ISIP Reading
Technical Report

Patricia Mathes, Ph.D.
Joseph Torgesen, Ph.D.
Jeannine Herron, Ph.D.

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

Istation's Indicators of Progress (ISIP)™
Reading Technical Report 2023

Patricia Mathes, Ph.D.
Joseph Torgesen, Ph.D.
Jeannine Herron, Ph.D.

Imagination Station – Istation

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

Table of Contents

Table of Tables

Table of Figures

Acknowledgements

The Istation's Indicators of Progress - Reading (ISIP-Reading) assessments were first developed under the guidance of Patricia Mathes, Ph.D., Jeanine Herron, Ph.D., and Joseph Torgesen, Ph.D., drawing on their decades of research experience in teaching and assessing the skills of reading, from foundational to more complex skills. ISIP Early Reading focused on foundational skills based on the science of reading from the National Reading Panel, and ISIP Advanced Reading built on this work and included more complexity in vocabulary, spelling, fluency, and reading comprehension. We are deeply indebted to these test authors for their work and innovation in creating a computer-adaptive assessment for progress monitoring and screening purposes.

In this edition, our goal was to bring together the ISIP Early Reading and ISIP Advanced Reading into one common scale. Therefore, this technical manual builds on the work from the ISIP Early Reading and ISIP Advanced Reading technical manuals.

There are several other differences between the prior editions of ISIP Reading and this edition. First, we added a vertical scale so that student progress can be better tracked over time. For this effort, we worked with Michael J. Young, Ph.D. He has decades of experience in vertical scaling, equating, and norming, and without his efforts the vertical scaling would not have been possible. He also wrote the chapter on vertical scaling.

With the new vertical scale, we updated our norms and norming procedures. Using stratified data sets created by the research team including psychometrician Chalie Patarapichayatham, Ph.D., and statistical analyst Sean Lewis, M.A., Dr. Young incorporated the most recent norming procedures for ISIP Reading. He advised us throughout the process on growth norming, and helped us determine which cohort would be the best to use for norming purposes, given that this norms update occurred while the COVID 19 pandemic was still present in the United States. He also wrote part of the technical manual that pertained to the norming.

Next, we offer solutions for growth reporting and growth assessment. We formed a working group within Istation consisting of Robert Rubin, Vickie Whitfield, Brian Duggan, Peter Jacobson and Gary White, who helped us with a solution for growth reporting. Peter Jacobson and Alicia Pruitt also assisted with the focus groups we held to determine the best path forward for growth reporting and determine the appropriate cohort for norming. Psychometrician Raffaela Wolf, Ph.D. came up with the frameworks for monitoring student growth and wrote the chapter in this manual on student growth. We are indebted to our focus group members for their time and expertise, it has resulted in a better solution for growth norming.

We have also updated our validity information, and we thank the school districts and other research partners that shared the data with us. The data were analyzed by Raffaela Wolf and Chalie Patarapichaytham who also wrote the sections for the differential item functioning, item drift, and special group studies.

The ISIP Reading is used across the country to help educators identify students at risk of reading failure, especially due to dyslexia. ISIP Reading can help detect the risk of dyslexia, and we have included a section on using ISIP Reading as a dyslexia screening tool. Tracy Larson, PsyD, helped us collect the validity data for ISIP as a dyslexia screening tool, and we thank her and the students and parents for their participation in our research.

Product managers Bonnie Martinez, Ph.D., and Peter Jacobson, MBA, worked on the final implementation of the new norms and scales and coordinated efforts to make sure that the product was available on time with no errors. Copyeditors Moniqa Paullet and Ross Frazier read the technical manual and all other documentation and made valuable suggestions that enhanced the clarity of the writing.

Working under the direction of Dr. Bill Fahle and Gary White, the digital test-development team consisted of Charles Middleton, David Smith, Gabriel Gallagher, Haley Aycock, Bryan Davis, and Matthew Kreideweis. The Quality Assurance team, led by David Pearson, helped check the reports, systems, and the norms. We are especially grateful to Kris Mackay, Greg Gibson, Jonathan Welch, and Subha Natarahan for their work in making sure that the reports and normative information are accurate. The

reporting team, led by Zachary Terry, implemented the new scale and norms flawlessly. Web engineers Christopher Corns, Harry Shoemaker, Mark Shyn, Amelia Moore, and Lindsay Padian worked tirelessly to ensure that all of the applications functioned well with the new scaling and reporting norms. The systems team, led by John Jeffus, made sure that the system continued to flow flawlessly throughout this major change to the assessment. Gary White and his team of data engineers helped us acquire the data we needed for analyses in chapter 5 for reporting item DIF and drift.

Our customer success team, led by Senior Vice President Christy Spivey and Director Alex Bardales answered customer questions and led the communication efforts so that Istation customers were aware of the changes. We are especially grateful to Tina Cole and Alicia Pruitt who helped with webinars and articles regarding the renorming. The product knowledge team, led by Vice President Vickie Whitfield, also worked with customer communication efforts.

We are also grateful to the Istation leadership team including Senior Vice President Chris Blevins, Vice President Vickie Whitfield, Chief Information Officer Robert Rubin, Chief Technology Officer Bill Lowrey, Chief Financial Officer Monika Flood, and Chief Product Officer Steve Jordan. Their support and guidance throughout the process made the job much easier.

Our former president and Chief Operating Officer Ossa Fisher and our Chief Executive Officer and Chairman Richard Collins provided us support, feedback, and encouragement. Their passion and enthusiasm for improving lives through educational technology is an inspiration.

Victoria Locke, Ph.D.

Vice President Research and Assessment

Chapter 1: Introduction and Test Development

Istation's Indicators of Progress (ISIP™) Reading is a computer-adaptative test (CAT) that provides continuous progress monitoring (CPM) as well as sophisticated reporting that gives detailed information to teachers, parents, and students in the critical domains of reading throughout the academic years.

ISIP Reading was built on decades of research on how to best teach and assess reading skills, building on the work from our authoring team of renowned literacy experts Joseph K. Torgesen, PhD; Patricia G. Mathes, PhD; and Jeannine Herron, PhD. The early reading assessment (ISIP ER) was designed for students in prekindergarten through grade 3 (Mathes, et al., 2016), and the advanced reading assessment (ISIP AR) was designed for students in grades 4 through 8 (Mathes, 2016). ISIP Reading is now combined into one assessment with a continuous vertical scale.

ISIP Reading: Goals for the Assessment Update

We had several goals with this revision of the ISIP Reading assessment. First, we wanted to revise the scaling of the assessment. Previously, the ISIP ER and ISIP AR were on separate scales. We linked the scales into a common vertical scale, which is described in chapter 2.

Second, we wanted to update our norms and sampling procedures. The normative sample was composed from millions of eligible students in our Istation database and we randomly selected a representative sample using a school-level socioeconomic index. We updated the norms using a continuous polynomial norming procedure, which is described in chapter 3.

Third, we wanted to improve the information available to teachers about students' different strengths in letter knowledge. This subtest is composed of letter sounds and letter recognition, and we composed separate scores for these two skills within the Letter Knowledge subtest. Those changes are described later in this chapter.

Fourth, we wanted to provide information on student growth. We offer three different lenses through which to view student growth: normative growth, a transition matrix, and expected growth pathways. These are described in detail in chapter 4.

We have also updated our reliability and validity information, which is described in chapter 5. It includes information on state summative assessments as well as other assessments.

The following sections in this chapter describe the purpose of ISIP Reading, a description of a computer-adaptive assessment, and a description of the ISIP subtests and what they measure. The purpose of this manual is to give updated technical information regarding the new features of ISIP Reading including vertical scaling, norming, growth norms, and updated reliability and validity. Complete details on the development of ISIP Reading, including the theory and process used in the assessment, are available in these technical manuals:

Mathes, P., Torgesen, J. & Herron, J. (2016). *Istation's Indicators of Progress (ISIP) Early Reading Technical Report: Computer Adaptive Testing System for continuous Progress Monitoring of Reading Growth for Students Pre-K through Grade 3*. Dallas, TX: Istation.

Mathes, P. (2016). *Istation's Indicators of Progress (ISIP) Advanced Reading Technical Report: Computer Adaptive Testing System for Continuous Progress Monitoring of Reading Growth for Students Grade 4 through Grade 8*. Dallas, TX: Istation.

Istation (2020). *Istation's Indicators of Progress (ISIP) Oral Reading Fluency Technical Report*. Dallas, TX: Istation.

Istation (2022a). *Istation's Indicators of Progress (ISIP) Rapid Auto Naming (ISIP RAN) Technical Report*. Dallas, TX: Istation.

Purpose of ISIP Reading

ISIP Reading provides teachers and other school personnel with easy-to-interpret reports that detail student strengths and deficits and provide links to teaching resources. Use of this data helps teachers make informed decisions regarding each student's response to targeted reading instruction and intervention strategies. Istation provides immediate results and insightful reports that help teachers plan instruction. The

Istation system also provides links to powerful, research-backed lessons teachers can use to differentiate instruction (Mathes et al., 2016).

ISIP Reading for prekindergarten through third grades provides growth information in the five critical domains of early reading, established by the National Reading Panel (2000): phonemic awareness, alphabetic knowledge and skills, fluency, vocabulary, and comprehension. For younger students, ISIP Reading is designed to identify children at risk for reading difficulties, provide continuous progress monitoring of skills that are predictors of reading success, and provide immediate reporting of assessment data on student learning needs, which facilitates differentiated instruction (Mathes et al., 2016).

For grades four to eight, ISIP Reading provides growth information for word analysis (spelling), text fluency, vocabulary, and comprehension. At these grades, when students are reading to learn, ISIP Reading is designed to identify specific reading needs of the older struggling reader. It provides continuous progress monitoring, and helps teachers differentiate instruction in upper elementary and middle school (Mathes, 2016).

ISIP Reading provides continuous measurement of reading skills as determined by the teacher or school district. Younger readers are often assessed monthly to monitor their progress to reading mastery. Older students may be assessed three months a year as a benchmark, or more often as a progress monitor.

The entire assessment battery requires approximately 30 minutes. Classroom and individual student results are available in real time, illustrating each student's past and present performance and skill growth. 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 (Mathes et al., 2016; Mathes, 2016).

Computer-Adaptive Testing

Recent advances in CAT mean that the assessment can adjust to the actual ability of each child. CAT replaces the need to create parallel forms. Assessments built on CAT tailor the assessment to match the performance abilities of individual students. Students who are achieving significantly above or below grade expectations can be accurately assessed to reflect their true abilities.

The ISIP Reading assessment also uses a Bayesian estimator in the CAT algorithm. The first time a student takes the assessment in an academic year, they receive an item that is at a median level of difficulty for their grade. If the student answers correctly, they receive a slightly more difficult item, and if they answer incorrectly, they receive a less difficult item, as depicted in Figure 1.1. The computer adapts with each item until it reaches reliability for the subtest. The next time the student takes the assessment, they start at an appropriate level of difficulty based on their ability score from the previous administration, essentially picking up where they left off. Thus, the CAT algorithm adapts within and across assessments to better estimate a student's reading ability.

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 item parameters for discrimination and difficulty have been determined, the CAT algorithm selects items based on each student's performance, selecting easier items if previous items are missed and harder items if the student answers correctly. Through this process of selecting items based on student performance, the algorithm 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 (Mathes et al., 2016; Mathes, 2016).

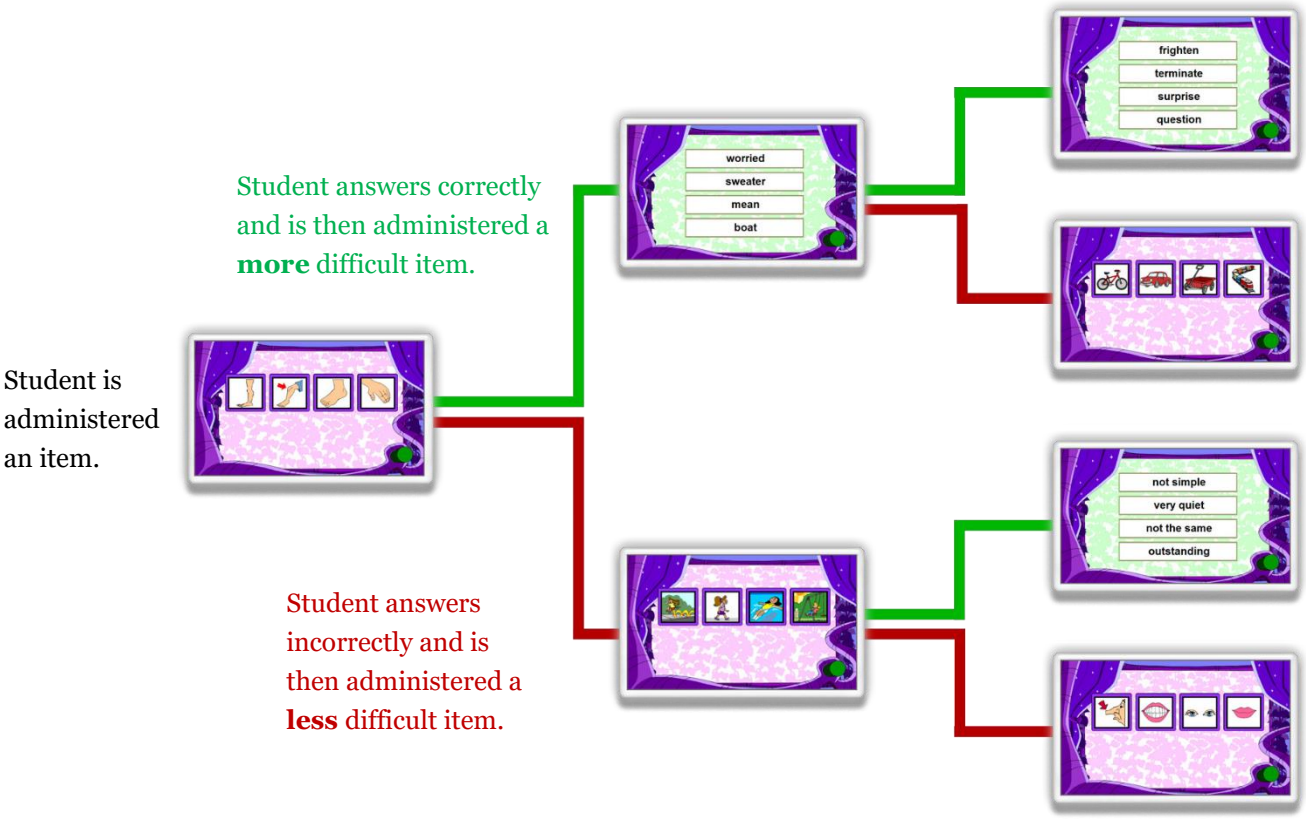

Figure 1.1. Branching pathway shows how a CAT adapts to student performance.

There are many advantages to using a CAT model rather than a more traditional parallel forms model. It is virtually impossible to create alternate forms of any truly parallel assessment. The reliability across forms will always be somewhat compromised, and the forms may have limitations for how low or how high they can assess a student's ability, depending on the depth of the item bank at any particular grade level. Conversely, in a CAT model, the assessments do not need to be identical in difficulty to previous and future assessments, and thus they may better be able to estimate ability and growth (Mathes et al., 2016). The use of CAT algorithms creates efficiencies in test administration, allowing the computer to adjust item difficulty while the student is taking the test, quickly zeroing in on ability level.

ISIP Reading Item Development

The purpose of ISIP Reading is to support teachers' instructional decisions and serve as a computer-adaptive universal screening and progress monitoring system. ISIP Reading

was first published in 2006 with parallel forms, with the CAT administration introduced in 2009. This early reading assessment focused on grades prekindergarten to third, while grades four to eight were added later in 2011. As standards have changed, new items have been introduced into the item pool.

Along with the authorship team, graduate students from the Institute for Evidence-Based Education at Southern Methodist University (SMU) were involved in item development. Students, depending on their grade, need to be assessed in listening comprehension, phonemic awareness, letter knowledge, alphabetic decoding, spelling, fluency, vocabulary, and reading comprehension, so the team searched for studies that focused on how to best assess each of these dimensions of reading as well as any possible confounds to the design of these assessments. After gaining clarity through the research, much time was spent defining models for each of the constructs and designing items to assess the models. The team further examined how each of the reading domains had been assessed in other widely accepted assessments. Enlightened with this information, the team met frequently to discuss the advantages and disadvantages of various formats and ideas of how best to assess each domain in order to reflect the model through computer administration.

Grades Prekindergarten to Three

In building the blueprint for the items within each domain, in terms of item types and number of items representing the span of skills development, the team reviewed the early release of the Common Core State Standards and certain state standards (California, Florida, New York, Virginia, and Texas) for grades kindergarten through 3 and grades 4 through 8 separately. Prekindergarten standards were also reviewed when available. The team listed the standards by grade and reading domain and then cross-referenced standards for each state, identifying standards that appeared in more than one state. Through this work, the key areas of each domain in which states expect students to demonstrate progress were determined.

