiles were then used to estimate the parameters of the beta distribution using R (R Core Team, 2018) for the analysis and the function `beta.parms.from.quantiles` (Blisle, 2017).[7]

One of the issues encountered at this stage involved using 3rd or 4th degree polynomial functions; although they produced closer fits to the empirical scale scores, they oftentimes resulted in the smoothed "bowing" or "reversing" across the time-of-year periods. That is, it was possible that students at a given scale score and percentile rank at midyear could find themselves scoring at a higher percentile rank for the same scale score at the end of the year. In order to meet substantive expectations of growth across the year, 2nd degree polynomial functions (parabolas) were used to constrain the growth to be monotonic, non-decreasing scale score values.

Generating and Reviewing the Percentiles for the New Norms

Once these parameters were estimated for a given period, they were used to generate for each percentile rank of the theoretical distribution (i.e., from 1 to 99) the corresponding scale score percentiles that would be used as norms. The final scale score percentiles for the new norms were rounded to the nearest whole numbers on the 100 to 900 ISIP Math reporting scale.

[7] As a technical point, It should be noted that the beta distribution only takes on values from 0 to 1. Thus, all of the smoothed percentiles on the ISIP Math scale needed to be rescaled from 100 to 900 to the 0 to 1 metric so that they could be used to estimate the parameters for the beta distribution. The final results were then scaled back to the original reporting scale.

The norms that were produced were reviewed to assess their reasonableness by examining scatterplots of the smoothed scaled score values, comparing the newly produced norms with the current norms and applying the current and new norms to current and previous years' student data. Additional checks examined the percentages of students that fell into the ISIP Math levels and instructional tiers under the current norms versus the new norms.

A major goal of this renorming effort was to update the norms and make them more rigorous than the previous norms. To evaluate whether we met this goal, we obtained data from an urban school district in Ohio that had used Istation and ISIP Math for several years and compared the students' performance on the ISIP Math and the Ohio AIR. The sample consisted of approximately 1,900 students in grades 3 through 8. The Ohio AIR is administered yearly to students in the state, who were then given a proficiency rating of limited, basic, proficient, accelerated, or advanced. We used data from the 2018 – 2019 school year, the same year that was used for norming purposes. We set a cut point at the *proficient* level or above and evaluated the percentage of students who scored in Tier 1 in the middle of the year using both the previous and updated norms to determine if the updated norms were better able to identify students at risk of not scoring *proficient* or higher. The second column in Table 6.8 shows the percent of students in the state who passed the Ohio AIR (Ohio Department of Education, 2019). The third column shows the percent of students in our data who scored at Tier 1 or higher under the previous norms, and fourth column shows those who scored at Tier 1 or higher in the updated norms.

Table 6.8. *Comparison of Ohio AIR Proficiency Rates, and the Percent of Students Who Reached Tier 1 in the Previous and Updated Norms and Met the Proficient Level or Higher*

Grade	% Passing Ohio AIR 2018-2019 SY	Previous Norms: Tier 1 students who scored *Proficient* or above	Updated Norms: Tier 1 students who scored *Proficient* or above
3	67.1%	57.9%	83.8%
4	74.3%	66.9%	76.2%
5	65.0%	51.4%	78.9%
6	60.1%	40.8%	57.1%
7	57.5%	37.7%	72.3%
8	57.3%	27.9%	54.5%

Using the tier system, under the previous norms, 57.9% of students scoring in Tier 1 in third grade also reached the proficient level or above on the Ohio AIR; however, with the updated norms, 83.8% reached the proficient level or above. In fourth grade, 74.3% of students in the state reached proficiency, and under the previous norms 66.9% of students reached the proficient level or above, while in the updated norms, 76.2% reached proficient or above. Similar patterns are apparent across grades 5 through 8. The updated norms appear to better identify students who may be at risk of not reaching the proficient or above status, indicating that we met the goal of having more rigorous norms.

Chapter 7: Reliability and Validity

During the initial development of the ISIP Math assessment, studies for test-retest reliability and concurrent and predictive validity work were conducted. We compared ISIP Math scores to scores from norm-referenced measures with good psychometric properties of similar constructs. ISIP Math scores were compared to STAR Math™, the Test of Early Mathematics Ability-Third Edition (TEMA-3), the Stanford Achievement Test – Tenth Edition (SAT10), and the State of Texas Assessments for Academic Readiness (STAAR).

Internal Consistency

Data for an internal consistency study came from three school districts in Texas. Students in the sample consisted of 13.8% African American, 36% Hispanic/Latino, 42.9% White, and 4.8% Asian, and the remainder were multiple race ethnicities, American Indian, or Native Hawaiian/Other or Pacific Islander. Over forty seven percent (47.9%) were receiving free or reduced-priced lunch (FRPL), and 48.7% were female, and 51.3% were male. We obtained criterion-related evidence using the STAR Math™, TEMA-3, SAT10, and STAAR.

STAR Math assesses a similar construct as ISIP 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.

The TEMA-3, which seeks to identify students significantly behind or ahead of peers in mathematical skills, was used as a criterion assessment for kindergarten through second 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 end-of-year 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 eighth grade. The criteria used within this study were identified by the National Center on Response to Intervention (NCRTI) in 2010 and include the following:

- 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 domain that 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 domain are believed to be accurate representations of the domain that they intend to measure.

Generalizability

Generalizability was analyzed to illustrate the extent to which the analytic sample for the study was comparable to the state and national population. Tables 7.1 — 7.3 shows the comparison of the analytic sample to the national distribution.

Table 7.1. *Comparison of Demographics for Race/Ethnicity for the State, National, and Recruited Sample*

	Statewide Distribution[a] %	National Distribution[bcd] %	Sample Distribution %
African American	12.61	15.60	13.76
Hispanic/Latino	52.22	24.88	36.05
American Indian/ Alaska Native	.39	1.05	.42
Asian	4.03	5.18	4.83
Native Hawaiian/Other or Pacific Islander	.14	-	.42
Two or More Races	2.05	3.02	2.23

[a] Texas Education Agency (2015).

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

[c] U.S. Department of Education, National Center for Education Statistics, Common Core of Data (2015).

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

Table 7.2. *Comparison of Demographics for Free/Reduced Priced Lunch Status for the State, National, and Recruited Sample*

	Statewide Distribution[a] %	National Distribution[bcd] %	Sample Distribution %
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, Common Core of Data (2012).

[c] U.S. Department of Education, National Center for Education Statistics, Common Core of Data (2015).

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

Table 7.3. *Comparison of Demographics for Gender for the State, National, and Recruited Sample.*

	Statewide Distribution[a] %	National Distribution[bcd] %	Sample Distribution %
Male	51.30	51.40	51.29
Female	48.70	48.60	48.71

[a] Texas Education Agency (2015).

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

[c] U.S. Department of Education, National Center for Education Statistics, Common Core of Data (2015).

[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. We conducted several validity studies and results are available in Table 7.4. 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.