Beyond these categories of skills, the standards that were analyzed also specified expectations for the level of refinement expected of students within each skill area for each grade. Using this information, the team created a flow chart by grade, illustrating each domain and the skills within and plotting expectations of skill development. This served as the foundation of the assessment blueprint. From this foundation, the numbers of items required were estimated for each domain at each grade level. Because this assessment was designed to be used universally with all students, it was recognized that a corpus of items in each domain was appropriate for students performing below

grade level as well as above grade level. Thus, the range of item types for ISIP Reading in grades pre-K to 3 was extended to include items with difficulties as low as the end of pre-K and as high as grade 6. Additionally, items were developed within each domain to represent easy, moderate, and hard items for each grade.

Grades Four to Eight

A similar process was used for grades 4 to 8, with a focus on assessing students in upper elementary and middle school. Older students need assessment in word analysis, fluency, vocabulary, and comprehension, so the team searched for studies that focused on how best to assess each of these four dimensions of reading as well as possible confounds to these design assessments. The results of this search provided great insight into the issues involved in assessing each of the four domains, as well as current thinking about how best to assess each domain. The authorship team was greatly influenced by Cutting and Scarborough's (2006) call to develop new instruments that correspond more closely to theoretical models of the constructs being measured. Thus, much time was spent defining our models for each of the four constructs and designing items to assess the models. It was further examined how each of the four domains of reading has been assessed in other widely accepted assessments. Enlightened with this information, the team met frequently to discuss the advantages and disadvantages of various formats and ideas for how best to assess each domain in order to reflect the model through computer administration of items.

This work was particularly helpful in guiding decisions on how to assess comprehension. Reading comprehension difficulties are found in as many as 15% of students, and these students may have adequate lower-level or surface processing deficits such as decoding, word recognition, fluency and/or language comprehension skills (Cain & Oakhill, 2007; Fletcher et al., 2007; Nation, 1999; Nation et al., 1999; Yuill & Oakhill, 1991). Understanding how students comprehend text at higher cognitive levels is necessary for advancement and intervention. There is consensus among the reviewed literature that reading comprehension assessments have been one-dimensional and have had little variation in reading material or response formats, and that current assessments provide little diagnostic information because they lack precision in measuring the underlying latent variables that comprise comprehension (Cutting & Scarborough, 2006; Deane et al., 2006; Fletcher, 2006; Francis et al., 2006; Millis et al., 2006; Rayner et al., 2006).

In building the blueprint for the items within each domain, in terms of item types and number of items representing the span of skills development, the state standards

for California, Florida, New York, and Texas were reviewed for grades 4 through 8. The standards were listed by grade and reading domain and then cross-referenced for each state to identify standards that appeared in more than one state. Through this work, the key areas of each domain in which states expect students to demonstrate progress were determined. Next, the team identified the big ideas that were consistent across all states and determined those big ideas could be summed up in three statements: (a) students should easily recognize increasingly complex words, (b) students should fluently process grade-level materials in a variety of genres, and (c) students should be able to derive meaning from grade-level texts representing a variety of genres.

The common skills associated with deriving meaning identified by all states examined included (a) determining a text's main ideas and how they are supported in the text, (b) analyzing text to determine the author's purpose, (c) analyzing plot structure and literary devices in a story, (d) identifying and explaining cause-and-effect relationships, (e) drawing conclusions and making predictions based on the text, (f) comparing and contrasting information in the text, (g) determining the sequence of events, and (h) distinguishing between fact and opinion. Embedded in these skills is knowledge of increasingly sophisticated vocabulary. Beyond these skill categories, the states that were analyzed also specified expectations for the level of refinement expected of students within each skill area for each grade. Using this information, a flow chart by grade was created, illustrating each domain, skills within each domain, and plotted skill-development expectations. This served as the foundation of the assessment blueprint.

From this foundation, the numbers of items required were estimated for each domain at each grade level. Because this assessment was designed to be used universally with all students, it was recognized that a corpus of items in each domain was appropriate for students performing below grade level as well as above grade level. Thus, the range of item types was extended to include items with difficulties as low as end of grade 2 and as high as college-level grade 14 words. Additionally, items were developed within each domain to represent easy, moderate, and hard items for each grade. While ultimately the item response theory (IRT) calibration work identified the difficulty of each item, the team was assured of having items representing the full achievement continuum for each domain.

With a blueprint in hand, the team developed items. ISIP Reading for grades 4 to 8 is composed of 3,100 items: 1,090 spelling items, 760 vocabulary items, 150 connected fluency stories, 220 comprehension passages, and 880 comprehension questions (4 per passage). Within the four domains, the complete item pool is distributed across the full continuum of middle school ability (i.e., grades 2-14). The use

of a CAT algorithm allows the computer to adjust to the student taking the test, and therefore the number of items are sufficient for grades 4–8.

ISIP Reading Assessment Domains

ISIP Reading tailors each assessment to the reading abilities of individual students while measuring progress in the five critical reading skill domains of phonemic awareness, alphabetic knowledge and skills, connected text fluency, vocabulary, comprehension and word analysis or spelling. ISIP Reading adapts to students' ability levels, and when they have reached a preset threshold of proficiency, they no longer receive the foundational reading subtests.

Within ISIP Reading for prekindergarten through grade 3, each subtest has both an accuracy component and a fluency component. Assessments that measure a student's accuracy and speed in performing a skill have long been studied by researchers, and they are a key component of measuring ability.

Fluency in cognitive processes is seen as a proxy for learning, such that as students learn a skill, the proficiency with which they perform the skill indicates how well they know or have learned the skill. To be fluent at higher-level processes of reading connected text, a student will also need to be fluent with foundational skills. Such fluency-based assessments have been proven to be efficient, reliable, and valid indicators of reading success (Fuchs et al., 2001; Good et al., 2001). Because each of the subtests has a fluency component, the tests are brief and, therefore, can be administered on a large scale without sacrificing valuable instruction time.

Phonemic Awareness

Phonemic awareness refers to the understanding that spoken words are comprised of individual sounds called phonemes. This awareness underpins how sound-symbols in printed words map onto spoken words. Deficits in phonemic awareness characterize most poor readers, including children, adolescents, or adults, and are not related to intelligence. Deficits can occur regardless of economic disadvantage or whether one is from non-English speaking backgrounds (Share & Stanovich, 1995). Difficulties in phonemic awareness are also noted in students at risk of dyslexia (Kirby, et al., 2003; Torgesen, et al., 1997;). The Phonemic Awareness subtest is comprised of two types of items: beginning sound and phonemic blending.

Beginning Sound

Beginning sound assesses a student's ability to recognize the initial sound in a word. In ISIP Reading, four items appear on the screen, as seen in Figure 1.2. The narrator says the name of each picture as the box around it highlights. The student is asked to click on the picture that has the same beginning sound as the sound produced by the narrator. The student may hover over a picture to hear the picture name repeated, and then they select the picture that matches the beginning sound.

Figure 1.2. Phonemic awareness/beginning sound items have four pictures to choose from.

Phonemic Blending

Phonemic blending assesses a student's ability to blend up to six phonemes into a word. Four items appear on the screen with a box in the middle of the items that contains an animated side view of a head, as depicted in Figure 1.3. The narrator says the name of each picture as the box around it highlights. The narrator says one of the words, phoneme by phoneme, as the animated head demonstrates production of each sound. The student is asked to click on the picture showing the word that has been said phoneme by phoneme. The student may move the mouse pointer over a picture to hear the picture name repeated. The highest level is a mix of five- and six-phoneme words to give the assessment a top range of ability.

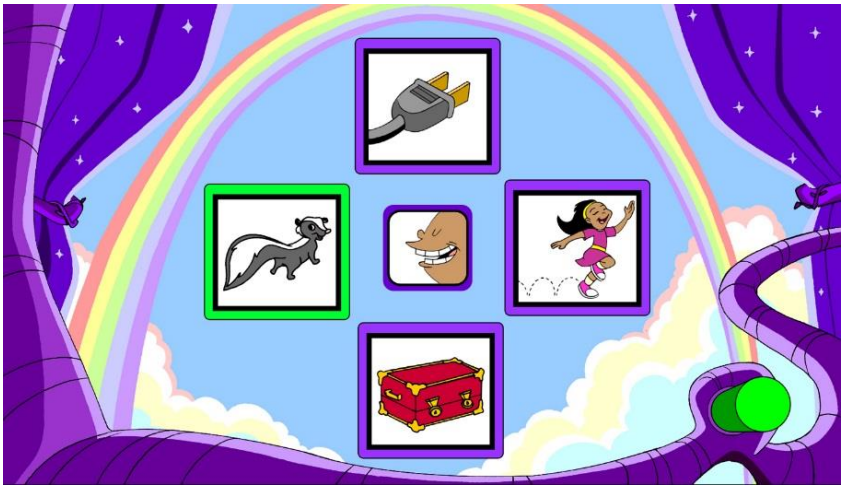

Figure 1.3. Phonemic awareness / phonemic blending item types have four answer choices.

Alphabetic Knowledge and Skills

Alphabetic knowledge and skills include knowing the symbols or combinations of symbols used to represent specific phonemes (i.e., letter knowledge) and using them to map print onto speech. The application of alphabetic knowledge and skills is exceedingly important because these skills facilitate word recognition. Reading problems for most children occur at the level of the single word because of difficulty with recognizing the letters and the sounds that go with them. Research in reading demonstrates that the best predictor of poor reading comprehension skills is deficient word-recognition ability (Shaywitz, 1996; Stanovich, 1991; Vellutino, 1991). Furthermore, alphabetic reading skills, especially alphabetic decoding (i.e., sounding out words), appear to account for individual differences in word recognition for both children and adults (Share, 1995).

ISIP assesses alphabetic knowledge with the Letter Knowledge and Alphabetic Decoding subtests.

Letter Knowledge

Letter knowledge represents the most basic level of phonics knowledge: whether students know the names and sounds represented by the letters of the alphabet. The Letter Knowledge subtest is comprised of two types of items: recognition of letter names and recognition of letter-sound correspondences. It is important to note that only the most frequent letter-sound correspondences are included in this subtest. More complex elements such as variant spellings, diphthongs, vowel teams, and r-controlled vowels are embedded in the Alphabetic Decoding and Spelling subtests (Mathes et al., 2016).

Letter recognition is a skill marked by how many letters a student can correctly identify in a minute. Five letters, in a combination of both uppercase and lowercase letters, appear on-screen at once, as depicted in Figure 1.4. The student is asked to identify the symbol for the letter name that is orally produced by the narrator. The assessment is timed to better assess fluency, and the timer is an important component of the subtest. As the academic year progresses, students should be able to identify more letters in a minute than they could at the beginning of the year. The new subscore for Letter Recognition will provide teachers more information regarding a student's ability to recognize the letters of the alphabet.

Letter sounds is a measure of alphabetic principle that assesses how many letter sounds a student can correctly identify in a minute. Five items, in a combination of both uppercase and lowercase letters, appear onscreen at once. The student is asked to identify the symbol for the letter sound that is orally produced by the narrator. Similar to Letter Recognition, the timer in Letter Sounds also helps to assess student fluency. Students should be able to recognize more letter sounds as the year progresses, and the new subscale score for Letter Sounds will give teachers more information regarding a student's progress in this essential skill.

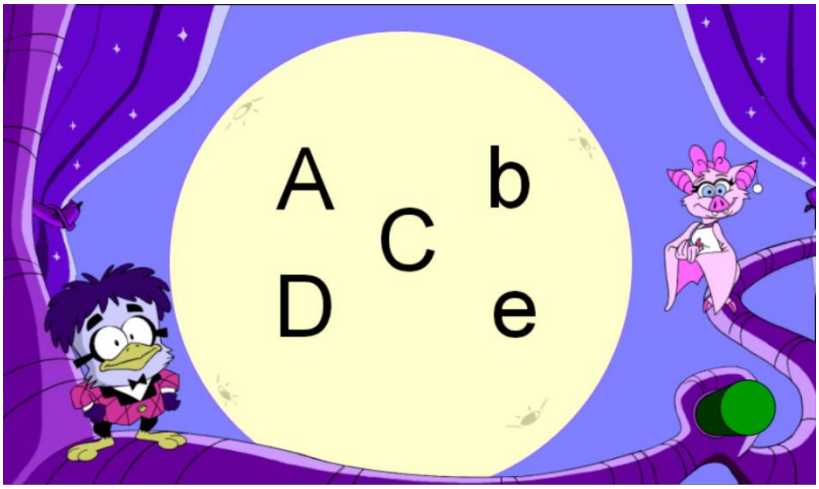

Figure 1.4. Letter Recognition contains both uppercase and lowercase letters.

Alphabetic Decoding

The Alphabetic Decoding subtest measures the ability to blend letters into nonsense words in which letters represent their most common sounds. By using nonsense words, the test more accurately assesses the ability to match letters to sounds and the ability to decode an unknown word when it is presented. For this subtest, four items appear on the screen, as seen in Figure 1.5. The student is asked to identify the

non-word that is orally pronounced by the narrator. Items for this subtest have been carefully constructed to move from easier to harder so that the subtest is appropriate across several grade levels (Mathes et al., 2016).

The sequence of difficulty moves in the following manner (Mathes et al., 2016):

1. two- or three-phoneme words composed of vc (vowel, consonant), cvc, or cv word types in which there is one-to-one letter-sound correspondence (e.g., *ib, maf, fe*);
2. three-phoneme words that include digraphs (e.g., *thil*) or diphthongs (loib);
3. three-phoneme words that include the cvce pattern with the silent e (e.g., *bave*) and four-or five-phoneme words with one-to-one letter-sound correspondence (e.g., *cvcc – kest; cvccc – kests*);
4. four-or five-phoneme words with simple blends (e.g., *ccvc – stam, stams*) and four- or five-phoneme words in which some phonemes are not represented by one letter (e.g., *caims, crame*);
5. four- or five-phoneme words with complex blends (e.g., *cccvc – streg*) and simple two-syllable words (e.g., *cvc/cvc – webbet; cv/cvc – tebet*) levels.

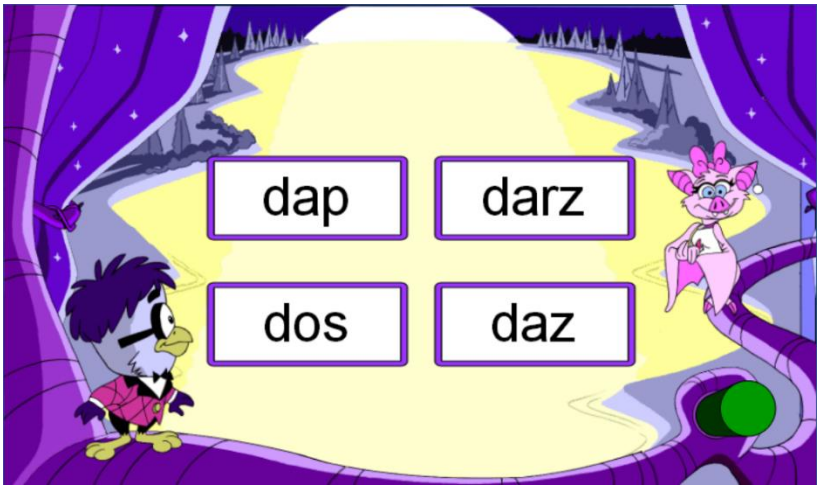

Figure 1.5. The Alphabetic Decoding subtest uses nonsense words.

Fluency

Beyond phonological and alphabetic knowledge, students must be able to read connected text with relative ease if the meaning of that text is to be understood and reading comprehension strategies are to develop and grow (Torgesen et al., , 2001). Fluency-building activities are important to use during instruction, as research indicates that fluency increases when it is included as part of classroom activities (Torgesen et al., 2001). Teachers need to know which students are not making desired gains in fluency so that they can incorporate necessary fluency strategies. ISIP addresses reading fluency with two subtests: Text Fluency and Oral Reading Fluency.

Text Fluency

Theory and Research. Successful fluent readers read connected text with both speed and understanding (Archer et al., 2003; Osborn et al,. 2003). In order to assess the full scope of fluency, measures need to incorporate both speed and meaning aspects of fluency. The maze task has been shown to be highly correlated to measures of both fluency and comprehension and has high reliability and concurrent validity (Brown-Chidsey et al., 2003; Fuchs & Fuchs, 1991; Jenkins et al., 1990; Shinn et al., ; Swain & Allinder, 1996). Delivered by computer, the maze task correlates highly to measures of oral reading fluency, comprehension measures, and high-stakes assessments (Kalinowski, 2009), and it is also associated with the risk of dyslexia in younger students (Locke et al., 2023).

Procedure. The Text Fluency subtest uses a maze task in which every fifth or sixth word is left blank from the text for grades 1 to 3, and for grades 4 to 8 there is a blank at every seventh word. All passages are near equivalent difficulty for the target grade. For each blank, the student is given three choices from which to choose the word that works in the sentence, as depicted in Figure 1.6, the student reads the text and then selects the correct maze responses for two minutes. This task has been shown to be highly correlated to measures of both fluency and comprehension and has high reliability and concurrent validity (Espin et al., 1989; Fuchs & Fuchs, 1990; Jenkins et al., 1990; Shinn et al., 1992). The Text Fluency subtest does not count toward the overall ISIP score.