Table 7.4. *Concurrent-Related Evidence for Validity*

Assessment	n	Coefficient
STAR Math (BOY) Grade 1	208	.66
STAR Math (BOY) Grade 2	185	.76
STAR Math (BOY) Grade 3	170	.71
STAR Math (BOY) Grade 4	81	.64
STAR Math (BOY) Grade 5	224	.55
STAR Math (BOY) Grade 6	174	.74
STAR Math (BOY) Grade 7	222	.61
STAR Math (BOY) Grade 8	165	.61
STAR Math (MOY) Grade 1	212	.77
STAR Math (MOY) Grade 2	183	.81
STAR Math (MOY) Grade 3	169	.75
STAR Math (MOY) Grade 4	69	.67
STAR Math (MOY) Grade 5	198	.71
STAR Math (MOY) Grade 6	173	.77
STAR Math (MOY) Grade 7	199	.60
STAR Math (MOY) Grade 8	167	.59
STAR Math (MOY) Grade 8	167	.59
STAR Math (EOY) Grade 1	213	.72
STAR Math (EOY) Grade 2	181	.75
STAR Math (EOY) Grade 3	167	.74
STAR Math (EOY) Grade 4	81	.78
STAR Math (EOY) Grade 5	235	.76
STAR Math (EOY) Grade 6	162	.80
STAR Math (EOY) Grade 7	211	.76
STAR Math (EOY) Grade 8	145	.61
STAR Math (EOY) Grade K	152	.49
STAR Math (EOY) Grade 1	210	.66
STAR Math (EOY) Grade 2	195	.69
SAT 10 Grade 3	196	.82
SAT 10 Grade 4	131	.82
SAT 10 Grade 5	250	.82
SAT 10 Grade 6	197	.83
SAT 10 Grade 7	146	.57
SAT 10 Grade 8	152	.67
SAT 10 PS Grade 3	196	.82
SAT 10 PS Grade 4	131	.82
SAT 10 PS Grade 5	250	.75
SAT 10 PS Grade 6	197	.83
SAT 10 PS Grade 7	146	.45
SAT 10 PS Grade 8	152	.65

SAT 10 P Grade 3	196	.69
SAT 10 P Grade 4	131	.71
SAT 10 P Grade 5	250	.78
SAT 10 P Grade 6	197	.74
SAT 10 P Grade 7	146	.58
SAT 10 P Grade 8	152	.54
STAAR Grade 3	190	.81
STAAR Grade 4	129	.80
STAAR Grade 5	241	.81
STAAR Grade 6	234	.85
STAAR Grade 7	192	.70
STAAR Grade 8	130	.68

It 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 complete battery and its two subtests (Problem Solving (PS) and Procedures (P)), and the STAAR.

Discussion

Reliability and validity are two important qualities of measurement data. Reliability can be thought of as consistency, either consistency between items within a testing instance or between 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 eighth grade students at three school districts in Texas during the 2015-2016 school year.

Sensitivity and Specificity

We also conducted classification accuracy to determine the sensitivity and specificity for detecting students at risk. The *sensitivity* of ISIP Math for kindergarten through second 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 second grade using TEMA-3 as the criterion assessment was lower, ranging 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 to 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 to .70 across grades, indicating that 29 to 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 second grade ISIP Math results provide partially convincing evidence for classification accuracy based on TEMA-3, while first 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.

Details from the full validity study are available. To review the complete validity study, “Imagination Station (Istation): Istation’s Indicators of Progress (ISIP) Math Validity Studies – Overview of Results,” visit the following webpage and click the link found under the Archive Technical Reports heading.
http://www.smu.edu/Simmons/Research/RME/Explore/Publications

ISIP Math Predictive Validity

In this ISIP Math update, we report additional evidence for validity with two separate studies. The first study was conducted for third grade students who were also assessed with the ACT Aspire, and the second study was conducted for third through eighth grade students who also took the Ohio AIR assessment.

ACT Aspire

Patarapichayatham & Locke (2020a) conducted an analysis of the ACT Aspire and the ISIP Math using data from the 2017-2018 and 2018-2019 school years. A full description of this study is available at www.istation.com/studies. This research sought to determine the relationship between second graders' ISIP Math EOY scores and their third-grade end-of-year performance on the ACT Aspire, and between third graders' ISIP Math MOY scores and their end-of-year performance on the third grade ACT Aspire. The ACT Aspire assessments are vertically scaled, and they are aligned with the standards that target college and career readiness (ACT, 2019). Previous scale scores for the ISIP Math were converted to the new scale using the methodology described in chapter 5.

All data for this analysis came from students in Arkansas. All races/ethnicities are represented in the sample. The Asian/Other group consists of students who were Asian, Native American or Native Hawaiian, and multi racial students. A full description is available in Table 7.5.

Table 7.5. *Demographic Characteristics of the ACT Aspire Study*

Category	Demographic	Math Grade 2 N=8,381	Math Grade 3 N=4,774
Gender	Female	48.9%	48.7%
Gender	Male	51.1%	51.3%
Race/Ethnicity	White	68.6%	66.9%
Race/Ethnicity	African American	12.4%	8.9%
Race/Ethnicity	Hispanic or Latino	13.3%	17.0%
Race/Ethnicity	Asian or Other	5.7%	7.2%
Language	English	88.6%	84.3%
Language	Spanish	9.9%	13.6%
Language	Other Language	1.5%	2.1%
Language Status	EL	9.7%	11.5%

We first conducted Pearson product-moment correlations between the ISIP Math and the ACT Aspire. These are available in Table 7.6. Correlations are strong for both second grade EOY and third grade MOY scores.

Table 7.6. *Pearson Product-Moment Correlation Coefficients between ISIP and ACT Aspire*

Grade	ISIP Math Correlation with ACT Aspire
2	.76
3	.77

Next, we used a multinomial logistic regression model to calculate probabilities for reaching Close, Ready, or Exceeding expectations. The probabilities of reaching *Ready* or above at key percentiles are available in Table 7.7.

Table 7.7. *Probabilities of reaching Ready or Higher on the ACT Aspire at Key Percentiles for 3rd Grade EOY*

Percentile Rank	Grade 2 ISIP EOY	Probability	Grade 3 ISIP MOY	Probability
20	462	.399	478	.427
40	493	.627	505	.680
60	520	.800	529	.858
80	550	.922	557	.965

Students who score in the 40th percentile or above have a greater than 60% probability of reaching *Ready* or higher. Students who score in the 60th percentile (level 4 or above) have an 80% probability of achieving *Ready* or higher.

Classification accuracy was also calculated for third grade students. A cut point at the 45th percentile rank was the best differentiation for all students and by subgroups for race/ethnicity and gender. The sensitivity was .83, indicating that 83% of students who met or exceeded the 45th percentile met *Ready* or higher. Specificity was .81, meaning that 81% of the students who did not meet the threshold did not meet *Ready* or higher standards. The positive predictive power was .89, and the negative predictive power was .72 (Patarapichyatham & Locke, 2020a).

Ohio AIR

The Ohio State Board of Education adopted the Common Core State Standards (CCSS) in both English language arts (ELA) and mathematics as Ohio's Learning Standards (Patarapichyatham & Locke, 2020b). The state of Ohio requires all students in grades 3 through 8 to take standardized tests in mathematics each year, and the Ohio Department of Education worked with Ohio educators and the American Institutes for Research (AIR) to develop the state assessments. Content advisory and sensitivity committees determined whether test items were suitable for the course, accurate, fair, and measured Ohio's Learning Standards (AIR, 2019).

The Ohio AIR performance levels are used to place students' assessment scores in one of five levels of achievement: level 1 – Limited, level 2 – Basic, level 3 – Proficient, level 4 – Accelerated, or level 5 – Advanced.

The sample for this study came from approximately 1,900 students in an urban school district in Ohio. A full description of this study is available at www.istation.com/studies.