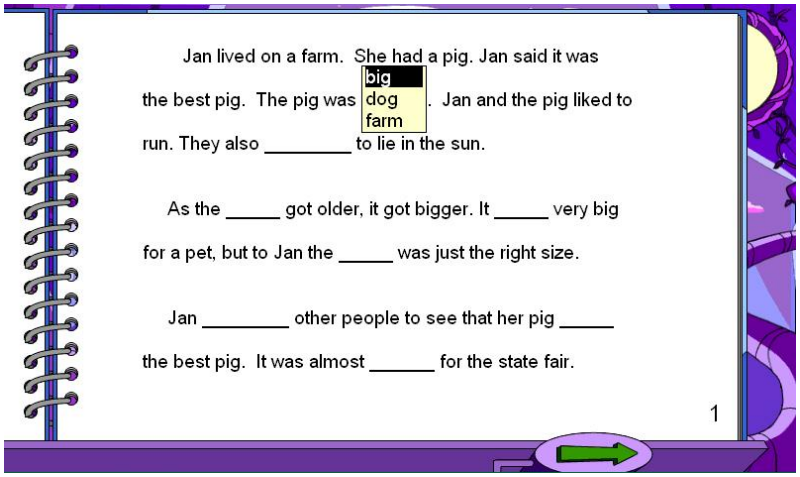

Figure 1.6. Text Fluency subtest uses a maze task, shown here.

Oral Reading Fluency

Oral reading fluency (ORF) is a measure of how accurately and quickly a student can read aloud, and it is associated with overall reading success (Hasbrouck & Tindal, 2017). ISIP ORF has grade-leveled passages that are a mixture of fiction and non-fiction. ISIP ORF randomly assigns three passages to students. Scoring is automated through the use of voice recognition technology, and a student's final score is the average of their two highest scores among the three passages.

ISIP ORF also provides a feature for teachers to listen to the passages and manually score them. The scoring measures both fluency and accuracy, and the teacher interface allows educators to take notes and document observations and remarks. Correlations are high between human and automated scoring, ranging from .97 to .99 (Istation, 2020). Norms are provided using the Hasbrouck and Tindal (2017) norms for grades 1-5. Istation also provides grade-level passages for kindergarten students so that teachers can begin to assess fluency as soon as students are able to read. ISIP ORF does not count toward the overall ISIP score, and it gives teachers valuable information about a student's oral reading abilities.

Vocabulary

The ultimate goal of all reading is to ensure that students comprehend what they read. Increasingly, there is a greater focus on the need to ensure that students possess adequate vocabulary and comprehension strategies to allow them to process text for meaning. This is especially true for students from lower socioeconomic backgrounds and from households in which English in not the primary language. Teachers need to know (a) if students have vocabulary deficits that place them at risk for failing to comprehend what they read, (b) if instruction is having the desired effect of increasing students' vocabulary knowledge, and (c) if students are making progress in comprehending increasingly challenging materials (Mathes et al., 2016).

Theory and Research. The importance of vocabulary knowledge in the development of reading skills has been extensively established in the literature (National Reading Panel, 2000). Vocabulary is especially important for students historically at risk of reading difficulties due to poverty and language background. Oral language in general and vocabulary in particular are critical to reading success (Hemphill & Tivnan, 2008; Pearson et al., 2007). Students need instruction that

accelerates their acquisition of new vocabulary and provides deep knowledge about words. Beck et al. (2002) suggest breaking words into the following three tiers:

- Tier 1 words are words that students are likely to know (e.g., sad and funny).
- Tier 2 words appear frequently in many contexts (e.g., regardless and compromise).
- Tier 3 words appear rarely in text or are content specific (e.g., irascible and biogenetics).

Beck and colleagues suggest that teachers focus vocabulary instruction on Tier 2 words drawn from content-area materials that contain words students are likely both to need (because they are encountered across contexts) and to learn well (because students will have repeated opportunities for practice and use). Tier 3 words represent a specific challenge to students since these words are the jargon of the content areas (Bravo & Cervette, 2008). ISIP Reading for prekindergarten through third grade focuses on a student's knowledge of Tier 2 vocabulary words, and the content for grades 4 through 8 focuses on both Tier 2 words (general vocabulary) and Tier 3 words (content-specific).

Vocabulary: Prekindergarten through Grade 3

The Tier 2 words in ISIP Reading have picture items beginning in prekindergarten. A picture appears onscreen, and the narrators asks the student to identify the picture that best illustrates the word they hear from the narrator, as seen in Figure 1.7. As students make progress in their skills, they begin to receive items that are synonyms. Four words appear on a screen, and the student is asked to identify the word that has the same or similar meaning as a target word.

Figure 1.7. Vocabulary picture item has four answer choices.

Vocabulary: Grades 4 through 8

For grades 4 through 8, Vocabulary items contain general vocabulary words as well as words that focus on content from social studies, science, and math. General vocabulary words were selected from national sources and state standards. Vocabulary words contain root origins from Latin, Greek, and Anglo-Saxon words. Both synonyms and antonyms are included in the answer choices, and affixes were used to construct distracters. Figure 1.8 shows a content vocabulary word from science. The narrator reads the stem for each item, and students can choose to hear the word choices by scrolling over each word on the screen. They can choose from among four possible answers by clicking on the selected answer.

There are four types of questions: (a) select the word that best matches a definition; (b) select the word most similar in meaning to the following word; (c), select the word that best describes a picture; and (d) select the word that is most similar in meaning to a target word. Distractor choices vary and include words that have a similar spelling or pronunciation, antonyms, and words with unrelated meanings. Since students acquire vocabulary best when it is used in a meaningful context, contextual questions are included.

The vocabulary content for grades 4 through 8 is more difficult than content for grades prekindergarten through 3, and vocabulary contains more Tier 3 words. This change in difficulty may impact a student's score. In the Vocabulary subtest, there may be some scores in grade 4 that achieve a higher percentile rank than a similar score in grade 3. This is due to the increased item difficulty in grades 4 to 8. More detail regarding the vertical scaling is available in chapter 2.

Figure 1.8. Vocabulary in grades 4 through 8 has general and content-specific vocabulary words.

Listening Comprehension

In this subtest, students are assessed on their ability to listen and understand grade-level sentences and paragraphs. This is accomplished through matching pictures to make meaning of what they have heard the narrator read aloud.

Matching sentences and pictures assesses a student's knowledge of semantic and syntactic information when pictures support what they are hearing read aloud. In this task, a sentence is read aloud and four pictures appear on the screen. The student identifies the picture that best illustrates the sentence's meaning.

Sentence and story completion measures a student's ability to use word meanings and word order to understand an orally read sentence or short story. In this task, a sentence or short story is read aloud and four pictures appear on the screen. One word is missing from the sentence or short story. The student must choose, from the four pictures, the word that best completes the sentence or story.

Reading Comprehension

Reading well is a demanding task requiring coordination of a diverse set of skills (Irwin, 1991). Struggling readers, even those with adequate word-level skills and acceptable fluency, often fail to use these skills as comprehension strategies, either because they do not monitor their comprehension or because they lack the necessary tools to identify and repair misunderstandings when they occur. Effective reading comprehension interventions have focused on helping students become strategic

readers by teaching them how to think while they are reading. Effective interventions have included single strategies such as finding the main idea and self-monitoring (e.g., Chan, 1991; Malone & Mastropieri, 1992; Mastropieri et al., 2003) and multi-component strategies that target reading sub-strategies (e.g., Jitendra et al., 2000; Schumaker et al., 1982). Additionally, student-led discussions of predictions, text structure, and summary development within interactive small groups have produced improvements in understanding and recalling expository text (Englert & Mariage, 1991). Reading comprehension assessments must provide information about specific comprehension abilities that can be strengthened or improved with appropriate instruction.

ISIP Reading assesses listening comprehension in prekindergarten and kindergarten, and students begin receiving the Reading Comprehension subtest in kindergarten when their scores meet a preset threshold. Reading Comprehension is a core subtest beginning in grade 1.

Grades Kindergarten to 3

ISIP Reading in grades kindergarten to 3 uses four broad areas of reading comprehension, which allows assessment of growth and provides diagnostic information to teachers to guide instruction. Students in kindergarten do not receive the Reading Comprehension subtest until their overall ISIP score reaches a predetermined threshold. Reading Comprehension in the early grades starts with matching sentences and pictures. Students read a sentence and identify the picture that best illustrates the story meaning, as depicted in Figure 1.9.

Figure 1.9. Items ask students to match sentences and pictures.

Sentence completion measures the student's ability to use word meanings and word order to understand a sentence, as depicted in Figure 1.10. A sentence, sentences, or a paragraph appears on the screen with a word missing. The student reads the text and must choose the word that best completes the sentence or text.

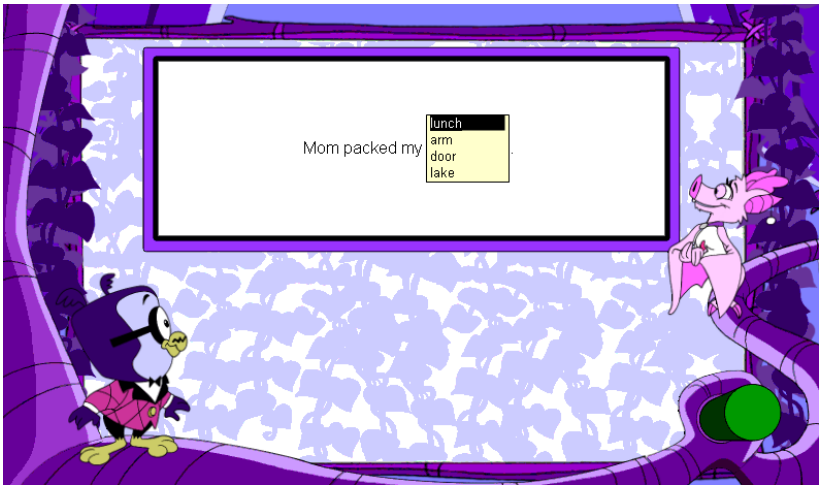

Figure 1.10. Sentence completion has students fill in the missing word based on the sentence.

Grades 4 to 8

Reading Comprehension for grades 4 through 8 becomes more advanced and assesses main idea, cause and effect, inference, and critical judgment. Passages were constructed to range in readability from grade 2.0 to 12.9 on the Flesch-Kincaid scale. After silently reading passages, students answer questions representing these four areas of comprehension ability.

Theory and Research. The underlying theory driving the reading comprehension subtest in grades 4 to 8 is that comprehension requires both low-level and high-level processing of text information. Deeper messages from the text come through when higher-level processing is used.

Higher-level cognitive processing during reading comprehension involves establishing connections among individual sentences, a concept known as local coherence (Van Dijk & Kintsch, 1983). This is evidenced in cause/effect and inference question types, where logical or causal relationships link sentences (McNamara & Magliano, 2009). Additionally, higher-level processing encapsulates the capacity to assimilate new data into pre-existing knowledge structures, creating what is termed

global coherence (Van Dijk & Kintsch, 1983). Within a testing scenario, global coherence is manifested through main idea, problem/outcome, and critical judgment question types. These require a comprehensive understanding of the text, synthesis of the overarching theme, and formation of higher-level judgements based on the entirety of the text (Zwaan & Singer, 2003) as well as being able to integrate new information into existing representations to establish global coherence of test (i.e., main idea, problem/outcome, and critical judgment question types) (Cain & Oakhill, 1999; Cain et al,. 2001; Oakhill 1982; Wixson & Peters, 1987).

All questions are designed to be dependent on information in the passage to avoid testing of background knowledge or having questions that can be answered without reading the text, a pitfall of other assessments (Keenan & Betjemann, 2006). During development as passages were being written, work was checked by asking high-performing middle-grade students the questions without asking them to read the passages. If the questions could be answered correctly, they were removed from the item bank. Since some comprehension measures are also linked to decoding ability (Keenan et al., 2008; Cutting & Scarborough, 2006), we solved this problem by matching text difficulty to a student's text reading ability, allowing the assessment of ability to be in processing text for meaning, not for decoding

All answer choices are related to the passage in some form. Also, because proficient memory has been associated with reading ability and skilled text comprehension (Cain, 2006; Daneman & Merikle, 1996; Sesma et al., 2009; Swanson et al., 2007), the text will not be available to students when they are answering questions. However, specific details that do not add to an understanding of the general or global coherence of the passage are not questioned. Thus, once students begin answering questions, they cannot see the passage again. Last, passages were written that include a range of structures found in both narrative and expository text, since comprehension failure has been linked to inadequate knowledge about how texts are structured (Perfetti, 1994). Understanding students' deficiencies in different types of text structures will help when intervening. Thus, a student's working memory is used.

Procedures. To complete the subtest, the students read a passage that appears on the screen, and they are told to read the passage for meaning. When they are ready, they turn the page, and the first of four multiple choice items appears. Students are not allowed to go back and review the passage. Figure 1.11 shows a reading comprehension item and sample question.

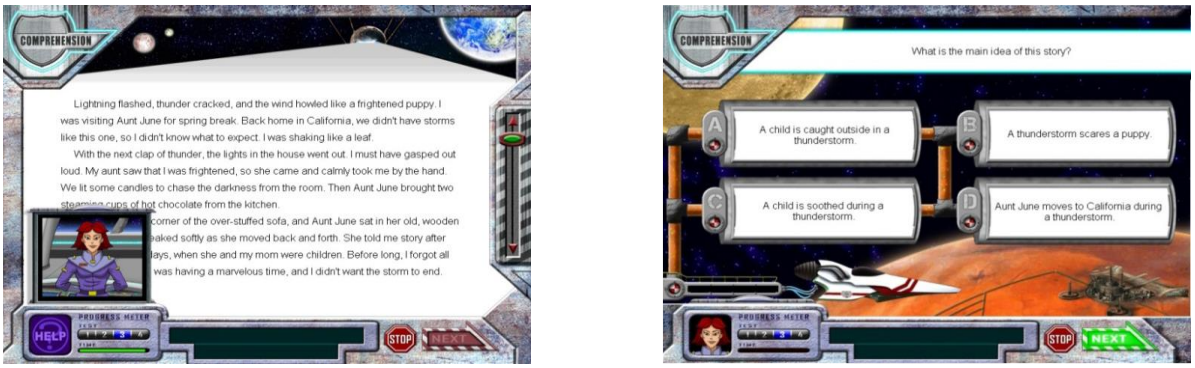

Figure 1.11. Reading comprehension in grades 4 through 8 provides the passage followed by four questions.

Word Analysis or Spelling

Learning to spell and learning to read rely on much of the same underlying knowledge, specifically the relationships between letters and sounds, and knowing the spelling of a word makes it accessible for fluent reading (Ehri, 2000; Snow et al., 2005). Young readers are often encouraged to use invented spelling as it helps them to develop their phonological awareness (Kilpatrick, 2015). Accurate and automatic identification of multisyllabic words is critical to comprehension of grade-level content-area texts (Deshler et al., 2001; Gersten et al., 2001) and distinguishes good readers from poor readers (Perfetti, 1986). Good readers use word components or parts — such as knowledge of syllable types, prefixes, suffixes, and roots — to identify long, multisyllabic words (Lenz & Hughes, 1990; Perfetti, 1986).

Targeted instruction in advanced word analysis can improve reading outcomes by teaching students strategies to effortlessly recognize increasingly complex words that they encounter in text (Scammacca et al., 2007). A valuable way to assess word analysis is with spelling. Correct spelling requires that a student possess a fully specified orthographic representation for each word, thus providing valuable information about the student's word-analysis skills (Bourassa & Treiman, 2001; Ehri, 2000; Ehri & Wilce, 1987; Graham, 2000; Perfetti, 1997).

Items go on a continuum from easy to difficult as defined in Alphabetic Decoding. Items also include the frequency of spelling patterns, with less frequent patterns being considered more difficult. The subtest also includes sight words that are frequently seen. These sight words are included because they can be difficult to spell phonetically but are important for reading fluency. In grades 4 through 8, students are asked to spell multisyllabic words that are carefully selected to contain the various aspects of syllables, affixes, and roots.

An array of letters appears on the screen, and the narrator asks the student to spell a specific word using the letters, as seen in Figure 1.12. In grades 4 through 8, the narrator says a word in a sentence and repeats it, and the students type the word.

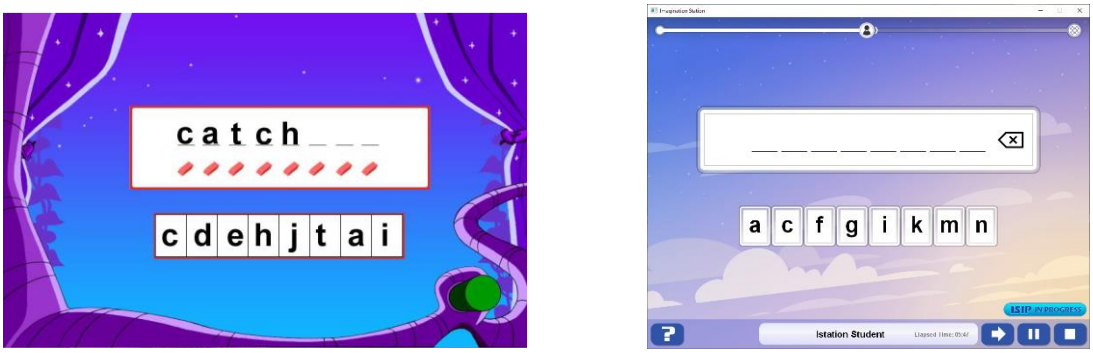

Figure 1.12. The Spelling subtest has students spell a word using an array of letters.

ISIP RAN

ISIP Rapid Auto Naming (RAN) is a digitally administered assessment that evaluates a student's ability to rapidly identify a series of pictures, letters, and numbers, as shown in Figure 1.13. There are several uses for RAN assessments, and they are primarily used to identify children at risk for reading and learning difficulties (Wolf & Denckla, 2005). In ISIP RAN, students are asked to name symbols, letters, and numbers. They go through a training process first, and responses are recorded. Teachers can score the responses afterwards at their convenience. Standard scores and percentile ranks are available for kindergarten through third grade and are given separately for symbols, letters, numbers, and a final composite standard score. Full details about ISIP

RAN are available in the ISIP RAN Technical Report (Istation, 2022).

Figure 1.13. ISIP RAN has objects, numbers, and letters.

ISIP Reading and Progression of Skills

ISIP Reading measures progress in each critical component of reading instruction in a manner appropriate to the underlying domain. There is a total of eight subtests that align to the five critical domains of reading, as shown in the table below. Of these subtests, six are built using a CAT algorithm, while two use parallel forms. Subtests that tailor items using CAT include Phonemic Awareness, Letter Knowledge, Alphabetic Decoding and Spelling, Vocabulary, and Reading Comprehension. Connected Text Fluency and Listening Comprehension are designed as parallel forms that measure end-of-grade-level expectations.

Within a classroom, students may have some variation in the exact subtest they are administered, and students' scores reflect these differences. For example, students whose performance scores indicate that they are not yet reading words will not be asked to read connected text. Similarly, students whose performance scores indicate that they read connected text fluently and with comprehension will not be asked to complete letter knowledge and phonemic awareness tasks. Most of these differences occur in the early grades. For example, in kindergarten students begin the year with Listening Comprehension, Phonemic Awareness, Vocabulary, and Letter Knowledge subtests. When their individual scores meet a preset threshold, they begin to receive Alphabetic Decoding, and as their score increases, they may begin to receive Reading Comprehension. In first grade, students start with Phonemic Awareness, Vocabulary, Alphabetic Decoding, Letter Knowledge, Reading Comprehension, and Spelling. As their individual scores meet preset thresholds, the foundational subtests of Phonemic Awareness and Letter Knowledge are dropped, and students begin to receive Text Fluency when their score is high enough on Alphabetic Decoding to suggest they can handle the task. Text Fluency is administered to all students beginning in grade 2, although Text Fluency does not count toward the overall score.

If students are struggling readers and their score falls below a preset threshold, they may begin to receive Phonemic Awareness, Alphabetic Decoding, and Letter Knowledge. Teachers may also administer these subtests separately, and if they do so, the student's performance on these subtests will count toward the overall score. The standard defaults are shown in Table 1.1.

Table 1.1. *Subtests Administered in ISIP Reading, by Grade*

Subtest	Prekindergarten	Kindergarten	Grade 1	Grades 2-8
Listening Comprehension	**Standard**	**Standard**	**Not standard** may be added	**Not standard** may be added
Letter Knowledge	**Standard**	**Standard**	**Standard** dropped after reaching threshold	**Not standard** may be added
Vocabulary	**Standard**	**Standard**	**Standard**	**Standard**
Phonemic Awareness	**Added after** reaching threshold	**Standard**	**Standard** dropped after reaching threshold	**Not standard** may be added
Alphabetic Decoding	**Added after** reaching threshold	**Added after** reaching threshold	**Standard**	**Not standard** may be added
Reading Comprehension	**Not assessed**	**Added after** reaching threshold	**Standard**	**Standard**
Spelling	**Not assessed**	**Not Standard** may be added	**Standard**	**Standard**
Text Fluency	**Not assessed**	**Not assessed**	**Not standard** may be added	**Standard**
Oral Reading Fluency	**Not assessed**	**Supplemental**	**Supplemental**	**Supplemental**
RAN	**Not assessed**	**Supplemental**	**Supplemental**	**Supplemental**

Teacher Friendly

ISIP Reading is teacher friendly. The assessment is computer based, requires little administration effort, and requires no teacher/examiner testing or manual scoring. Teachers monitor student performance during assessment periods to ensure results' reliability. In particular, teachers are alerted to observe specific students identified by ISIP Reading as experiencing difficulties as they complete ISIP Reading. They 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 Reading battery or individual skill assessments.

Student Friendly

ISIP Reading is also student friendly. The original, classic versions of ISIP Reading are administered in game-like sessions. Students in prekindergarten through grade 3 play a fast-paced computer game called "Show What You Know." In the beginning of the session, an animated bird named Alex Treebeak enters the screen with his assistant, Batana White, a white bat. Alex announces to the student in a game-show announcer voice, "It's time to play… Show What You Know!" A curtain pulls back to show the first game. Alex announces the game quickly, and the assessment begins. At the end of the assessment, the student sees an animated graph of progress. Each assessment proceeds in a similar fashion.

For students in grades 4 through 8, the assessment feels like playing a computer game called "Right Stuff University." In the beginning of the session, an animated character named Commander North enters the screen. The commander announces to the student in an authoritative voice, "Welcome to the Right Stuff University! We are looking for cadets with the right stuff. You will embark on a series of missions to prove your strengths." Students choose a trainer to guide them through their missions. Once a trainer is chosen, students begin their assessment missions. Each assessment proceeds with instruction from the chosen trainer.

Alternate Themes

Different backgrounds are available for students in grades 2 through 8. These themes, known as *Skyline* and *Night,* offer student choice in selecting the background, which increases student agency. Students enjoy picking their background, and it helps to alleviate student fatigue. The different backgrounds are engaging without impacting

the item content or how it is delivered, as seen in Figure 1.14 — similar to guidelines recommended by Dadey et al. (2018). Research with the different backgrounds showed that there were either no significant differences from the classic versions or that the effect sizes were .05 or less, and these effect sizes diminished over time, indicating a novelty effect (Patarapichayatham & Locke, 2022a; Patarapichayatham et al., 2021a). These different backgrounds help older, struggling students by giving them an opportunity to take the foundational subtests in a background that is more age-appropriate. They also provide fewer distractions for students who may have attention deficits.

Figure 1.14. Depiction of alternate backgrounds for classic, skyline, and night themes.

The CAT Algorithm

The initial Item Response Theory (IRT) study determined the item characteristics of item discrimination and item difficulty (for greater detail, see Mathes et al., 2016; Mathes, 2016). The Overall ISIP score is based on a student's ability, or theta score. ISIP uses a two-parameter model, which is preferred over a single-parameter model. IRT models estimate a single latent trait (ability), and this trait is assumed to account for response behavior. These models provide response probabilities based on the test taker's ability and the item parameters of discrimination and difficulty. The Bayesian approach incorporates prior knowledge about the student. We used Gauss-Hermite quadrature with 88 nodes from −7 to +7. The algorithm is as follows:

1. Assign an initial ability estimate to the student.
2. Ask the question that gives the most information based on the current ability estimate.
3. Re-estimate the ability of the test taker.
4. If stopping criteria is met, stop. Otherwise, go to step 2.

Stopping criteria has a minimum of five items and a maximum of 20 items. We end ISIP Reading when the ability score's standard error drops below a preset threshold or when four consecutive items do not reduce the standard error by a preset threshold.

Thus, the number of items administered may vary across administrations for the same student and across a classroom of students.

Reliability for a CAT is computed by this formula:

$$\rho^2 = 1 - [\mathrm{SE}(\theta)^2]$$

where θ is the student ability. In prekindergarten through grade 3 the reliability is .891, and in grades 4 through 8 it is .868, indicating that ISIP Reading is very reliable. A more detailed explanation is available in chapter 5.

Conclusion

The ISIP Reading assessment is based on recommendations by the National Reading Panel (2000). This chapter gives an overview of the ISIP Reading assessment and the goals for the 2022 update. The remainder of this report focuses on the reading update rather than the development of the original assessment. Previous information is consolidated to orient the reader of this report, and greater detail is available in our prior technical reports.

Chapter 2. Vertical Scaling

Introduction

In the prior editions of ISIP Reading, Istation assessed reading via two different computer-adaptive tests: Istation's Indicators of Progress Early Reading (ISIP ER) for students in prekindergarten through grade 3 and ISIP Advanced Reading (AR) for students in grades 4 through 8.

These assessments were reported out on separate vertical scales (Patz & Yao, 2006; Tong & Kolen, 2010; Carlson, 2011; Young & Tong, 2015). Item response theory (IRT) was used to calibrate the ER and the AR items separately to produce within-grade scales. This was followed by linking together the different ER within-grade scales to produce a single ER vertical scale that spanned the range from pre-K to grade 3. A similar process was used to link together the AR scales for grades 4 through 8 to form the AR vertical scale (Patz & Yao, 2006; Tong & Kolen, 2010; Carlson, 2011; Young & Tong, 2015). The results were two separate and non-comparable vertical scales.

In order to allow for the monitoring of students' reading progress from the ER test to the AR test, Istation decided to link the separate vertical scales to form a single vertical scale that would go from pre-K through grade 8. A special *bridge study* to link the ER and AR scales (i.e., "bridge the gap") was carried out to accomplish this goal.

The following sections present the procedures used to create the new ISIP Reading vertical scale. These include the design used to collect the student data for the study and prepare it for analysis; the approach used to link together the separate ER and AR scales, and steps incorporated to create the coefficients needed to do this; the application of the linking coefficients to create the new reading vertical scale and convert it to an appropriate reporting scale; and validating the resulting scale.

Methods

Data Collection Design

The data for the bridge study were collected in April and May 2021. The grade 3 and grade 4 students from the participating schools were randomly assigned to take either the ER or the AR test delivered using the Istation platform in what is called a *random equivalent groups design* (see Kolen, 2007; Kolen & Brennan, 2004).

In this design, a single group of examinees is randomly assigned to one of two test forms, usually via a spiraling process. Thus, in the study, a sample of grade 3 students was randomly assigned to take either the ER or AR test. Similarly, a sample of grade 4 students was also randomly assigned to take either the ER or AR test as well. Figure 2.1 below shows a schematic of this data collection design as applied to the bridge study.

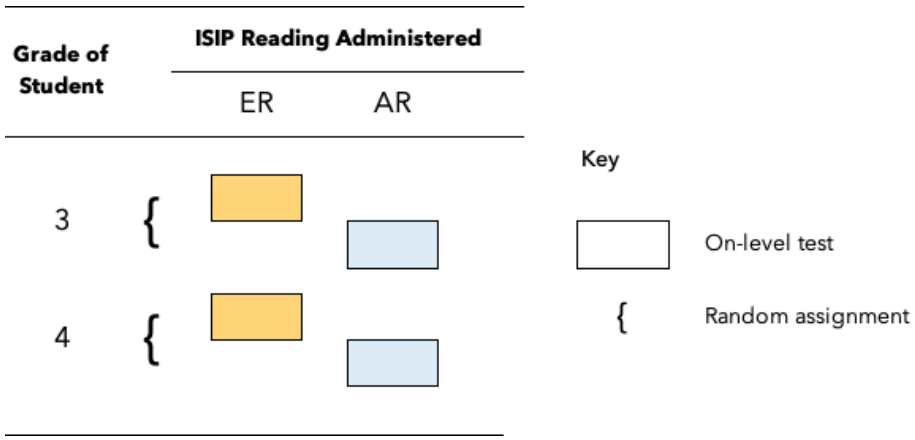

Figure 2.1. Schematic of the random equivalent groups design used in the ISIP Reading bridge study.

The rows of the figure denote the grades of the students in the study while the columns show the test-level of the items administered to them. The boxes in the figure represent the groups of students taking the test content targeted to their level, and the different colors indicate the differences in test content between the ISIP ER (orange) and ISIP AR (blue).

Comparing the performance of grade 3 students on the ER and AR tests, we would expect that the students would, on average, find the AR test to be harder than the ER test. Similarly, the grade 4 students would find the ER test (once again, on average) to be easier than the AR test.

The proper implementation of this process leads to two groups of examinees such that the groups are randomly equivalent with respect to their ability. The differential performance of the examinee groups on the test that they were each assigned is then used to create the statistical adjustment to put the two tests on the same scale.

Analysis Procedures

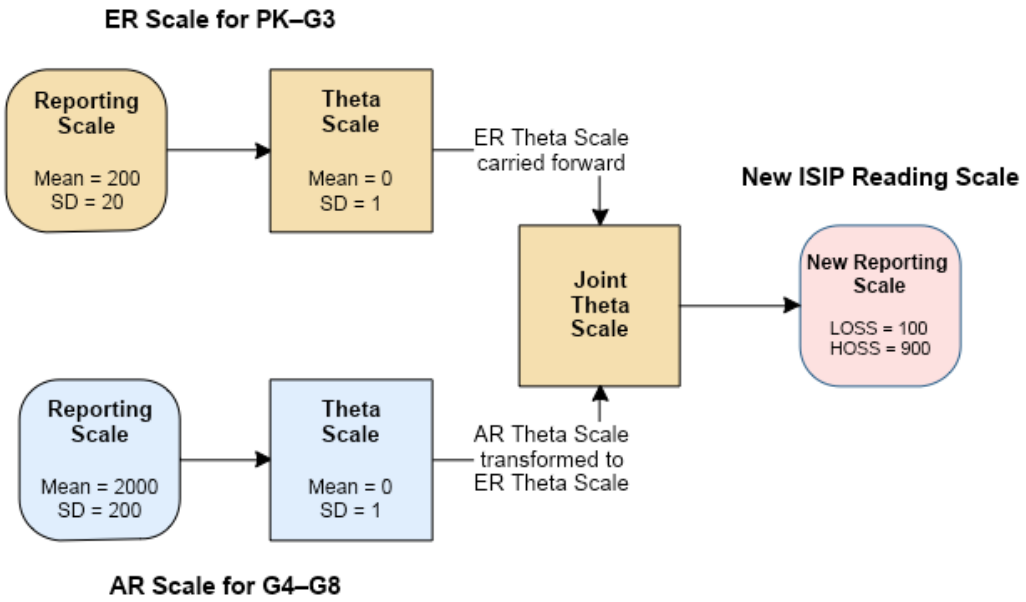

Figure 2.2. This schematic outlines the process used to link the ISIP ER and ISIP AR scales in order to create a new combined scale. The two boxes on the top row represent the ER scale while the two boxes on the bottom represent the AR scale. The box on the right represents the new ISIP Reading reporting score that is derived by linking those scales together.

Given the data collection design described above, the process used to link the ER and AR scales is shown in Figure 2.2 above. After cleaning up the data, we used these steps:

1. Prepare the data collected from the bridge study for analysis, validate the assumptions of the random equivalent groups design, and calculate the statistics needed for the following steps;
2. Convert the ER and AR reporting scale scores back to their IRT theta scale scores;
3. Use the data to create the linking constants that are used to transform the AR theta scale to the ER theta scale;
4. Convert this joint ER/transformed AR theta scale to a final reporting scale and validate the new reading scale with student results.

These steps are described in more detail in the following sections.

Preparation for Data Analysis

The data collected in the bridge study included these fields:

- The test administration date and unique student identifier;
- The grade of the student and the bridge-study test taken (i.e., ER or AR);
- The student's overall ISIP Reading test scale score for the random equivalent groups data collection gathered in the April-May 2021 window;
- The student's ISIP Reading test scale scores for the on-grade test the student took in January 2021.

It is important to clarify the two kinds of test score data that were collected as they were used for different purposes (see Figure 2.3).

	Reading Test Scale Score	
Grade	On-Grade	Random Equivalent Groups
3	ER	ER
		AR
4	AR	ER
		AR

Figure 2.3. Test score data collected during ISIP Reading bridge study

As can be seen in the figure, each student with complete case data in the study had two different scale scores. Each student had, of course, the score on the test they had been assigned to as a part of the *random equivalent groups* data collection design. In addition, at each grade, all of the students in that grade had a score for an *on-grade* ISIP Reading test that they took in January 2021.

Specifically, each grade 3 student had a score on the ER test, while each grade 4 student had a score on the AR test. Recall that the purpose of the random equivalent groups design was to have the ability distributions of the students assigned to either the ER or AR be the same. By having these on-grade ISIP scores available, it was possible to check this assumption and see if the random assignment of students to tests had been properly done and that, in fact, there were no significant differences between ER and AR

group abilities for a given grade. Specifically, this was done by looking at the difference in test score means for ER versus AR groups on the January 2021 assessment for grade 3 and grade 4 students separately, and expressing those differences as effect sizes.

The additional data preparation steps included:

- checking all variables for out-of-bound and missing values;
- removing any duplicate student cases or cases with completely missing item response strings; and
- removing student cases that were incomplete due to missing a January 2021 on-grade test score.

Once this had been done, the data from the random equivalent groups collection were used to create the statistics needed for subsequent steps in order to link together the ER and AR scales.