Table 7.8 describes the demographics of the sample. More than 50% of the students were African American or Black (AA), and the remainder were White, Hispanic/Latino, or Others. These data were from the 2018-2019 school year, and the scores were converted from the previous scale to the current scale. We used MOY scores to calculate the proficiency projection.

Table 7.8. *Demographic Characteristics of the Ohio AIR Study*

Grade	N	AA	Hispanic/ Latino	White	Others	Female	Male	EL: Yes	EL: No
3	337	53.9%	19.0%	17.6%	9.5%	46.3%	53.7%	3.5%	96.5%
4	399	54.0%	17.9%	14.9%	13.2%	45.1%	54.9%	5.8%	94.2%
5	335	55.9%	19.9%	15.5%	8.7%	44.7%	55.3%	5.1%	94.9%
6	311	59.4%	19.3%	13.0%	8.3%	49.6%	50.4%	4.6%	95.4%
7	290	55.7%	23.1%	14.1%	7.1%	50.9%	49.1%	4.8%	95.2%
8	236	56.2%	26.2%	11.2%	6.4%	52.8%	47.2%	5.6%	94.4%

Pearson product-moment correlations between ISIP Math and Ohio AIR were computed by grade. Correlations are available in Table 7.7. Correlations are moderate to strong across grades 3 through 8.

Table 7.7. *Correlations Between ISIP Math and Ohio AIR*

Grade	Correlation
3	0.70
4	0.66
5	0.73
6	0.76
7	0.69
8	0.50

Probabilities for reaching the *Proficient* level or above on the Ohio AIR are available in Table 7.8. The table shows key percentiles that correspond to the cut points for the ISIP levels, their associated scores at MOY, and the probability of reaching *Proficient* or higher on the Ohio AIR.

Table 7.8. *Probability of Reaching Proficient or Higher on the Ohio AIR at Key Percentiles of ISIP Math*

Grade	Percentile Rank	Score	Probability
3	20	474	.228
3	40	500	.506
3	60	523	.781
3	80	549	.970
4	20	491	.206
4	40	518	.438
4	60	540	.756
4	80	567	.926
5	20	504	.185
5	40	532	.543
5	60	555	.816
5	80	583	.957
6	20	512	.068
6	40	543	.256
6	60	568	.510
6	80	597	.776
7	20	518	.099
7	40	552	.298
7	60	579	.558
7	80	610	.810
8	20	527	.171
8	40	562	.444
8	60	591	.713
8	80	624	.918

Across grade levels, the probability of reaching *Proficient* or higher increases as the student's percentile rank or level increases. Students in grades 3 and 5 have a greater than 50% probability of meeting *Proficient* or higher at level 2, while in the other grades a higher level is needed to obtain greater than a 50% probability of reaching *Proficient*.

Student Growth Expectations for Regular Education, Special Education, and Students with Disabilities

Students enrolled in special education services (SPED) have growth patterns in ISIP Math that differ from students in general education (Non-SPED). To determine if ISIP Math can detect differences in growth for students who are receiving special education services, we used data from the 2018-2019 school year and computed the gain scores between the beginning of the year (BOY) and middle of the year (MOY), as well as the MOY and end of the year (EOY). We obtained data from school districts that provided special education status. If there was no status reported, the observations were deleted from the analysis, and only students with a score at all three data points were included. These differences were computed for all students, and then for those that have an identified disability in the Istation database. Table 7.9 shows mean scores and the gain scores between students who were identified as receiving special education versus general education students.

Table 7.9. *Mean and Gain Scores for Students in General and Special Education, 2018-2019 School Year*

Grade	Students	Sample	BOY Mean	MOY Mean	EOY Mean	BOY to MOY Gain	MOY to EOY Gain	BOY to EOY Gain
K	Non-SPED	7,038	301.66	368.04	425.42	66.38	57.37	123.75
K	SPED	803	273.31	324.89	377.37	51.58	52.48	104.06
1	Non-SPED	8,762	395.23	449.00	490.66	53.76	41.67	95.43
1	SPED	1,326	368.18	408.68	442.38	40.49	33.70	74.19
2	Non-SPED	8,444	455.59	480.69	494.10	25.10	13.41	38.52
2	SPED	1,310	435.82	453.36	464.82	17.54	11.46	29.00
3	Non-SPED	6,728	481.22	501.89	520.40	20.67	18.51	39.18
3	SPED	1,140	456.13	466.07	482.88	9.94	16.81	26.75
4	Non-SPED	6,298	507.30	526.03	543.78	18.73	17.75	36.48
4	SPED	1,079	485.83	495.17	507.89	9.34	12.72	22.06
5	Non-SPED	4,625	509.08	535.19	553.23	26.11	18.03	44.15
5	SPED	822	489.72	501.55	513.87	11.82	12.32	24.14
6	Non-SPED	1,657	541.27	563.88	575.96	22.61	12.08	34.69
6	SPED	245	497.81	499.06	513.52	1.25	14.46	15.71
7	Non-SPED	708	541.86	554.41	564.04	12.55	9.63	22.18
7	SPED	129	501.56	504.03	509.65	2.47	5.62	8.09
8	Non-SPED	643	547.17	570.75	571.94	23.57	1.19	24.77
8	SPED	92	504.58	512.99	521.04	8.41	8.05	16.45

Students in special education have gains across the year and by grade, and thus teachers, parents, and administrators can expect student growth in mathematics. Their gains are lower than general education students. In the lower grades of kindergarten through second grade, students in special education have scores at EOY that are like scores for students in general education at MOY. Starting in third grade, students enrolled in special education have scores at EOY that are like BOY scores for students in general education. In the middle school years, students enrolled in special education have scores that are lower than the BOY scores for students in general education.

Results by type of disability are shown in Table 7.10. Note that only results from a sample size of 30 or larger are reported in this study.

Table 7.10. *Mean and Gain Scores for Students in Special Education by Type of Disability, 2018-2019 School Year*