Conversion of Reporting Scales to Theta Scales

After completing the data cleaning and check of the random equivalent groups assumption, the next step was to convert the separate ER and AR reporting scales back to their original IRT or *theta* scales. This was done because the ER and AR reporting scales were quite different. The original ER scale was set to have a mean of 200 and a standard deviation of 20, while the AR scale was set with a mean of 2,000 and a standard deviation of 200. The scores on the ER and AR reporting scales, SS_{ER} and SS_{AR} respectively, were transformed back to their theta scales using

$$\theta_{ER} = (SS_{ER} - 200)/20$$

and

$$\theta_{AR} = (SS_{AR} - 2000)/200.$$

Transforming the reporting scales in this way to theta scales allowed both scales to have means of zero and standard deviations of one, meaning differences in performance between grades on a given test are shown in terms of standard deviation units.

Determining Linking Constants

When the same group of students has taken two tests, the scales can be linked using the process of *linear equating* (Kolen & Brennan, 2004). That is, we can find

linking constants (i.e., a slope and an intercept) that can be used to transform the measures of student ability from one scale into another.[1] In the context of the bridge study, the data were used to create the linking constants needed to transform the AR scale scores into the ER scale scores. The linear transformation from the AR scale to the ER scale is derived by first setting the *standardized* or *z-scores* of the ER and AR theta scales equal to each other:

$$z_{ER} = (\theta_{ER} - Mean(\theta_{ER})/SD(\theta_{ER}) = (\theta_{AR} - Mean(\theta_{AR})/SD(\theta_{AR}) = z_{AR},$$

and solving for θ_{ER} in terms of θ_{AR}, and the sample means and standard deviations of the ER and AR theta scores.

Thus,

$$\theta_{ER} = (SD(\theta_{ER})/SD(\theta_{AR})) \cdot \theta_{AR} + \big(Mean(\theta_{ER}) - (SD(\theta_{ER})/SD(\theta_{AR})) \cdot Mean(\theta_{AR})\big).$$

Letting $A = SD(\theta_{ER})/SD(\theta_{AR})$

and

$$B = Mean(\theta_{AR}) - (SD(\theta_{ER})/SD(\theta_{AR})) \cdot Mean(\theta_{ER})$$

$$= Mean(\theta_{AR}) - A \cdot Mean(\theta_{ER}).$$

We then have that

$$\theta_{ER} = A \cdot \theta_{AR} + B.$$

which is the linear transformation from the AR theta scale to the ER theta scale with slope *A* and intercept *B*. Given this result, all student scores can be reported on the same vertical scale using either the ER scale for the Early Reading test or the transformed AR-to-ER scale for the Advanced Reading test.

Creating and Validating the New ISIP Reading Reporting Scale

The final step in the process was to take the ER theta/AR-to-ER theta scale and transform it to a final scale that would be more appropriate for reporting ISIP Reading scores across all of the grades from pre-K through grade 8. Essentially, we needed to choose a linear transformation from the ER theta/AR-to-ER theta scale to the new reporting scale:

[1] Note that the transformation can also be used to transform the item difficulty and discrimination parameters from one scale to the other.

$$SS_{New\ Scale} = A_{New\ Scale} \cdot \theta_{ER,\ AR \rightarrow ER} + B_{New\ Scale}$$

Although the choices of the slope and intercept to create the new scale were arbitrary, there were two main considerations that guided their selection. First, the new ISIP Reading scale needed to have a range of values that was distinct from the ranges of the original scales. This needed to be done in order to emphasize the fact that ISIP Reading was now on a single vertical scale that was different from the individual ER and AR 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 to have an underlying vertical scale that would support the full range of reading achievement on the current tests while allowing for flexibility in describing the floor and ceiling of reading achievement going forward.

The approach used to derive the new ISIP Reading scale was to choose two points on the theta scale to represent the effective minimum (*Theta Low*) and effective maximum (*Theta High*) on that scale and map those scores into the choices of the *LOSS* and *HOSS* for the reporting scale.

Thus, the slope and the intercept for the new reporting scale would be given by

$$A_{New\ Scale} = \frac{HOSS - LOSS}{\theta_{High} - \theta_{Low}}$$

and

$$B_{New\ Scale} = HOSS - A_{New\ Scale} \cdot \theta_{High}.$$

Once the transformation of the joint ER/AR-to-ER theta scale to the reporting scale had been made, the final step in the process was to validate the reporting 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;
- the grade-to-grade variability; and
- the separation of grade distributions.

These attributes were examined using ISIP Reading student data taken from the January 2021 test administration for the entire population of students. The grade-by-grade means and standard deviations were used to assess the average growth and variability of the student distributions on the new reporting scale. The separation of the

grade distributions was examined using an index proposed by Yen (1986) that is the effect size for the grade-to-grade differences.

Results

Descriptive Statistics: Original ER and AR Scales

The *n*-counts of the students for the data as originally collected, the final n-counts after the data cleaning process described above, and descriptive statistics on the original ER and AR scales are shown in Table 2.1.

Table 2.1. *N-Counts and Descriptive Statistics for the ER/AR Bridge Study Scale Scores*

Assigned Test	Grade	Original N-counts	Final N-counts	Scale Scores Mean	Scale Scores SD
ER	3	375	292	248.0	20.0
	4	475	371	250.3	18.6
	Total	850	663	249.3	19.2
AR	3	324	214	1754.9	168.1
	4	442	306	1852.0	184.1
	Total	766	520	1812.1	183.8
Both Tests	Total	1,616	1,183		

The table shows that, as expected, grade 4 students performed better, on average, than grade 3 students on each of the tests. We can better see this when we look at the effect sizes associated with the difference in test scores' means across grades within an assigned test (Table 2.2).

Based on Cohen's (1988) commonly used criteria for categorizing the magnitude of effect sizes[2] , the difference in the grade 3 and grade 4 means is small on the ER test and medium on the AR test. These results may be due to the 3rd and 4th graders being much closer in performance with respect to the familiar material on the grade 3 ER test, but further apart in performance on what would be new material for the 3rd graders on the grade 4 AR.

[2] See Grissom and Kim (2012, pp. 127–132) for an extensive discussion of Cohen's criteria and some of the pitfalls that can be encountered if using them mechanically.

Table 2.2. *N-Counts, Descriptive Statistics, and Effect Sizes for Differences Between Grades on ISIP Reading Tests*

Assigned Test	Grade	N	Mean	SD	Mean Diff.*	Pooled SD	Effect Size
Early Reading	3	292	248.0	20.0	−2.3	19.2	−0.12
	4	371	250.3	18.6			
Advanced Reading	3	214	1754.9	168.1	−97.1	177.7	−0.55
	4	306	1852.0	184.1			

*Grade 3 scale score minus grade 4 scale score

Validating the Random Assignment of Students to Tests

Table 2.3 looks at the students' results on the on-grade reading test they were administered in January 2021.

Table 2.3. *N-Counts, Descriptive Statistics, and Effect Sizes for Differences on the January 2021 On-Grade ISIP Reading Tests*

Grade	Assigned Group	N	Mean	SD	Mean Diff.*	Pooled SD	Effect Size
3	Early Reading	292	243.5	21.3	−1.4	22.3	−0.06
	Advanced Reading	214	242.1	23.6			
4	Early Reading	371	1816.5	208.1	−5.1	206.3	−0.02
	Advanced Reading	306	1811.4	204.0			

*Early Reading test scale score minus Advanced Reading test scale score

Recall that the goal of the random equivalent groups design is to produce two groups of students of the same ability level at each grade, with one group taking the ER test and the other group the AR test. This time, effect sizes were used to examine the difference between the mean test scores for students assigned to the ER and AR groups at grade 3 and at grade 4.

The resulting effect sizes are negligibly small, showing very little to no difference in the ability levels of the students assigned to the different tests in each grade. This result provides evidence that the random assignment of students to tests during the bridge study was successfully carried out and that the groups are randomly equivalent.

Descriptive Statistics: ER and AR Theta Scales

The reporting scale statistics in Table 2.1 are shown in Table 2.4 on the theta scales of the respective tests.

Table 2.4. *Descriptive Statistics for the ER/AR Bridge Study on the Separate Theta Scales*

Test	Grade	N-count	Mean	SD
Early Reading	3	292	2.40	1.00
	4	371	2.51	0.93
		663	2.46	0.96
Advanced Reading	3	214	–1.23	0.84
	4	306	–0.74	0.92
		520	–0.94	0.92

The large, positive values of the means for the ER test indicate that the upper part of the scale (i.e., grade 3 of the pre-K to grade 3 ER test span) is being measured. Similarly, the means of the AR test are negative, indicating that the lower end of the AR scale in grade 4 is being tested.

Developing the Final Scale

Calculating the Equating Coefficients

Given the ER and AR theta scale statistics, the next step was to create the linking coefficients (i.e., the slope and intercept of the linear transformation) needed to place scores from the AR theta scale on the ER theta scale. The data collection for the bridge study allowed for these coefficients to be calculated in three different ways using (a) only the grade 3 data,(b) only the grade 4 data, or (c) both the grade 3 and grade 4 data together. Table 2.5 shows all three of these solutions.

Table 2.5. *Initial Linking Coefficients for Placing the AR Theta Scale on the ER Theta Scale*

Data	Test	N-count	Mean	SD	Slope (A)	Intercept (B)
Grade 3	Early Reading	292	2.40	1.00	1.188	3.855
	Advanced Reading	214	−1.23	0.84		
Grade 4	Early Reading	371	2.51	0.93	1.009	3.260
	Advanced Reading	306	−0.74	0.92		
Both Grades	Early Reading	663	2.46	0.96	1.046	3.446
	Advanced Reading	520	−0.94	0.92		

The three solutions were applied to ISIP Reading user data and the then-current ISIP Reading norms. The review of the solutions focused on the means and standard deviations of the student performance by grade and test administration timeframe (i.e., beginning, middle, and end of year) and at key percentiles tied to tiers of response to intervention (RTI) support. The decision was made to use the data from both grades and use the slope and intercept of that solution for the linking coefficients.

Transformation to the Final Reporting Scale and Validation

The transformation of the joint ER/AR-to-ER theta scale to the reporting scale was determined by setting the low and high values of the theta scale to −3.50 and 9.00 respectively, and the LOSS and HOSS of the new reporting scale to 100 and 900. This resulted in the slope and the intercept for the new reporting scale being given by

$$A_{New\ Scale} = \frac{HOSS - LOSS}{\theta_{High} - \theta_{Low}} = \frac{900 - 100}{9.0 - (-3.5)} = 64.00$$

and

$$B_{New\ Scale} = HOSS - A_{New\ Scale} \cdot \theta_{High} = 900 - 64.00 \cdot 9.0 = 324.00.$$

The attributes of the final reporting scale were examined using ISIP Reading student data taken from the January 2021 test administration. The student scores from this administration were transformed from their original grade-specific scales to the new ISIP Reading reporting vertical scale. The final transformations were, therefore, for the new scale scores, and for the standard error of the new scale scores. The results of applying these transformations are shown in Figure 2.4 and Table 2.6.

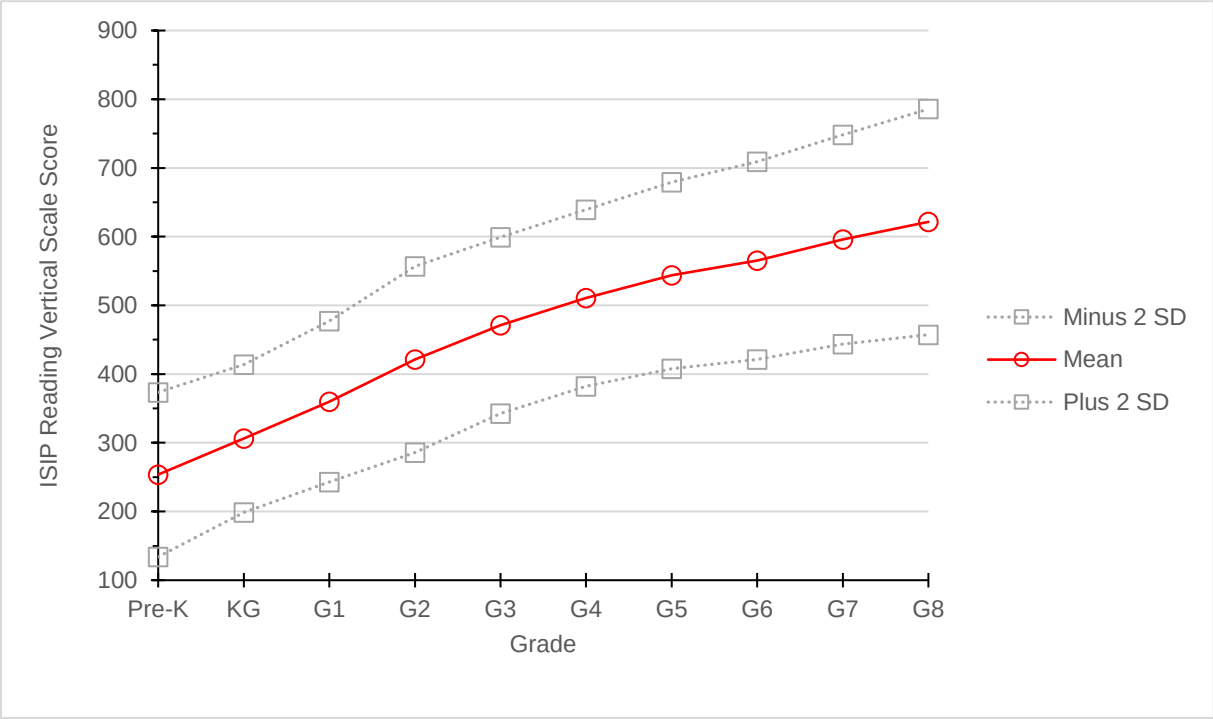

Figure 2.4. ISIP Reading final reporting growth curves: Scale score means and confidence bands by grade level (Source: ISIP Reading January 2021 test administration data)

Table 2.6. *ISIP Reading Vertical Scale Grade-to-Grade Growth, Variability, and Distribution Separation (Source: ISIP Reading January 2021 test administration data)*

Grade –Lower	Grade – Upper	Difference	Pooled SD	Effect Size
PK	K	52.6	54.6	0.96
K	1	53.7	56.5	0.95
1	2	61.4	63.3	0.97
2	3	49.5	66.1	0.75
3	4	39.7	64.2	0.62
4	5	32.9	66.1	0.50
5	6	21.6	69.1	0.31
6	7	31.0	73.7	0.42
7	8	25.4	78.7	0.32

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, and generally lessening from grade 2 through grade 8. This pattern is similar to that seen for other vertically scaled achievement tests such as the *Stanford Achievement Test* series (10th Edition)(Young & Tong, 2015).

The grade-to-grade variability tends to increase across the grades, and the effect-size measures decrease. The effects sizes decrease from nearly a full standard deviation (0.95 to 0.75) at the lower grades to less than half a standard deviation (0.31 to 0.42) at the higher 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.

The evidence suggests that the final reporting scale for ISIP Reading successfully joins the originally separate Early Reading and Advanced Reading vertical scales and should be of use to teachers, administrators, and parents in monitoring student progress across the entire pre-K to grade 8 range.

Chapter 3: Norming

Determining Norms

The purpose of having normative information is to provide data for evaluating student performance. Norms show the scores that are typical for a student in a particular grade, and they help identify students at risk of failure in their current, or future, grade. Norms also help identify student proficiency and help educators see their students' performance in a larger context. To fulfill these purposes, norms need to be recent, relevant, and reliable. This chapter will describe the process we used for selecting the normative sample and the norming procedures. We will also provide information on how the norms function within decision consistency, with special group studies, and using ISIP as a dyslexia screener.

The COVID-19 Dilemma

Istation most recently updated its assessment norms using data from the 2014-2015 school year. Re-norming an assessment typically takes place every four to five years. Therefore, by the 2019-2020 school year, it was time to update the norms. However, both the 2019-2020 and the subsequent 2020-2021 school years were disrupted by the pandemic caused by the spread of the novel coronavirus 2019 (COVID-19). Student performance during the pandemic was erratic as schools moved to remote learning in the spring semester of 2020. Ohio was the first state to close schools on March 12, 2020, and was closely followed by other states (Grossmann, Reckhow, Strunk, & Turner, 2021). In the fall of 2020, there was wide variation in instruction across the country, depending on local conditions. Some schools stayed virtual with instruction across the school year primarily conducted at home via conferencing technology. Other schools had hybrid options, where some students attended in person and others attended remotely, and some schools had predominately in-person instruction (Kelly, 2021). The patterns changed throughout the year as waves of the pandemic ebbed and flowed (Grossmann et al., 2021).

Students experienced greater-than-expected learning lags at the beginning of the year, and the lags were exacerbated throughout the 2020-2021 school year (Kuhfeld, Soland, & Lewis, 2022). Many assessments were also administered at home, and overall

student performance lagged behind previous years (Patarapichayatham, Locke, & Lewis, 2021b). Istation determined that data from the 2020-2021 school year would not yield results that would be relevant; therefore, we delayed re-norming another year in anticipation that the pandemic would wane and student performance would begin to recover.

The 2021-2022 school year encompassed the second half of the Delta variant (late summer, early fall 2021) and the first Omicron wave (winter 2021/2022). There were marked absences from school by both teachers and students. In some instances the teacher shortages were so severe that states turned to the National Guard to provide substitute teachers and bus drivers; other states turned to police to serve as substitutes (Nierenberg, 2022). Mean scores on the ISIP Reading assessment in the early grades continued to lag those of previous school years, while students in the upper grades appeared to begin to recover (Patarapichayatham & Locke, 2022).