Grade	Disability Type	Sample	BOY Mean	MOY Mean	EOY Mean	BOY to MOY Gain	MOY to EOY Gain	BOY to EOY Gain
K	ID	35	245.93	242.20	263.99	-3.73	21.79	18.07
K	AU	53	259.76	337.06	401.77	77.30	64.72	142.01
K	DD	151	252.42	302.42	352.82	50.00	50.40	100.40
K	OHI	40	284.10	300.07	324.80	15.96	24.73	40.70
K	SI	508	278.51	331.25	384.25	52.74	53.00	105.74
K	SL	49	289.71	335.74	389.18	46.03	53.44	99.47
1	ID	49	328.50	334.49	341.66	5.99	7.18	13.16
1	LD & SLD	99	352.79	384.43	413.57	31.64	29.14	60.78
1	AU	67	349.92	367.87	413.85	17.95	45.98	63.93
1	DD	154	358.79	390.71	426.42	31.92	35.71	67.63
1	OHI	69	362.76	388.80	411.44	26.03	22.65	48.68
1	SI	688	382.23	423.05	459.73	40.82	36.68	77.50
1	SL	50	377.28	415.48	460.77	38.20	45.29	83.49
2	ID	64	409.98	418.21	427.80	8.23	9.59	17.82
2	LD & SLD	245	429.54	440.32	452.20	10.78	11.88	22.66
2	AU	75	434.29	444.09	463.06	9.80	18.96	28.77
2	DD	132	429.03	445.29	454.54	16.26	9.26	25.52
2	OHI	126	433.12	442.11	454.48	8.99	12.37	21.36
2	SI	472	445.65	467.10	480.19	21.46	13.09	34.54
3	ID	56	434.15	403.71	433.12	-30.44	29.41	-1.03
3	LD & SLD	313	445.40	452.49	467.49	7.09	15.00	22.09
3	AU	63	445.26	446.87	467.44	1.61	20.57	22.18
3	DD	69	448.96	453.95	462.70	4.98	8.75	13.74
3	OHI	120	444.45	443.00	458.42	-1.45	15.41	13.96
3	SI	238	468.31	481.14	499.11	12.83	17.98	30.80
4	ID	69	457.67	437.95	446.88	-19.72	8.93	-10.79
4	LD & SLD	377	473.92	480.93	490.12	7.01	9.19	16.20
4	AU	61	474.17	471.96	479.50	-2.21	7.54	5.33
4	OHI	140	467.03	473.58	480.14	6.55	6.56	13.11
4	SI	154	493.64	503.61	517.88	9.97	14.27	24.24
5	ID	34	471.95	460.51	462.36	-11.44	1.85	-9.59
5	LD & SLD	389	477.58	485.68	496.34	8.10	10.67	18.77
5	AU	52	479.33	497.26	512.77	17.92	15.52	33.44
5	OHI	112	479.61	480.36	495.07	0.75	14.72	15.46
5	SI	60	493.84	500.25	515.93	6.41	15.69	22.10
6	LD & SLD	162	491.23	500.70	510.87	9.47	10.17	19.64
6	OHI	53	488.11	485.75	489.67	-2.36	3.92	1.55
7	LD & SLD	100	505.00	506.51	516.79	1.51	10.28	11.78
7	OHI	30	505.49	509.12	496.08	3.63	-13.04	-9.41
8	LD & SLD	97	504.51	511.97	508.17	7.46	-3.79	3.66

ID – Intellectual Disabilities
LD, SLD – Learning Disabilities, Specific Learning Disabilities
AU – Autism Spectrum
OHI – Other Health Impairment
SI – Speech Impairment
SLI – Speech Language Impairment
DD – Developmental Delay

Students with intellectual disabilities (ID) typically have the lowest math scores and the least growth in mathematics, and in some grades these students have negative growth. Students with a speech impairment and those with autism have the most growth in math. For most students enrolled in special education, teachers can expect growth in math ability using the ISIP Math assessment. Special attention may need to be given to ID students, especially if they are showing negative growth. These students are capable of growth in mathematics but may need intensive intervention.

Conclusion

This chapter investigated the reliability and validity of ISIP Math. In all studies, correlations with other measures were moderate to strong, and the research also demonstrates that the assessment can be used to identify students at risk of not meeting grade-level expectations at the end of the year. ISIP Math also has utility with students in special education, and while these students may need intensive intervention, teachers can expect growth for these students, although the year-end growth may be less than for students who are not receiving special education.

Chapter 8: Growth

Introduction

Student achievement is typically evaluated in terms of a test score from a single test administration, whereas student growth can be conceptualized as change in academic performance over multiple administration periods. Monitoring growth can be used to gauge how well a student is performing relative to his or her peers. More specifically, growth may be used as a means to promote accountability, inform data-based decision-making, and foster partnerships within and between schools and districts. Monitoring individual student growth allows educators to determine whether students — and correspondingly teachers and schools — are making adequate annual progress toward state or national standards. As a result, monitoring student growth may improve student learning and inform decisions regarding classroom instruction and intervention (January et al., 2018; Jenkins et al., 2007; Pentimonti et al., 2017).

When educators think about student growth, there are certain questions they seek to answer, including:

- How much do my students need to grow to make a year's worth of progress?
- If my students start out in Tier 3, how many will grow into Tier 2 or Tier 1?
- How much do my students need to grow to maintain proficiency or to achieve more than a year's worth of growth?
- How are my students growing in comparison to other students? Is their growth faster or slower?

Istation provides three different approaches to view student growth across the school year to answer these questions. The first method is to view it as normative growth, which considers the growth a student needs to make to maintain the same percentile level. This method provides an answer to how much students need to grow to achieve a year's worth of progress. The second method is to view groups of students in a transition matrix. This method provides information based on expected changes in performance categories for a group of students throughout the school year. The third method we provide is based on performance pathways of growth. It is similar to student growth percentiles and attempts to answer the question regarding rates of growth and whether the student's growth is accelerating or decelerating in comparison to other students who started at the same level (Betebenner, 2011).

Expected Growth

Normative Growth by Decile at the Beginning of the Year

Istation's normative growth is based on information that allows us to evaluate the extent to which students' growth may be considered faster or slower than their academic peers with similar beginning-of-the-year (BOY) scores. By comparing how much growth a student has made relative to normed growth deciles, educators can make inferences about whether a student is making adequate progress or may need additional support or instruction. For example, if a student's growth on overall math exceeds the growth of 70% of their similarly scoring peers, this likely implies that the student is receiving adequate instruction. Students with scores in lower deciles may require additional support (<40th percentile).

BOY scale scores that were collected from the 2018-2019 normed sample were divided into 10 initial status groups for ISIP Math Overall Score. These groups indicate whether a student scored…

- at or below the 10th percentile,
- at or above the 11th percentile but below the 21st percentile,
- at or above the 21st percentile but below the 31st percentile,
- at or above the 31st percentile but below the 41st percentile,
- at or above the 41st percentile but below the 51st percentile,
- at or above the 51st percentile but below the 61st percentile,
- at or above the 61st percentile but below the 71st percentile,
- at or above the 71st percentile but below the 81st percentile,
- at or above the 81st percentile but below the 91st percentile, or
- at or above the 91st percentile.

After using percentile ranks to create decile categories for students' BOY scores, we calculated expected growth between BOY scores and end-of-the-year (EOY) scores for each decile. Tables 8.1 to 8.4 show the growth that would be expected in ISIP Math Overall scores by grade and decile. This information can be used to identify whether a student's growth may be considered faster or slower than their academic peers with similar BOY scores. In the elementary grades, for example, students starting out at a lower achievement level tend to demonstrate greater growth compared to students in upper grades.

Table 8.1: *Normative Growth for ISIP Math Overall for Grades Prekindergarten to 3, by Grade and Decile at the Beginning of the Year (September to April)*

BOY Percentile Rank	Decile	Pre-K Norm Growth	Kindergarten Norm Growth	1st Grade Norm Growth	2nd Grade Norm Growth	3rd Grade Norm Growth
1-10	1	26	22	19	17	17
11-20	2	46	33	28	24	25
21-30	3	58	44	35	29	31
31-40	4	64	52	41	35	35
41-50	5	61	61	47	39	40
51-60	6	56	70	52	43	44
61-70	7	54	80	58	47	49
71-80	8	61	91	65	52	54
81-90	9	89	106	74	60	61
91-99	10	140	138	94	75	76

Table 8.2: *Normative Growth for ISIP Math Overall for Grades Prekindergarten to 3, by Grade and Decile at the Beginning of the Year (September to May)*

BOY Percentile Rank	Decile	Pre-K Norm Growth	Kindergarten Norm Growth	1st Grade Norm Growth	2nd Grade Norm Growth	3rd Grade Norm Growth
1-10	1	29	24	20	18	18
11-20	2	52	37	29	24	25
21-30	3	67	49	36	30	32
31-40	4	73	58	42	35	36
41-50	5	70	68	48	39	41
51-60	6	64	78	53	44	45
61-70	7	62	89	59	48	49
71-80	8	70	101	66	53	55
81-90	9	102	118	75	61	62
91-99	10	160	153	95	76	77

Table 8.3: *Normative Growth for ISIP Math Overall for Grades 4 to 8, by Grade and Decile at the Beginning of the Year (September to April)*

BOY Percentile Rank	Decile	4th Grade Norm Growth	5th Grade Norm Growth	6th Grade Norm Growth	7th Grade Norm Growth	8th Grade Norm Growth
1-10	1	17	19	19	19	20
11-20	2	25	27	26	27	29
21-30	3	30	32	32	34	35
31-40	4	35	36	37	38	39
41-50	5	40	41	41	43	44
51-60	6	45	45	46	47	48
61-70	7	48	49	50	52	53
71-80	8	53	55	55	56	58
81-90	9	60	61	62	63	64
91-99	10	75	76	75	76	76

Table 8.4: *Normative Growth for ISIP Math Overall for Grades 4 to 8, by Grade and Decile at the Beginning of the Year (September to May)*

BOY Percentile Rank	Decile	4th Grade Norm Growth	5th Grade Norm Growth	6th Grade Norm Growth	7th Grade Norm Growth	8th Grade Norm Growth
1-10	1	18	19	20	19	21
11-20	2	25	27	27	28	29
21-30	3	31	32	33	34	35
31-40	4	36	37	37	39	40
41-50	5	40	41	42	43	45
51-60	6	45	46	47	48	49
61-70	7	49	50	51	53	54
71-80	8	54	56	56	57	59
81-90	9	61	62	63	64	65
91-99	10	76	77	76	77	77

Normative growth can inform several education-related activities. Educators can use these growth resources to evaluate students’ current achievement status. They may also use these resources to guide individualized instruction and to aid in setting achievement and growth objectives for students in a particular school. Normative growth provides an opportunity to support conversations about attainment patterns as educators can gage whether students made gains consistent with that of other students in the same grade with similar performance at the beginning of the year. This is useful because it provides the extent and magnitude by which a student's growth exceeded or fell short of the growth observed for other students with similar performance at the beginning of the year.

Transition Matrix Model

The transition matrix model depicts student growth in terms of movement in performance level categories rather than evaluating changes in scale score points throughout the academic school year (Castellano & Ho, 2013a). Istation uses a decile framework that expresses gains as the change in performance from the beginning of the year (September) to the end of the year (April). BOY and EOY scale scores that were collected from the 2018-2019 normed sample were divided into 10 initial status groups for the ISIP Math Overall Score. These groups indicate whether a student scored…

- below the 10th percentile,
- at or above the 10th percentile but below the 20th percentile,
- at or above the 20th percentile but below the 30th percentile,
- at or above the 30th percentile but below the 40th percentile,
- at or above the 40th percentile but below the 50th percentile,
- at or above the 50th percentile but below the 60th percentile,
- at or above the 60th percentile but below the 70th percentile,
- at or above the 70th percentile but below the 80th percentile,
- at or above the 80th percentile but below the 90th percentile, or
- at or above the 90th percentile.

After creating the groups, a transition matrix was computed to evaluate the movement in performance level categories from BOY to EOY for the ISIP Math Overall score for pre-K to grade 8. Initial analyses showed that the patterns were similar in grades 6 -8, and therefore we combined these grades into one sample for the transition matrix. In the tables below, the numeric values in the gray cells with an asterisk next to them reflect the percentage of students in the normed sample that maintained the same decile level category from BOY to EOY. The cells below the shaded values with an asterisk next to them correspond to cases in which a student regresses down one or more deciles in BOY and EOY. Similarly, the cells above the numeric values with an asterisk next to them represent growth or gaining one or more decile levels from BOY to EOY.

Tables 8.5 to 8.14 illustrate the change in performance categories from BOY to EOY for the ISIP Math Overall score. In general, students in the lower decile categories show growth by gaining a level or two between BOY and EOY. Students who placed in the upper decile categories mostly remain in the same category between BOY and EOY. There is more movement between levels for students who placed in the 30th to 79th decile categories in BOY. While some students remain in the initial decile category, there are also more balanced percentages of students who either gain a level or drop a level.

Table 8.5: *Pre-K Change in Performance Categories BOY-EOY for ISIP Math Overall by Decile Category*

BOY Decile Category	EOY 1-9	EOY 10-19	EOY 20-29	EOY 30-39	EOY 40-49	EOY 50-59	EOY 60-69	EOY 70-79	EOY 80-89	EOY 90-99
1-9	18.42*	15.79	10.53	10.53	5.26	5.26	5.26	13.16	13.16	2.63
10-19	15.91	13.64*	13.64	13.64	13.64	11.36	11.36	0	4.55	2.27
20-29	4.55	9.09	9.09*	22.73	18.18	4.55	9.09	9.09	9.09	4.55
30-39	16.22	13.51	5.41	2.7*	5.41	8.11	18.92	8.11	8.11	13.51
40-49	2.7	5.41	10.81	18.92	10.81*	16.22	10.81	10.81	8.11	5.41
50-59	20	6.67	10	3.33	10	10*	6.67	16.67	13.33	3.33
60-69	2.63	5.26	13.16	7.89	15.79	15.79	5.26*	5.26	13.16	15.79
70-79	0	12.5	12.5	4.17	4.17	4.17	12.5	16.67*	16.67	16.67
80-89	10	12.5	5	5	12.5	15	5	10	10*	15
90-99	7.69	5.13	12.82	5.13	2.56	15.38	10.26	12.82	10.26	17.95*

Table 8.6: *Kindergarten Change in Performance Categories BOY-EOY for ISIP Math Overall*

BOY Decile Category	EOY 1-9	EOY 10-19	EOY 20-29	EOY 30-39	EOY 40-49	EOY 50-59	EOY 60-69	EOY 70-79	EOY 80-89	EOY 90-99
1-9	15.03*	11.9	11.48	10.23	7.72	9.6	10.44	10.65	7.52	5.43
10-19	21.57	22.22*	15.03	10.46	7.63	7.63	3.7	3.92	2.4	5.45
20-29	12.88	14.05	16.63*	15.93	9.6	10.07	8.9	5.39	4.68	1.87
30-39	9.57	14.11	11.72	13.4*	12.68	11.48	11	7.18	6.7	2.15
40-49	8.78	8.29	9.51	8.05	12.2*	11.71	11.95	14.15	9.76	5.61
50-59	6.65	6.39	7.16	7.16	14.07	12.28*	13.3	9.97	13.81	9.21
60-69	7.26	5.87	6.15	11.17	6.98	12.57	11.73*	14.8	15.08	8.38
70-79	4.41	5.08	6.1	8.47	11.19	8.47	14.58	12.88*	14.24	14.58
80-89	2.9	3.23	6.13	6.13	7.1	9.68	10	13.55	16.13*	25.16
90-99	2.14	2.14	4.29	7.5	8.21	4.64	6.43	12.86	15	36.79*

Table 8.7: *First Grade Change in Performance Categories BOY-EOY for ISIP Math Overall*