We recognize that norms are only as good as the sample of students in the data set, and poor student performance can result in norms that are too easy and misidentify students. For example, in previous years their score may have placed them in a category for students at risk of reading failure, while norms that are too easy may categorize them for being on track for reading success. Conversely, above-average student performance may result in norms that are too difficult. Istation had planned to use the 2021-2022 school year data, but by the middle of the year it became evident that not all grades and not all areas of the country were at a place where student performance had reached pre-pandemic levels.

Before deciding to use 2021-2022 school year data or to use the most recent pre-pandemic school year (2018-2019), we held several focus group discussions with teachers, instructional coaches, administrators, superintendents, and state-level administration professionals. We conducted internal analyses on Istation data, reviewed the research literature on norm performance during a disrupted school year, attended national conferences where student performance during the pandemic was discussed, and also consulted with psychometric experts in the field.

The focus group participants fell into two different camps. One camp advocated for using the 2021-2022 school year data, reasoning that students were scoring lower than previous cohorts and the lag would persist and represent a new normal. These professionals thought that our norms needed to reflect reality rather than a pre-pandemic ideal.

The other camp advocated for using the 2018-2019 school year data. This group expressed concern that using pandemic data for norms would result in norms that are too easy. Students who would previously score in the 30th percentile might score higher and enter Tier 1, thus giving a false picture of their reading ability. Since one of the purposes of norms is to provide information regarding student performance, this was a primary concern.

After reviewing the student achievement data from the 2021-2022 and 2018-2019 school years, we observed that particularly in the younger grades, using the 2021-2022 data would produce norms that may under-identify students at risk of reading failure. We therefore decided to use the 2018-2019 school year data to provide the most relevant information.

We have one exception to this data set. We wanted to provide a normative update for the Alphabetic Decoding subtest for kindergarten students. This subtest is administered to kindergarten students after they have achieved a preset threshold of reading proficiency. We ran a study in the 2021-2022 school year where, beginning in January (period 5), all kindergarten students received the Alphabetic Decoding subtest, and it was only included in the Overall score if the students had met the preset threshold. This was to provide normative information for this subtest for when it is used as part of the dyslexia screener. We did not have this data available in the 2018-2019 school year, and since student performance above the 60th percentile was not dramatically different across the two norming years, we determined that it was reasonable to use the 2021-2022 data to norm Alphabetic Decoding for kindergarten, for periods 5-9 (January-May).

Sampling Methodology

Since users of Istation may differ somewhat from the national population, we stratified the sample using a school socioeconomic index. The school socioeconomic index is derived from sociodemographic information at the school and the surrounding area. Post-stratification helps to reduce the bias in sampling as long as the variables have a relationship with the outcome (Jagers, 1986). In this instance, socioeconomic status at the school level has a long-established relationship with student academic performance. We created a post-stratification index using enrollment information from the National Center for Education Statistics (NCES) for the 2019-2020 school year. We also used information from the American Community Survey five-year period estimates

from the US Census Bureau to capture child/family poverty in the area around the school.

Construction of the school stratification index

To create the school index (SI), we relied on research regarding the school challenge index (SCI), designed by researchers at the Northwest Evaluation Association, who based it on the similar schools index in California (Thum & Hauser, 2015). Using a linear regression model, the outcome variable was the percentage of students at the school receiving free or reduced-price lunch (FRPL). Predictor variables included school level information from the NCES including number of teachers, the teacher-student ratio, Title I status, and the racial/ethnic composition within the school, which typically is highly correlated with socioeconomic status. We developed the sample frame by using data from the NCES. We compared this list with schools that were in the Istation database, and we added those schools to the frame if they were not included in the NCES data file.

We also wanted to account for variance in the neighborhood or surrounding area as a predictor. We added additional information for the child/family poverty rate in the zip code for the school location. These data came from the American Community Survey 2015-2019 five-year period estimates data set (US Census Bureau, 2021) available from the Integrated Public Use Microdata Series (IPUMS) at the University of Minnesota (Ruggles, Flood, Goeken, Schouweiler, & Sobek, 2022). All population rates were transformed into logit units as these rates are typically not normally distributed. There were 97,310 schools in the final data file. Since we were using administrative data from the NCES, there were missing data in the sample frame. Missing data were imputed using predicted means matching in *R* statistical software.

The variables used in the construction of the index as predictor variables in a regression model consists of a variety of sociodemographic information that included the following:

Free or Reduced-Price Lunch (FRPL)	*Percentage of students eligible for free or reduced-price lunch (LN transformation)*
Percent White	*Percentage of students who are non-Hispanic White (LN transformation)*
Percent Black or African American	*Percentage of students who are non-Hispanic Black or African American (LN transformation)*
Percent Hispanic	*Percent of students who are of Hispanic origin of any race (LN transformation)*
Teacher	*Total number of full-time Teachers (LN transformation)*
Teacher-Student Ratio	*Ratio of teachers per student (LN transformation)*
Locale of school	*Whether the school is 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*
Child/Family Poverty	*Child/Family Poverty at the school zip code (LN transformation)*

Results from the regression model are available in Table 3.1. In the regression model, the child/family poverty in the zip code, Title I status, locale, and racial/ethnic composition at the school were the strongest predictors. Using the predicted variable for the outcome measure, we rescaled them to create a normal curve equivalent.

Table 3.1. *Results from the Regression Model to Construct the School Index*

Measure	Coefficient	t	p
Intercept	−0.386	−42.389	< .001
Student enrollment: % White	−0.096	−66.667	< .001
Student enrollment: % Black or African American	0.081	69.075	< .001
Student enrollment: % Hispanic or Latino	0.049	38.597	< .001
Number of Teachers	−0.027	−24.982	< .001
Teacher-Student Ratio	−0.018	−14.597	< .001
Locale: Suburban	−0.061	−13.755	< .001
Locale: Rural	0.215	43.237	< .001
Elementary Schools	−0.048	−9.945	< .001
Middle Schools	0.021	3.525	< .001
High Schools	0.009	1.593	.11
Title 1	0.333	68.776	< .001
Title 1: School-wide program	0.341	73.175	< .001
BIE	0.190	4.873	< .001
Charter	−0.175	−26.499	< .001
Magnet	−0.028	−3.085	< .001
Regular school	−0.095	−14.711	< .001
Region: Northeast	−0.090	−16.842	< .001
Region: Midwest	0.003	0.573	.57
Region: West	0.026	5.364	< .001
Child/family poverty	0.182	105.351	< .001
Model R^2 = .52			

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

Next, the SI was divided into eighths of the distribution, or octiles. A low value indicated schools with greater challenges due to the characteristics of the school and neighborhood. Private and parochial schools were identified as a separate category. The percent of students on free or reduced-price lunch and the rate for child/family poverty change by SI level, as seen in Table 3.2.

Table 3.2. *Percent of Students Receiving Free or Reduced-Price Lunch (FRPL) and Child/Family Poverty by School Index (SI)*

SI	Mean FRPL	Mean Child/Family Poverty
1	88%	32%
2	79%	23%
3	69%	19%
4	59%	17%
5	51%	15%
6	42%	12%
7	32%	8%
8	21%	4%

We also calculated the means of the January overall score for the previous score and the new scale, and calculated means by SI. Results are available in Table 3.3. Notably, the mean student performance went up across all eight SI levels. Students in private and parochial schools performed similarly to students in SI 4-6, and they scored below students in wealthier public schools. The one exception to this is in seventh grade, where students in private and parochial schools scored lower than public school students.

Table 3.3. *Means of Scale Score Performance by School Index (SI)*

Grade	Scale	SI 1	SI 2	SI 3	SI 4	SI 5	SI 6	SI 7	SI 8	Private/ Parochial
Pre-K	Previous Scale	171.74	172.83	173.64	174.03	175.06	176.48	178.43	175.10	175.54
	Current Scale	233.61	237.15	239.72	240.89	244.31	248.71	254.99	244.36	245.77
K	Previous Scale	190.58	192.82	193.89	195.11	196.77	196.07	197.67	199.25	195.88
	Current Scale	293.87	301.01	304.44	308.35	313.64	311.44	316.54	321.59	310.81
1	Previous Scale	205.87	207.88	210.12	211.50	213.64	214.91	217.75	220.09	215.28
	Current Scale	342.80	349.21	356.37	360.80	367.66	371.72	380.79	388.30	372.89
2	Previous Scale	224.97	227.65	229.90	231.17	233.67	235.45	237.81	240.41	235.42
	Current Scale	403.89	412.48	419.68	423.73	431.74	437.43	445.00	453.33	437.34
3	Previous Scale	238.31	240.98	243.69	245.33	247.92	250.12	252.12	255.58	250.16
	Current Scale	446.58	455.14	463.80	469.07	477.36	484.40	490.79	501.84	484.50
4	Previous Scale	1836.19	1859.16	1882.65	1901.58	1927.42	1948.25	1974.84	2015.08	1932.01
	Current Scale	489.73	497.41	505.27	511.60	520.25	527.22	536.12	549.58	521.78
5	Previous Scale	1914.84	1936.28	1967.54	1989.49	2017.86	2041.10	2068.84	2110.75	2025.28
	Current Scale	516.04	523.21	533.68	541.02	550.51	558.28	567.57	581.59	552.98
6	Previous Scale	2104.94	2003.72	2035.96	2040.80	2083.84	2087.11	2105.84	2069.37	2103.04
	Current Scale	545.78	556.59	558.19	572.59	573.68	579.94	567.74	579.00	579.62
7	Previous Scale	2176.59	2186.78	2231.91	2176.49	2185.59	2213.99	2264.98	2200.10	2147.69
	Current Scale	603.62	607.01	622.10	603.58	606.61	616.12	633.11	611.47	593.96
8	Previous Scale	2147.69	2176.59	2186.78	2231.91	2176.49	2185.59	2213.99	2264.98	2159.37
	Current Scale	593.96	603.62	607.01	622.10	603.58	606.61	616.12	633.11	597.83

Sample targets and selection

The next step was to compute sampling targets using the enrollment data for the elementary and middle schools. We used the total enrollment for elementary schools for grades K-5, and for grades 6-8 we used the total enrollment for middle schools as denoted in the NCES file. Sixth grade was included in middle school. We compared the targets to the enrollment in the Istation Reading program and determined that to achieve a representative sample, we would need to do some post-stratification. Due to

potential sample bias, we used slightly different procedures in grades K-5, 6-8, and prekindergarten.

Procedures for the sample targets for the elementary and middle school samples

To calculate the targets for the elementary school sample, we calculated the number of students in each SI level for elementary school enrollment and calculated the percentage of students. We also decided to add approximately 1-2% of private school students on top of the public school enrollment targets. Therefore, the targets Table 3.4 for elementary school students add up to 100%. Similarly, for the targets for the middle school sample, we calculated the number of students in each SI for middle schools.

Procedures for the sample targets for the prekindergarten sample

Methods varied slightly in the prekindergarten sample. Among 4-year-old children, approximately 54.5% were enrolled in school, according to the US Census American Community Survey 2020 one-year period estimates, available from IPUMS. Of these students, 39.5% were enrolled in private or parochial schools, and 60.5% were enrolled in public schools (Ruggles et al., 2022). We therefore included 39.5% of our sample from private and parochial schools, and the remaining 60.5% were stratified by the SI levels. We determined targets for the public school sample by calculating the percentage of public school enrollment by SI level for the number of students enrolled in prekindergarten. Therefore, for the prekindergarten targets, the SI level and the private school targets add up to 100%.

Stratification procedures

Since the ISIP Reading assessment is used for progress monitoring and benchmarking, selecting only observations that have complete data would bias the sample. Benchmarking months can vary state by state and year by year. Istation divides the instructional months depending on the first day of school, and the first month is considered Period 0. Subsequent months are numbered sequentially. In the Istation database, we observed that benchmarking typically occurs in periods 0, 1, and 2 for the beginning of the year (BOY); months 4, 5, and 6 for middle of the year (MOY); and months 7, 8, and 9 for the end of the year (EOY). Student observations were selected as eligible for norming if they had at least three observations: one each in BOY, MOY, and

EOY. This helps to account for those schools that benchmark at different times. This reduced the number of eligible observations.

Using the targets derived from the NCES data file described above, we conducted sampling by grade. Sampling without replacement was conducted for grades prekindergarten through 5, and with replacement in grades 6 through 8. We used the data sets that contained eligible observations, and we randomly selected a sample of students according to the targets derived from the NCES enrollment data. Extreme outliers based on middle-of-the-year scores were eliminated.

Next, we checked the mean performance of the sample by grade to determine if there was a continuous progression within the grades and that the sample performance met our expectations. We noted in kindergarten and grades 6 through 8 that the random selection of students resulted in a sample that performed lower than the adjacent grades. This was particularly noteworthy in grades 6 through 8, where students were performing at a mean in the 30th to 39th percentiles. Using this random selection of students could result in sample bias. We therefore changed the criteria for inclusion in the kindergarten and grades 6 through 8 samples. We sampled within both SI and the Istation performance level categories (quintiles) to create a sample that was more typically achieving. Sampling with replacement was held to 4.2% of the sample in grades 6 through 8. In grades 6 through 8, if there were not enough students in an SI level, then we increased the sample in an adjacent SI level. Results of the stratification are available in Table 3.4 for the normative sample and in Table 3.5 for the kindergarten Alphabetic Decoding sample. In the normative sample, there were over 835,000 students enrolled in 5,926 schools in 42 states. The kindergarten Alphabetic Decoding sample had 74,700 students enrolled in 2,780 schools in 33 states.

Table 3.4. *Percent of Public School Students by School Index (SI) Octile and of Private/Parochial School Students for the 2018-2019 School Year*

Targets/ Sample	N	SI 1	SI 2	SI 3	SI 4	SI 5	SI 6	SI 7	SI 8	Private/ Parochial
Prekindergarten Targets (ACS and NCES)	N/A	9.8%	10.3%	9.1%	8.1%	6.3%	5.7%	5.6%	5.7%	39.5%
Prekindergarten Sample	9,100	10.9%	11.3%	9.9%	9.1%	7.0%	6.3%	3.9%	2.4%	39.2%
Grades K-5 NCES Targets	N/A	12.3%	15.01%	15.03%	12.9%	9.8%	8.8%	11.5%	14.7%	1-2%
Kindergarten Sample	144,752	12.1%	15.3%	15.2%	13.0%	10.2%	8.1%	10.4%	13.4%	2.4%
Grade 1 Sample	169,623	11.9%	14.8%	14.8%	13.7%	10.4%	8.5%	10.4%	13.8%	1.7%
Grade 2 Sample	150,000	12.3%	14.8%	14.8%	12.6%	9.5%	8.6%	11.5%	14.3%	1.7%
Grade 3 Sample	125,000	12.3%	15.0%	15.1%	13.0%	9.9%	8.6%	10.6%	13.9%	1.6%
Grade 4 Sample	100,000	12.1%	14.8%	14.8%	12.7%	9.6%	8.6%	11.4%	14.5%	1.5%
Grade 5 Sample	80,000	12.0%	14.7%	14.7%	12.7%	9.6%	8.6%	11.3%	14.4%	1.9%
Grades 6-8 NCES Targets	N/A	11.83%	12.56%	10.87%	9.37%	11.57%	13.73%	16.11%	13.95%	1-2%
Grade 6 Sample	25,000	9.2%	9.3%	11.6%	8.6%	20.6%	7.3%	20.0%	12.5%	1.0%
Grade 7 Sample	25,000	10.1%	8.7%	10.9%	12.0%	16.9%	9.6%	18.1%	12.7%	1.1%
Grade 8 Sample	8,000	11.9%	8.3%	11.8%	14.7%	22.0%	4.6%	15.9%	9.9%	1.0%

Table 3.5. *Percent of Public School Students by School Index (SI) Octile and of Private/Parochial School Students for the Kindergarten Alphabetic Decoding Sample for the 2021-2022 School Year*

Targets/ Sample	N	SI 1	SI 2	SI 3	SI 4	SI 5	SI 6	SI 7	SI 8	Private/ Parochial
Grades K-5 NCES Targets	N/A	12.3%	15.01%	15.03%	12.9%	9.8%	8.8%	11.5%	14.7%	1-2%
Kindergarten Sample	74,700	12.2%	15.4%	15.5%	13.2%	10.6%	8.1%	10.1%	12.5%	2.4%

Norming Analysis

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 since their inception.

As described in chapter 1 of this report, ISIP Reading produces a scale score that describes overall student performance in reading. In addition, the assessment also produces scores for the subtests that are a part the Early and Advanced levels[1]. In all, 65 sets of norms were needed for the assessment, and these are summarized by grade and domain in Table 3.6.