BOY Decile Category	EOY 1-9	EOY 10-19	EOY 20-29	EOY 30-39	EOY 40-49	EOY 50-59	EOY 60-69	EOY 70-79	EOY 80-89	EOY 90-99
1-9	28.52*	17.33	14.98	12.27	7.76	7.04	3.61	3.25	3.25	1.99
10-19	20.48	18.88*	14.66	11.45	11.04	4.62	4.02	5.62	4.02	5.22
20-29	12.5	19.07	11.65*	15.68	14.19	8.9	7.2	3.6	4.03	3.18
30-39	9.09	11.31	13.3	13.75*	13.3	12.42	11.09	7.1	4.66	3.99
40-49	6.18	7.13	12.11	12.59	12.59*	13.06	13.3	12.59	8.08	2.38
50-59	4.48	6.5	7.17	9.19	13.45	15.02*	12.56	15.02	8.74	7.85
60-69	3.29	4.46	6.34	5.63	11.5	15.02	14.08*	14.08	17.37	8.22
70-79	2.65	3.45	7.16	5.31	7.96	8.22	14.06	18.04*	17.24	15.92
80-89	1.47	2.64	5.28	5.57	4.99	8.8	12.02	13.2	20.82*	25.22
90-99	1.1	2.48	1.93	4.13	5.23	7.16	9.37	13.5	18.18	36.91*

Table 8.8: *Second Grade Change in Performance Categories BOY-EOY for ISIP Math Overall*

BOY Decile Category	EOY 1-9	EOY 10-19	EOY 20-29	EOY 30-39	EOY 40-49	EOY 50-59	EOY 60-69	EOY 70-79	EOY 80-89	EOY 90-99
1-9	37.41*	23.02	14.39	10.25	5.4	3.96	1.44	2.16	0.9	1.08
10-19	15.61	16.4*	20.75	14.43	11.86	6.32	5.73	2.37	3.95	2.57
20-29	10.43	12.27	14.52*	15.75	15.13	12.27	7.57	3.48	4.09	4.5
30-39	6.43	10.79	11	15.56*	16.6	14.32	11.41	5.81	3.94	4.15
40-49	3.89	9	9	10.63	12.07*	13.29	14.72	14.52	8.59	4.29
50-59	3.73	6.47	7.71	5.22	12.19	14.18*	14.68	13.18	13.43	9.2
60-69	1.76	4.69	6.74	6.45	7.33	13.49	16.72*	17.89	17.3	7.62
70-79	2.91	4.85	6.31	6.8	5.58	6.55	11.65	19.17*	20.15	16.02
80-89	2.52	2.52	4.2	5.32	7	7.56	9.24	18.77	19.61*	23.25
90-99	3.63	2.42	2.42	3.63	3.02	6.04	8.46	11.48	20.54	38.37*

Table 8.9: *Third Grade Change in Performance Categories BOY-EOY for ISIP Math Overall*

BOY Decile	EOY	EOY	EOY	EOY	EOY	EOY	EOY	EOY	EOY	EOY
Category	1-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
1-9	43.82*	21.04	16.41	7.92	4.05	3.47	1.54	0.77	0.39	0.58
10-19	16.93	21.65*	16.14	16.54	11.42	8.86	3.54	1.97	1.38	1.57
20-29	9.73	15.16	15.61*	15.84	16.06	11.31	7.01	5.43	2.71	1.13
30-39	5.2	10.81	12.47	13.93*	14.14	16.63	10.81	9.15	4.16	2.7
40-49	5.2	6.19	6.93	12.38	11.63*	16.34	14.85	15.35	7.43	3.71
50-59	3.1	5.49	9.79	9.79	12.17	13.6*	16.23	15.51	10.26	4.06
60-69	2.06	4.37	4.88	8.23	10.03	13.62	13.37*	13.88	16.71	12.85
70-79	0.84	3.9	2.79	6.41	6.41	8.64	15.88	16.16*	22.84	16.16
80-89	0.57	1.42	3.12	5.1	6.52	6.8	10.76	15.3	23.8*	26.63
90-99	1.12	1.96	1.96	3.36	3.36	7.56	8.68	11.48	17.09	43.42*

Table 8.10: *Fourth Grade Change in Performance Categories BOY-EOY for ISIP Math Overall*

BOY Decile	EOY	EOY	EOY	EOY	EOY	EOY	EOY	EOY	EOY	EOY
Category	1-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
1-9	51.35*	23.51	12.7	6.22	2.43	0.81	0.81	1.08	0.81	0.27
10-19	18.36	22.03*	24.01	13.56	10.45	6.5	1.98	1.98	0.28	0.85
20-29	9.2	18.1	18.68*	17.24	10.06	12.36	6.9	3.45	2.3	1.72
30-39	2.64	9.57	8.91	18.81*	17.49	16.83	9.9	8.25	4.29	3.3
40-49	3.19	6.96	6.67	12.75	17.1*	16.81	14.78	10.43	7.83	3.48
50-59	1.17	2.73	7.81	9.38	14.84	16.02*	14.06	15.23	10.94	7.81
60-69	0.97	3.57	6.17	7.79	12.34	10.71	20.45*	15.91	14.94	7.14
70-79	1.02	0.68	1.71	5.12	6.83	11.95	16.04	18.43*	21.16	17.06
80-89	1.48	2.21	2.21	3.32	5.9	8.86	10.33	18.45	21.77*	25.46
90-99	0	0.43	2.6	1.73	2.6	3.9	8.23	9.52	24.68	46.32*

Table 8.11: *Fifth Grade Change in Performance Categories BOY-EOY for ISIP Math Overall*

BOY Decile Category	EOY 1-9	EOY 10-19	EOY 20-29	EOY 30-39	EOY 40-49	EOY 50-59	EOY 60-69	EOY 70-79	EOY 80-89	EOY 90-99
1-9	48.31*	24.58	13.28	6.5	3.67	1.41	0.85	0.56	0.28	0.56
10-19	14.64	27.41*	22.12	13.08	9.97	4.98	4.36	1.56	1.25	0.62
20-29	10.04	13.38	18.59*	18.96	18.22	7.06	8.18	4.09	0.74	0.74
30-39	4.47	8.59	11.68	12.37*	17.18	15.81	15.46	5.84	5.5	3.09
40-49	2.64	6.04	13.96	12.83	9.81*	17.74	15.85	9.06	7.17	4.91
50-59	0.69	2.78	5.9	10.42	14.24	18.4*	14.58	15.63	7.29	10.07
60-69	0	1.94	3.4	8.25	11.65	13.59	13.59*	17.96	16.02	13.59
70-79	0.43	0.43	3.03	4.76	9.96	9.52	12.99	17.32*	20.35	21.21
80-89	0	0	1.85	3.24	3.7	6.94	9.72	19.44	28.24*	26.85
90-99	0	0	1.02	1.02	1.02	4.57	7.61	17.77	30.46	36.55*

Table 8.15: *Six through Eighth Grade Combined Change in Performance Categories BOY-EOY for ISIP Math Overall*

BOY Decile Category	EOY 1-9	EOY 10-19	EOY 20-29	EOY 30-39	EOY 40-49	EOY 50-59	EOY 60-69	EOY 70-79	EOY 80-89	EOY 90-99
1-9	54.24*	16.95	15.25	10.17	1.69	1.69	0	0	0	0
10-19	20.31	28.13*	12.5	4.69	10.94	10.94	6.25	3.13	3.13	0
20-29	15.09	15.09	18.87*	5.66	11.32	11.32	5.66	5.66	5.66	5.66
30-39	8.62	10.34	18.97	20.69*	10.34	12.07	6.9	6.9	1.72	3.45
40-49	1.67	8.33	10	20	16.67*	10	5	10	13.33	5
50-59	1.59	7.94	11.11	19.05	9.52	11.11*	14.29	6.35	12.7	6.35
60-69	0	7.41	5.56	5.56	9.26	14.81	14.81*	18.52	11.11	12.96
70-79	0	0	0	6.9	17.24	13.79	17.24	15.52*	13.79	15.52
80-89	0	0	4.84	6.45	6.45	8.06	14.52	22.58	20.97*	16.13
90-99	0	1.85	1.85	0	5.56	7.41	16.67	11.11	24.07	31.48*

This information may be useful to educators as it illustrates what they can expect at the class or school level in terms of movement throughout performance levels from BOY to EOY. More specifically, the transition matrix provides an insight into the percentage of students on track to maintain or reach proficiency. It should be noted that a change in categories can be associated with a wide range of actual gains depending on the student's standing within the category regions.

Transitions through past categories can also support predictions about a student's future category location under the assumption that transitions across categories will resume in a linear pattern over time. For example, if a student improves one decile level between BOY and EOY in grade 3, it might be reasonable to assume that the student will improve one or more decile categories in grade 4. In this hypothetical context, the transition matrix functions as a coarse trajectory model, where an increase in one decile category is extrapolated and assumed to resume to future time points.

Another useful feature of the transition matrix is that average values for groups of students are interpretable as a type of average growth. For example, the matrix cells correspond to the number of decile categories a student has gained or lost; thus the average over all students is the average gain in decile categories for that particular group.

Expected Growth Pathways

Expected growth pathways are another feature that allows educators to monitor and compare the overall math skill development of their students over the course of the school year to the growth of a nationally representative sample of students with similar achievement at BOY. Expected growth pathways may be used to set growth targets and monitor student progress. By comparing how much a student has gained relative to normed growth pathways, educators can make inferences about whether a student is making adequate progress.

A nationally representative 2018-2019 normed sample was used for students in pre-K through grade 8. BOY ISIP Math Overall scores were placed into five BOY status groups. These BOY groups are linked to Istation's instructional levels, which are set to identify students at risk for developing reading deficiencies.

These instructional levels indicate whether a student at the beginning of the year scored…

- at or below the 20th percentile,
- at or above the 21st percentile but below the 41st percentile,
- at or above the 41st percentile but below the 61st percentile,
- at or above the 61st percentile but below the 81st percentile, or
- at or above the 81st percentile.

After assigning BOY scores to BOY status groups, a gain score was computed for each student by subtracting the BOY overall reading score from the EOY overall reading score. The resulting gain scores were used to create percentile gains by dividing gain scores into quantiles within each BOY status group. Higher percentile gains indicate that the student showed more growth relative to other students in the same BOY status group. Labels were then assigned to expected growth pathways within each BOY status group where a gain score falling between the 41st and 60th percentiles can be classified as falling within the typical growth pathway. Similarly, scores that fall between the 61st and 80th percentiles can be classified as above typical, whereas scores above the 80th percentile can be classified as accelerated. Table 8.12 summarizes the growth descriptions.

Table 8.12. *Pathway Growth Descriptions*

Pathways	Percentile Range	Growth Descriptor
1	≤40th	Below Typical
2	41st - 60th	Typical
3	61st - 80th	Above Typical
4	>80th	Accelerated

Expected growth pathways provide a metric that accounts for differing patterns of growth across grades and BOY ability level. Table 8.13 illustrates these expected growth pathways within each BOY instructional group for ISIP Math Overall scores. Similar to the transition matrix, we combined grades 6 – 8 because growth pathways were similar for these grades.

One intuitive finding is that students starting out in a lower BOY instructional group are expected to demonstrate greater growth than students who are already in a higher BOY instructional group within the same grade. Similarly, expected growth is greater for students in the elementary grades compared to students in upper grades. Additional analyses were conducted to examine the impact of prescribed growth goals and student location at the EOY. In general, students who were in level 1 or level 2 in the BOY status group moved up two levels by setting an accelerated target. Setting an above-typical target usually results in moving up one level, whereas a typical target usually results in staying within the same level. These findings are particularly consistent in the early grades where students have much more room to improve their skill sets.

Table 8.13: *Expected Growth Pathways (Gains) for ISIP Math Overall BOY-EOY by ISIP Instructional Levels*

BOY Status Group Percentile	Growth Pathway	Pre-K	K	Grade 1	Grade 2	Grade 3	Grade 4	Grade 5	Grade 6-8
<21st	Below Typical	<150	<132	<104	<38	<40	<25	<20	<14
	Typical	150-185	132-192	104-134	38-57	40-56	25-41	20-37	14-32
	Above Typical	186-246	193-247	135-180	58-92	57-86	42-67	38-62	33-62
	Accelerated	≥247	≥248	≥181	≥93	≥87	≥68	≥63	≥63
21st - 40th	Below Typical	<165	<127	<95	<32	<31	<23	<29	<15
	Typical	165-192	127-178	95-123	32-52	31-45	23-38	29-48	15-32
	Above Typical	197-226	179-222	124-158	53-81	46-67	39-61	49-70	33-59
	Accelerated	≥227	≥223	≥159	≥82	≥68	≥62	≥71	≥60
41st - 60th	Below Typical	<129	<122	<98	<34	<31	<23	<35	<6
	Typical	129-164	122-158	98-125	34-56	31-45	23-38	35-57	6-32
	Above Typical	165-214	159-199	126-150	57-83	46-68	39-58	58-78	33-55
	Accelerated	≥215	≥200	≥151	≥84	≥69	≥59	≥79	≥56
61st - 80th	Below Typical	<125	<111	<87	<31	<33	<31	<41	<9
	Typical	125-138	111-145	87-111	31-50	33-50	31-48	41-56	9-33
	Above Typical	139-184	146-190	112-141	51-75	51-73	49-70	57-77	34-58
	Accelerated	≥185	≥191	≥142	≥76	≥74	≥71	≥78	≥59
>80th	Below Typical	<64	<31	<50	<20	<40	<37	<36	<16
	Typical	64-96	31-88	50-83	20-42	40-60	37-56	36-55	16-43
	Above Typical	97-133	89-134	84-116	43-73	61-83	57-82	56-77	44-72
	Accelerated	≥137	≥135	≥117	≥74	≥84	≥83	≥78	≥73

Expected growth pathways can inform decisions about instruction and intervention by providing normative information regarding growth, which may be particularly useful in schools that implement multi-tiered systems of support. Educators can use this type of growth information to evaluate the extent to which the instructional approach is working or whether modifications are necessary to meet students' needs. Data based on a one-time assessment do not support this type of decision-making because these data refer to students' status rather than their growth. Expected growth pathways can be used to identify how quickly students are growing even if they are not on track to meet predefined criteria such as criterion related standards.

Pathways of growth promote inferences that account for students' initial status, which is key to interpreting growth, since growth is often related to BOY performance but not necessarily in an intuative manner. When comparing a given pathway of growth (e.g., *Typical*) across BOY instructional levels, students with the highest BOY scores (i.e., those in level 5) tend to improve less over the course of the year than students in level 1 at the beginning of the year.