Table 3.6. *ISIP Reading Norms Developed by Grade and Test/Subtest*

Test	PK	K	1	2	3	4	5	6	7	8
Overall Reading[1]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Phonemic Awareness	✓	✓	✓							
Letter Knowledge	✓	✓	✓	✓						
Letter Recognition[2]	✓	✓	✓							
Letter Sounds[2]	✓	✓	✓							
Alphabetic Decoding	✓	✓	✓							
Listening Comprehension	✓	✓	✓							
Reading Comprehension			✓	✓	✓	✓	✓	✓	✓	✓
Spelling			✓	✓	✓	✓	✓	✓	✓	✓
Vocabulary	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Text Fluency			✓	✓	✓	✓	✓	✓	✓	✓

1 The Overall Reading score is a composite of the scores obtained on the tests taken.
2 Letter Recognition and Letter Sounds are subtests producing scores derived from the Letter Knowledge subtest.

Three kinds of norm-referenced scores are reported for ISIP Reading, namely, *percentile ranks* (PRs), *levels*, and *instructional tier goals*. The percentile rank shows the percentage of students in the norm group that was 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 five levels — 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 who score at or below the 20th percentile are placed in Tier 3 and are at significant risk of not meeting grade-level expectations. Students who score in Tier 2 (between the 20th and 40th percentiles) are said to be at some risk of not meeting grade-level expectations. Finally, students who score above the 40th percentile (Tier 1) are said to be on track to meet grade-level expectations. Istation divided Tier 1 into three separate levels to give greater differentiation for students in this tier, especially those between the 41st and 60th percentiles, who may get overlooked in the classroom.

Thus, in terms of the analyses needed, only the PRs associated with the tests shown in Table 3.6 needed to be calculated, as cut scores for the levels and instructional tier goals would follow by definition.

Data Preparation

The data collected in the study were broken up into grade-specific files that contained the stratified samples of the students' test scores. The files included fields showing…

- the unique student identifier, their grade, and the unique identifier for the school the student attended;
- various demographic variables such as gender and race/ethnicity; and
- the SI of the school composite index used to post-stratify schools as described above.

For each test administration period, the student's CAT scale score, scale score standard error, percentile rank on the existing norm set, and performance level were captured for each assessment. The ISIP administration periods are denoted in Table 3.7.

Table 3.7. *ISIP Reading 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

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

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

Finally, a cleaned data file was produced for each of the ISIP Reading tests to be used for further analyses. All of these data preparation steps were written in the SAS software language environment (SAS Institute Inc., 2019) and executed using WPS Workbench (World Programming Limited, 2022).

Norming Approach

ISIP Reading requires normative information to be developed for each period of the school year. In traditional approaches to norming, 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 the time period as a continuous variable to predict test scores, one can use information taken from across the entire year to estimate norms. A robust approach called *nonparametric continuous norming using Taylor polynomials* was used to develop the new norms for ISIP Reading (Lenhard et al., 2018). In this approach, the scale scores on a test are modeled as a continuous function of a student's location in the distribution of scale scores on a test (i.e., their percentile or normalized standard score) and an explanatory variable such as age or grade.

In the context of ISIP Reading, the student scale scores for a given test were modeled as a function of percentiles that had been transformed into normalized scores (L) and of the grade and administration period (A) of the test as

$$SS = f(L, A) = \sum_{s,t=0}^{k} c_{st} L^s A^t,$$

with the c_{st} as the coefficients of a polynomial multiple regression analysis with the $L^s A^t$ terms as the independent variables, and the value k representing the common, maximum power of the terms[3]. This multiple regression equation is fit stepwise with all of the powers and products of the model shown above, and the final Taylor polynomial function is defined by choosing the significant variables from the stepwise regression and using the unstandardized beta weights as the c_{st} constants in the polynomial (Lenhard et al., 2018, p. 116). Using this fitted Taylor polynomial, we can then find the scale score (SS) for each percentile rank corresponding to L within each administration period A and produce a comprehensive set of norms tables.

In developing the ISIP norms for a given reading test, this approach was applied to either the data taken from all of the administration periods of a single grade or from all of the periods across a set of grades. For instance, when the data clearly showed that

[3] For example, for $k = 2$, the independent variables of the regression would be L, L^2, A, A^2, LA, LA^2, L^2A, and L^2A^2, as $L^0 = A^0 = 1$, $L^1 = L$, and $A^1 = A$.

there was a summer drop-off in the test scores from the end of one grade to the start of the next, then the norms for these grades were produced separately. This was done so that the students' actual performance on the test was not smoothed away to produce an unreal depiction of student growth[4].

Example of the Norming Process

Examining the data

This norming approach using Taylor polynomials is demonstrated using the overall reading scores from the grade 3 sample described above. The empirical scale score distributions for these data were examined using descriptive statistics (Table 3.8), boxplots (Figure 3.1), and histograms (Figure 3.2).

These data show that grade 3 overall reading growth from the beginning to the end of the year (August to May) is a bit over 40 scale score points or roughly 0.6 standard deviations. The distributions are slightly negatively skewed, and the positive kurtosis values show that the tails of the distribution are heavier than those of a normal distribution. The boxplots provide a better understanding of why the kurtosis is positive, as they clearly show the length of the distributions' tails and presence of a large number of outliers. Both the boxplots and the histograms show how the distribution shifts upward over the course of the year. This is especially clear with the histograms, where the period-by-period scale score distributions across the year can be compared with a fixed, normal distribution in the middle of the year.

Table 3.8. *ISIP Reading Grade 3 Normative Sample: Overall Scale Score Descriptive Statistics by Time of Year*

Time of Year	Mean	SD	Skewness	Kurtosis
August	447.2	58.0	−0.04	1.04
September	449.7	58.8	−0.19	1.01
October	455.7	61.6	−0.31	1.08
November	462.4	64.1	−0.22	1.27
December	467.5	63.6	−0.31	1.29
January	472.7	65.4	−0.21	0.96

[4] An excellent discussion regarding the reporting of norms vís-a-vís the phenomenon of summer drop-off can be found in the technical manual for the *California Achievement Test, Fifth Edition* (CTB/McGraw-Hill, 1996, pp. 312–313).

February	475.0	65.4	−0.29	1.52
March	479.9	68.0	-0.20	1.60
April	483.1	67.4	-0.29	1.77
May	489.2	69.4	-0.15	1.81

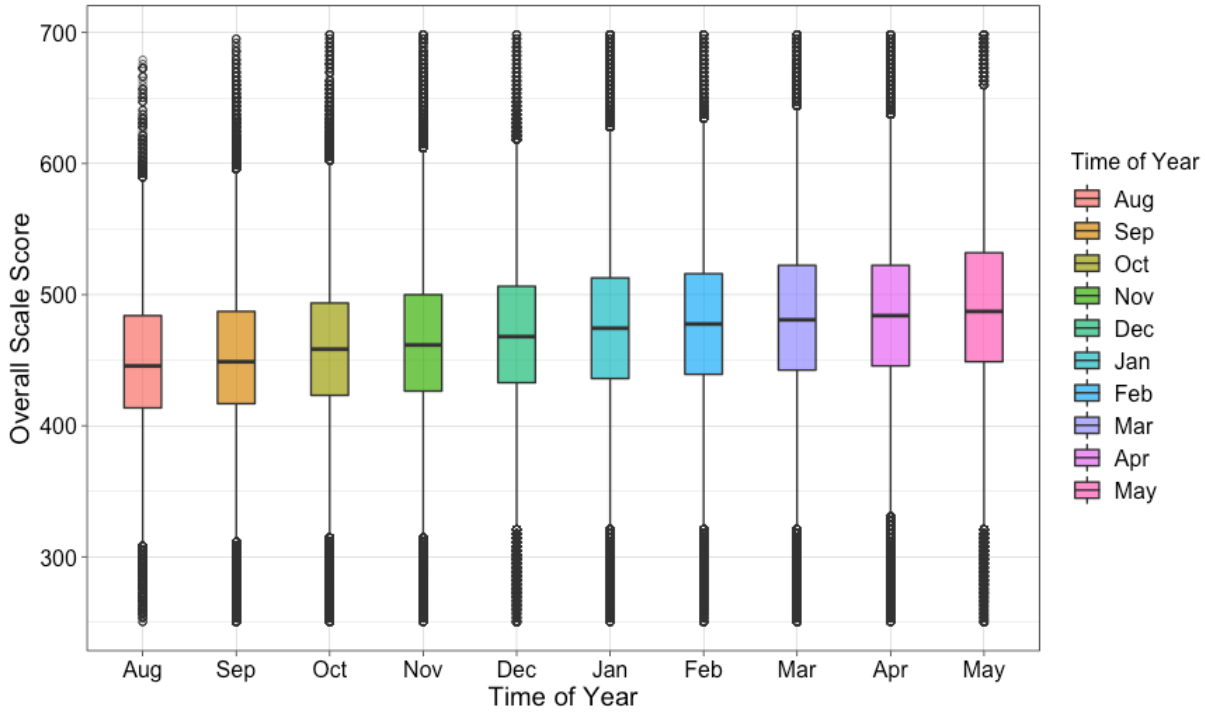

Figure 3.1. Boxplots of ISIP Reading grade 3 scale score normative samples by time of year

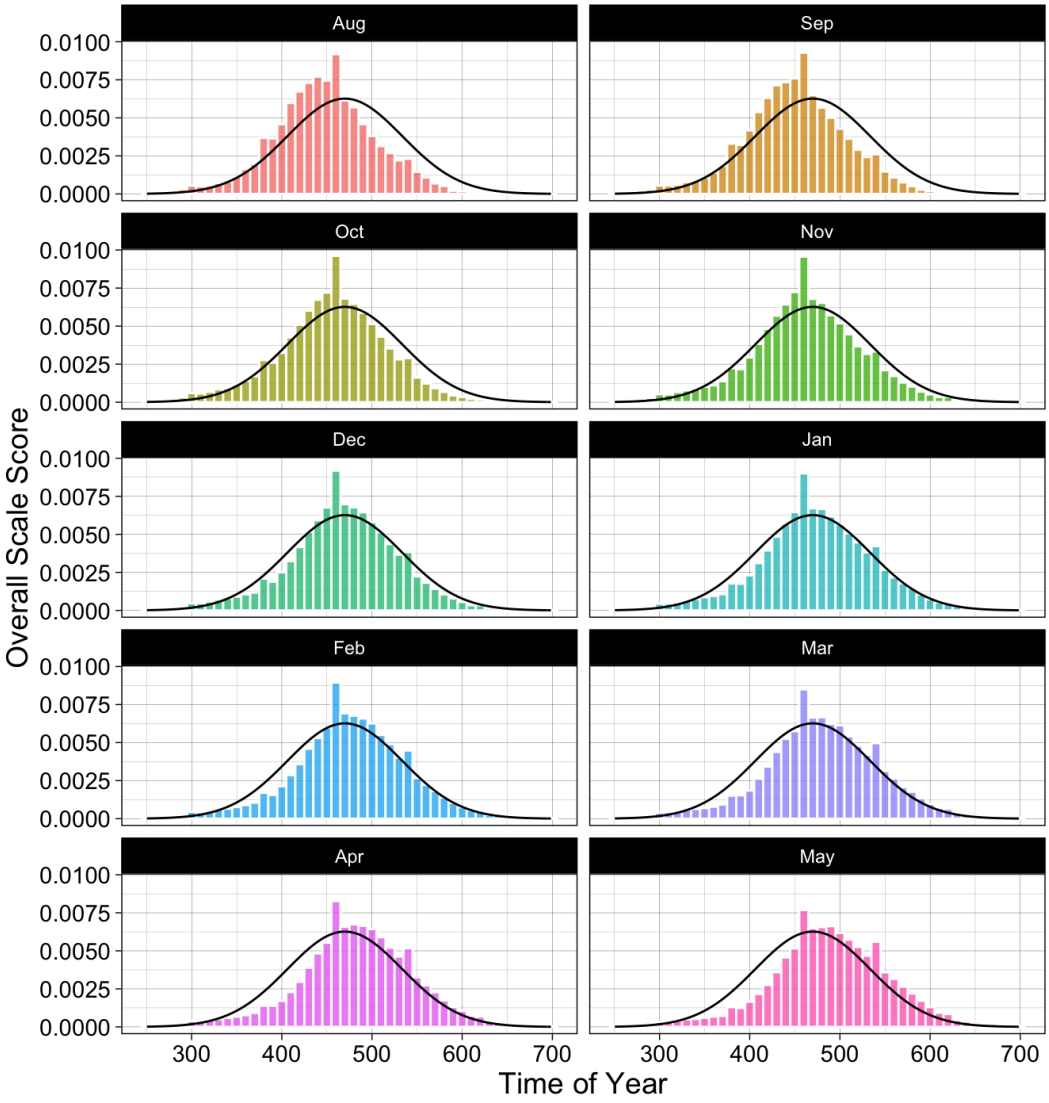

Figure 3.2. ISIP Reading grade 3 overall scale score normative samples by time of year with normal distribution overlay

Fitting the Taylor polynomials

R software 4.2.1 (R Core Team, 2021) was the environment for modeling the students' test score distributions. The R tidyverse package (Wickham, 2021) was used to prepare plots such as those seen in Figures 3.1 and 3.2.

The R package cNORM 3.1 (Lenhard et al., 2018) was used to transform the percentiles into normalized scores, calculate the powers of these locations (L) and of the grade-administration periods (A) and their products, fit the polynomial regressions using a stepwise procedure, and output the final regression equations and diagnostics. Selected output from this package for modeling of the ISIP Reading grade 3 overall scores is shown below in Figures 3.3 and 3.4.

```
User specified solution: 5 terms

R-Square Adj. = 0.997718

Final regression model: raw ~ L1 + L2 + L4 + A2 + L1A1

Regression function: raw ~ -120.3154071 + (20.87968902*L1) +
 (-0.2228828241*L2) + (1.296600841e-05*L4) + (-0.1218117621*A2) +
(0.1177085667*L1A1)

Raw Score RMSE = 3.14874
```

Figure 3.3. cNORM output of the polynomial regression for the ISIP Reading Grade 3 overall scale scores.

Figure 3.3 shows that the final regression model fitted to these data consisted of five terms and included an intercept term, powers of the normalized scores (L, L^2, and L^4), the grade-administration period squared (A^2), and the interaction between the normalized scores and the grade-administration period (LA). The *adjusted* R^2 is a goodness-of-fit measure that has been adjusted for number of predictors in the model. Here, the adjusted R^2 is greater than 0.99 and indicates that more than 99% of the variance in these data were captured by the model. The *root mean square error* (RMSE) is the standard deviation of the residuals of the model (i.e., of the prediction errors). While higher adjusted R^2 and lower RMSE values indicate better fit, we rarely used models with more terms than shown here to avoid overfitting the model and obtaining flawed results.

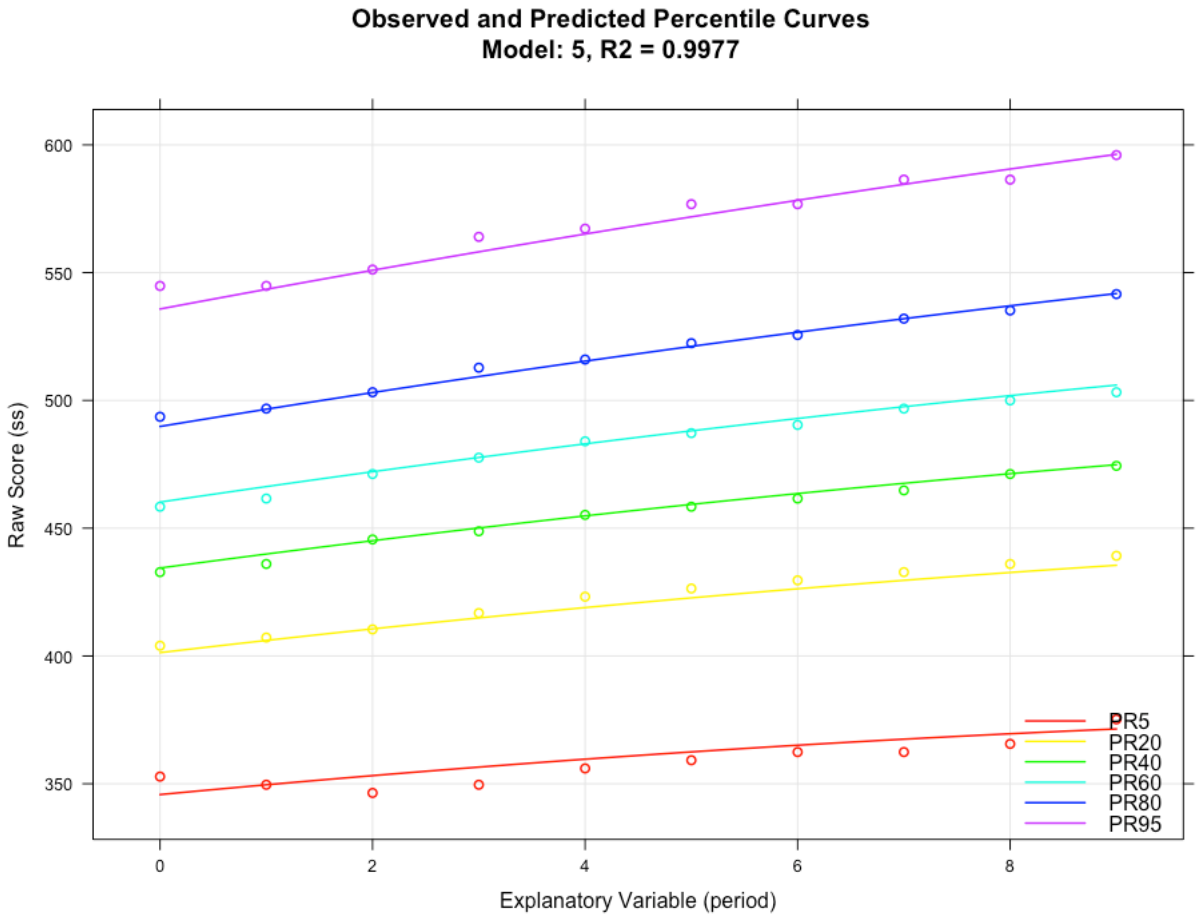

Figure 3.4. Observed and predicted percentile curves for the regression model fitted to the ISIP Reading overall grade 3 scale scores

The results of the model are shown in Figure 3.4. Here, the percentile rank (PR) curves for the 5th and 95th percentiles were plotted to provide a robust indication of the overall reading score distributions by period. The curves for PR20, PR40, PR60, and PR80 were plotted as these are the cut scores that are used to define the ISIP performance levels and because PR20 and PR40 are used to define the instructional tier goals. In general, the curves show a reasonable progression of the PRs across the periods with no anomalies such as reversals.