References

Baker, F. B., & Kim, S.-H. (Eds.). (2004). *Item response theory: Parameter estimation techniques* (2nd edition). CRC Press.

Betebenner, D.W. (2011). New directions in student growth: The Colorado growth model. Paper presented at the National Conference on Student Assessment, Orlando, FL, June 19, 2011. Retrieved March 29, 2012, from http://ccsso.confex.com/ccsso/2011/webprogram/ Session2199.html.

Blisle, P. (2017). *Beta.parms.from.quantiles* (1.3) [Computer software]. Division of Clinical Epidemiology, McGill University Health Science Center.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51.

Carlson, J. E. (2011). Statistical methods for vertical linking. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling, and linking*. Springer.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective* (2nd edition). Chapman and Hall/CRC.

Castellano, K. E. & Ho, A. D. (2013a). A practitioner's guide to growth models. A paper commissioned by the Technical Issues in Large-Scale Assessment (TILSA) and Accountability Systems & Reporting (ASR) State Collaboratives on Assessment and Student Standards, Council of Chief State School Officers

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.

Cook, J. R., & Stefanski, L. A. (1993). *Simulation-extrapolation estimation in parametric measurement error models* (No. 2224R; Institute of Statistics Mimeograph Series, p. 31). North Carolina State University.

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

DeMaris, A. (1992). *Logit modeling: Practical applications*. SAGE Publications, Inc.

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

Engec, N. (1998). *Logistic regression and item response theory: Estimation item and ability parameters by using logistic regression in IRT.* Louisiana State University.

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. (2007). 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., & 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.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

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

Hardin, J. W., Schmiediche, H., & Carroll, R. J. (2003). The simulation extrapolation method for fitting generalized linear models with additive measurement error. *The Stata Journal*, *3*(4), 373–385.

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.

Hill, S., Ketterlin-Geller, L. R., & Gifford, D. B. (2012). Imagination Station (Istation): Universal Screener Instrument Development for Grade 3 (Tech. Rep. No. 12-04). Dallas, TX: Southern Methodist University, Research in Mathematics Education.

Hosmer, Jr., D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd edition). Wiley.

Huynh, H., & Meyer, P. (2010). Use of robust z in detecting unstable items in item response theory models. *Practical Assessment, Research, and Evaluation, 15*(Article 2). https://doi.org/10.7275/YCX6-E864

Istation. (2018). *Istation's Indicators of Progress (ISIP) math technical report*. Author.

Jagers, P. (1986). Post-Stratification against Bias in Sampling. *International Statistical Review / Revue Internationale de Statistique, 54*(2), 159-167. doi:10.2307/1403141

January, S.A. A., Van Norman, E. R., Christ, T. J., Ardoin, S. P., Eckert, T. L., & White, M. J. (2018). Progress monitoring in reading: Comparison of weekly, bimonthly, and monthly assessments for students at risk for reading difficulties in Grades 2–4. School Psychology Review, 47(1), 83–94. https://doi.org/10.17105/SPR-2017-0009.V47-1

Jenkins, J. R., Hudson, R. F., & Johnson, E. S. (2007). Screening for at-risk readers in a response to intervention framework. School Psychology Review, 37(4), 582–600.

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

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155–186). American Council on Education and Praeger.

Kolen, M. J. (2007). Data collection designs and linking procedures. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 31–56). Springer.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd edition). Springer.

Lederer, W., & Küchenhoff, H. (2006). A short intro to the SIMEX and MC-SIMEX .pdf. *R News*, *6*(4).

Lenhard, A., Lenhard, W., Suggate, S., & Segerer, R. (2018). A continuous solution to the norming problem. *Assessment*, *25*(1), 112–125. https://doi.org/10.1177/1073191116656437

Lenhard, W., & Seibold, H. (2019). *simex: SIMEX- And MCSIMEX-algorithm for measurement error models* (R package version 1.8) [Computer software]. https://CRAN.R-project.org/package=simex

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.

Nitko, A. J., & Brookhart, S. M. (2011). *Educational assessment of students* (6th ed.). Pearson.

Oehlert, G. W. (1992). A note on the delta method. *The American Statistician*, *46*(1), 27–29. https://doi.org/10.1080/00031305.1992.10475842.

Ohio Department of Education (2019). School and district results 2018-2019. Retrieved from https://reportcardstorage.education.ohio.gov/search/State_Report_Card.pdf

Patarapichayatham, C., Kamata, A. & Kanjanawasee, S. (2012). Evaluation of model selection strategies for cross-level two-way differential item functioning analysis. *Education and Psychological Measurement, 72*, 1, 44-51.

Patarapichayatham, C., & Locke, V. (2020a). Linking the ACT Aspire Assessments to ISIP Reading and Math. www.istation.com/studies.

Patarapichayatham, C., & Locke, V. (2020b). Linking the Ohio AIR to ISIP. www.istation.com/studies.

Patz, R. J., & Yao, L. (2007). Vertical scaling: Statistical models for measuring growth and achievement. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics* (Vol. 26). Elsevier.

Pentimonti, J. M., Walker, M. . A., & Zumeta, R. E. (2017). The selection and use of screening and progress monitoring tools in data-based decision making within an MTSS framework. Perspectives on Language and Literacy, 43(3), 34–40.

R Core Team. (2018). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing.

Shang, Y. (2012). Measurement Error Adjustment Using the SIMEX Method: An Application to Student Growth Percentiles: Measurement Error Adjustment Using the SIMEX Method. *Journal of Educational Measurement*, *49*(4), 446–465. https://doi.org/10.1111/j.1745-3984.2012.00186.x

Shang, Y., VanIwaarden, A., & Betebenner, D. W. (2015). Covariate Measurement Error Correction for Student Growth Percentiles Using the SIMEX Method. *Educational Measurement: Issues and Practice*, *34*(1), 4–14. https://doi.org/10.1111/emip.12058

Shaw, P., & Keogh, R. (2017, August). *Understanding and tackling measurement error: SIMEX* [STRATOS Initiative Workshop, Understanding and tackling measurement error: A whistle stop tour of modern practical methods.]. *Education for Statistics in Practice,* CEN-IBS, Vienna, Austria.

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.

Stefanski, L. A., & Cook, J. R. (1995). Simulation-Extrapolation: The Measurement Error Jackknife. *Journal of the American Statistical Association*, *90*(432), 1247–1256. https://doi.org/10.2307/2291515

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement*, *40*(4), 331–352. https://doi.org/10.1111/j.1745-3984.2003.tb01150.x

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

Thum, Y. M., & Hauser, C. H. (2015). *NWEA 2015 MAP Norms for Student and School Achievement Status and Growth.*

Tong, Y., & Kolen, M. J. (2010). Scaling. *Educational Measurement: Issues and Practice*, *29*(4), 39–48. https://doi.org/10.1111/j.1745-3992.2010.00192.x

Young, M. J., & Tong, Y. (2015). Vertical scaling. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd edition, pp. 450–456). Routledge.

Voncken, L., Albers, C. J., & Timmerman, M. E. (2019). Model selection in continuous test norming with GAMLSS. *Assessment*, *26*(7), 1329–1346. https://doi.org/10.1177/1073191117715113

Zhu, J., & Chen, H.-Y. (2011). Utility of inferential norming with smaller sample sizes. *Journal of Psychoeducational Assessment*, *29*(6), 570–580. https://doi.org/10.1177/0734282910396323