Generating and Reviewing the Percentiles for the New Norms

Once the polynomial regressions for all of the grades on a given test had been completed and reviewed, the regression equations were used to generate the corresponding scale score percentiles that would be used as norms for each percentile rank of the theoretical distribution (i.e., from 1 to 99). The final scale score percentiles for the new norms were rounded to the nearest whole numbers on the 100 to 900 ISIP Reading reporting scale.

The norms produced were reviewed in the following ways:

- Each set of norms was examined to see if they were monotonically increasing across the periods within a grade (i.e., there were no reversals of the percentiles across periods).
- The entire set of observed and predicted percentile curves for a test was examined to see if the progression of the norms across all the grades was reasonable.
- The newly produced reading norms were compared with the current set of norms to see their differences and were applied to current student data. Differences in the student achievement needed to reach a particular PR were noted, and additional checks examined the percentages of students that fell into the ISIP Reading levels and instructional tiers under the current norms versus the new norms.

Chapter 4: Growth

Introduction

Student achievement is typically derived from a score from a single test administration, whereas student growth can be conceptualized as change in academic performance over time. Monitoring growth can enrich the understanding of how well a student is performing. For example, growth may be used as a tool to promote accountability, inform data-based decision-making, and foster collaboration within and between schools and districts. Being able to monitor individual student growth allows educators to determine whether students — and correspondingly teachers and schools — are making adequate annual progress toward state or national standards. Furthermore, 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 ways 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 throughout the school year. The third method we provide is based on performance categories 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 reading exceeds the growth of 80% 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.

BOY scale scores that were collected from the 2018-2019 normed sample were divided into 10 initial status groups for ISIP Reading 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 4.1 to 4.4 show the growth that would be expected in ISIP Reading 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 4.1. *Normative Growth for ISIP Reading 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	62	61	36	32	23
11-20	2	65	66	43	39	27
21-30	3	66	70	48	42	30
31-40	4	68	70	52	44	32
41-50	5	70	71	56	44	33
51-60	6	71	70	59	44	35
61-70	7	73	70	62	44	37
71-80	8	74	69	65	43	40
81-90	9	75	67	67	42	44
91-99	10	75	76	64	50	53

Table 4.2. *Normative Growth for ISIP Reading 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	68	65	41	37	25
11-20	2	72	71	49	45	30
21-30	3	74	75	54	48	33
31-40	4	77	75	60	50	35
41-50	5	78	76	63	50	37
51-60	6	80	76	67	50	39
61-70	7	81	75	71	50	42
71-80	8	83	74	74	49	45
81-90	9	83	72	76	48	50
91-99	10	80	83	73	57	59

Table 4.3. *Normative Growth for ISIP Reading 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	22	15	18	18	16
11-20	2	24	18	22	20	18
21-30	3	26	18	23	22	20
31-40	4	27	19	25	23	22
41-50	5	29	21	26	24	22
51-60	6	30	22	27	25	23
61-70	7	32	23	28	26	24
71-80	8	34	24	28	27	25
81-90	9	36	26	30	28	26
91-99	10	42	30	31	29	27

Table 4.4. *Normative Growth for ISIP Reading 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	25	18	22	20	18
11-20	2	28	20	25	23	21
21-30	3	30	21	27	25	23
31-40	4	31	22	28	26	24
41-50	5	33	24	30	27	25
51-60	6	35	25	31	28	27
61-70	7	36	27	31	30	27
71-80	8	39	28	32	31	29
81-90	9	41	29	34	31	30
91-99	10	48	34	35	33	31

Normative growth can inform several education-related activities. Educators can use these growth resources to evaluate students' achievement and growth. They may also use these resources to guide individualized instruction and to aid in setting achievement and growth goals for students in a school. Normative growth provides an opportunity to support conversations about achievement patterns as educators can evaluate whether students made growth 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 characterizes student growth in terms of changes in performance level categories rather than evaluating changes in scale score points throughout the 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 (May). BOY and EOY scale scores that were collected from the 2018-2019 normed sample were divided into 10 initial status groups for the ISIP Reading 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 change in performance level categories from BOY to EOY for the ISIP Reading Overall score for pre-K to grade 8. For example, in the table 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 goes 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 moving up one or more decile levels from BOY to EOY.

Tables 4.5 to 4.14 illustrate the change in performance categories from BOY to EOY for the ISIP Reading Overall score. In general, students in the lower decile categories show growth by moving up a level or two between BOY and EOY. Students

who placed in the upper decile categories mostly stay 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 4.5. *Pre-K Change in Performance Categories BOY-EOY for ISIP Reading 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	37.88*	23.33	12.73	6.06	7.88	5.45	3.03	0.91	1.52	1.21
10-19	16.03	20.92*	15.49	8.7	11.96	9.51	7.61	4.89	2.72	2.17
20-29	9.68	15.77	13.62*	15.05	11.83	13.26	6.45	7.17	3.23	3.94
30-39	4.81	11.14	13.67	10.89*	13.42	7.34	13.16	9.37	9.11	7.09
40-49	2.89	9.92	9.92	14.46	11.98*	13.22	10.33	8.26	11.57	7.44
50-59	3.02	5.74	9.97	8.16	12.69	15.41*	12.99	10.27	11.18	10.57
60-69	2.09	5.07	6.57	8.06	12.54	12.54	13.73*	11.34	11.34	16.72
70-79	1.99	7.62	8.28	8.61	8.94	13.25	11.92	9.27*	15.89	14.24
80-89	1.33	3.67	3.67	4	8.67	11.33	12	10	20.33*	25
90-99	2.38	3.74	3.74	4.42	8.16	7.14	12.24	8.16	18.71	31.29*

Table 4.6. *Kindergarten Change in Performance Categories BOY-EOY for ISIP Reading 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	48.42*	22.88	15.08	5.41	4.49	1.73	0.9	0.56	0.26	0.26
10-19	22.1	23.52*	22.15	10.67	10.91	4.96	2.9	1.52	0.65	0.61
20-29	12.33	18.76	21.89*	13.55	14.26	8.2	5.34	3.01	1.57	1.08
30-39	7.45	13.31	19.17	13.26*	17.43	10.79	8.83	5.48	2.68	1.6
40-49	4.54	9.17	14.81	12.42	18.14*	12.65	12.16	8.73	4.84	2.54
50-59	2.95	6.39	11.12	10.63	16.72	13.41*	14.58	12.31	7.52	4.36
60-69	1.69	4.2	8.49	7.83	15.14	13.15	15.01*	15.87	11.94	6.68
70-79	1.08	2.68	5.61	6.38	11.8	12.01	15.85	17.53*	16.2	10.87
80-89	0.87	1.49	3.6	4.02	8.18	8.93	12.48	17.27	22.37*	20.79
90-99	1.04	1.78	2.64	2.57	4.96	5.53	8.58	12.86	20.98	39.06*

Table 4.7. *First Grade Change in Performance Categories BOY-EOY for ISIP Reading 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	66.22*	23.33	6.19	2.62	0.73	0.31	0.23	0.08	0.11	0.17
10-19	21.72	33.13*	21.3	13.58	5.69	2.95	0.9	0.27	0.25	0.21
20-29	8.14	22.33	22.35*	20.7	12.73	8.05	3.43	1.22	0.61	0.44
30-39	3.53	12.45	16.68	21.53*	16.67	15.64	7.62	3.35	1.75	0.77
40-49	1.9	7.42	11.24	17.1	16.24*	20.03	13.86	7.15	3.73	1.33
50-59	0.98	4.33	7.54	13.07	14.68	20.1*	17.42	12.34	7.29	2.25
60-69	0.58	2.48	4.55	9.1	10.92	18.93	20.26*	16.83	12.39	3.96
70-79	0.43	1.55	3	5.76	7.35	13.95	19.56	19.61*	20.6	8.18
80-89	0.27	0.79	1.32	2.79	3.99	8.5	13.76	19.1	30.5*	18.99
90-99	0.26	0.53	0.94	1.5	1.63	3.46	6.08	9.92	25.91	49.78*

Table 4.8. *Second Grade Change in Performance Categories BOY-EOY for ISIP Reading 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	66.31*	23.63	6.87	1.95	0.46	0.31	0.18	0.05	0.06	0.19
10-19	22.51	34.7*	23.49	11.35	4.26	1.88	0.87	0.5	0.18	0.26
20-29	6.92	22.3	26.4*	21.7	11.08	6.34	3.16	1.2	0.56	0.33
30-39	2.6	12.13	21.46	22.74*	16.59	12.57	7.16	3.3	0.86	0.6
40-49	1.04	5.86	13.28	20.76	19.06*	16.94	13.46	6.51	2.21	0.88
50-59	0.74	2.84	7.55	15.21	18.2	19.27*	17.57	11.76	4.96	1.91
60-69	0.26	1.09	3.94	9.44	13.19	18.08	21.34*	19.39	10.17	3.1
70-79	0.17	0.51	1.8	4.4	7.83	12.83	20.05	24.7*	20.19	7.53
80-89	0.17	0.37	0.85	1.91	3.55	6.59	13.08	21.76	32.11*	19.61
90-99	0.28	0.25	0.38	0.69	1.18	1.94	4.11	8.58	24.4	58.18*

Table 4.9. *Third Grade Change in Performance Categories BOY-EOY for ISIP Reading 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	61.36*	26.01	9.23	2.24	0.53	0.29	0.09	0.08	0.05	0.12
10-19	22.87	33.61*	26.61	10.34	3.64	1.73	0.5	0.23	0.21	0.26
20-29	8.05	22.04	29.82*	20.32	10.3	5.89	2.13	0.91	0.26	0.29
30-39	2.84	10.81	23.01	23*	17.19	13.66	5.87	2.46	0.65	0.5
40-49	1.1	4.69	14.16	19.76	19.18*	20.35	12.38	5.52	2.01	0.84
50-59	0.46	2.64	7.29	13.12	16.69	23.46*	19.12	11.49	4.61	1.11
60-69	0.25	0.92	3.99	7.94	11.94	21.46	23.33*	18.43	9.21	2.54
70-79	0.12	0.48	1.48	3.89	6.62	15.05	21.45	26.3*	19.37	5.23
80-89	0.08	0.19	0.58	1.28	2.65	6.85	14.07	25.04	32.83*	16.41
90-99	0.12	0.12	0.34	0.35	0.73	1.76	3.27	8.34	22.48	62.47*

Table 4.10. *Fourth Grade Change in Performance Categories BOY-EOY for ISIP Reading 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	67.89*	22.74	6.22	1.83	0.62	0.34	0.14	0.06	0.06	0.09
10-19	21.06	36.29*	24.1	11.13	4.39	1.64	0.65	0.45	0.17	0.14
20-29	6.94	23.26	26.6*	20.94	12.52	5.19	2.72	1.1	0.51	0.23
30-39	2.46	11.09	20.02	23.48*	19.9	11.75	6.68	3.08	1.11	0.42
40-49	0.95	4.71	12.41	19.99	22.42*	17.24	12.83	6.54	2.1	0.81
50-59	0.31	1.64	5.97	13.49	19.03	21.16*	19.43	11.81	5.33	1.82
60-69	0.18	0.88	2.51	6.86	13.54	19.49	23.17*	19.05	11.29	3.04
70-79	0.16	0.35	0.92	2.85	7.16	12.56	21.44	26.26*	21.52	6.78
80-89	0.08	0.2	0.36	0.83	2.6	6.08	14.18	23	32.83*	19.84
90-99	0.05	0.08	0.13	0.22	0.62	1.16	2.87	8.15	22.85	63.87*

Table 4.11. *Fifth Grade Change in Performance Categories BOY-EOY for ISIP Reading 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	70.13*	22.94	4.5	1.31	0.44	0.18	0.16	0.04	0.1	0.22
10-19	19.42	38.14*	23.69	12.04	3.96	1.45	0.78	0.15	0.19	0.17
20-29	5.76	22.89	27.91*	22.89	11.34	5.36	2.23	1.03	0.25	0.34
30-39	2.19	11.26	20.51	26.1*	19.14	11.66	5.56	2.37	0.76	0.44
40-49	0.98	4.32	11.62	20.38	23.56*	18.54	12.27	5.3	2.12	0.9
50-59	0.48	1.67	5.36	12.34	19.78	21.57*	20.39	12.4	4.8	1.2
60-69	0.35	0.65	2.17	6.23	12.02	20.43	24.57*	19.49	11.21	2.88
70-79	0.27	0.31	0.69	2.72	6.54	12.08	21.61	27.03*	22.16	6.58
80-89	0.1	0.18	0.34	0.96	1.87	4.94	11.63	23.71	36.18*	20.1
90-99	0.04	0.2	0.12	0.38	0.54	1.05	2.34	6.9	22.98	65.44*

Table 4.12. *Sixth Grade Change in Performance Categories BOY-EOY for ISIP Reading 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	69.97*	21.59	5.73	1.47	0.59	0.15	0.29	0.07	0.07	0.07
10-19	19.56	37.73*	23.52	11.94	4.03	1.61	0.81	0.22	0.22	0.37
20-29	5.28	23.2	28.85*	22.45	11.67	4.68	1.93	1.12	0.37	0.45
30-39	2.64	9.83	19.81	23.77*	19.08	13.35	7.12	3.01	0.66	0.73
40-49	1.21	3.8	11.47	19.29	22.78*	18.07	12.98	6.61	2.73	1.06
50-59	0.45	2.15	6.24	11.07	18.13	23.63*	20.88	11	5.35	1.11
60-69	0.38	0.83	2.49	7.33	12.61	19.18	23.34*	19.18	11.18	3.47
70-79	0.22	0.79	1.08	2.8	5.6	11.85	22.92	26.08*	22.56	6.11
80-89	0.15	0.08	0.38	0.98	1.81	5.82	11.71	24.55	31.42*	23.11
90-99	0	0.23	0.53	0.3	0.99	1.14	2.44	6.47	23.46	64.43*

Table 4.13. *Seventh Grade Change in Performance Categories BOY-EOY for ISIP Reading 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	63.64*	25.96	6.56	2.36	0.44	0.26	0	0.09	0.09	0.61
10-19	20.95	35.48*	25.35	11.8	4.58	0.88	0.09	0.18	0	0.7
20-29	6.82	20.9	29.62*	22.37	12	4.4	2.25	0.95	0.26	0.43
30-39	3	9.89	17.76	25.88*	21.29	13.6	5.39	1.33	0.97	0.88
40-49	1.4	3.78	8.87	20.54	24.41*	19.49	13.52	5.27	1.58	1.14
50-59	1.21	1.31	4.39	9.43	20.26	23.9*	20.35	12.89	4.95	1.31
60-69	1.16	1.34	2.4	5.97	11.31	20.39	23.86*	19.41	11.22	2.94
70-79	0.62	1.33	1.15	2.3	5.48	12.02	23.17	27.94*	20.25	5.75
80-89	0.27	0.18	0.53	0.71	1.42	3.36	9.65	26.28	39.47*	18.14
90-99	0.27	0	0.27	0.18	0.45	1.35	1.9	6.68	21.57	67.33*

Table 4.14. *Eighth Grade Change in Performance Categories BOY-EOY for ISIP Reading 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	54.56*	33.92	9.44	1.44	0.16	0.16	0.16	0	0	0.16
10-19	19.91	32.13*	32.29	10.82	3.45	0.31	0.47	0.16	0.31	0.16
20-29	12.73	18.17	27.33*	24.38	10.4	3.88	0.93	0.47	1.09	0.62
30-39	4.52	11.47	12.6	26.49*	22.94	12.6	5.01	2.58	0.81	0.97
40-49	2.61	2.61	6.86	17.97	25.98*	22.22	12.42	5.07	2.12	2.12
50-59	2.02	1.86	3.72	11.47	18.14	24.34*	23.41	9.46	3.1	2.48
60-69	1.64	0.65	0.82	4.42	10.97	19.97	25.86*	22.91	8.18	4.58
70-79	0.81	0.98	1.63	2.44	4.56	11.73	24.27	27.36*	21.17	5.05
80-89	0.79	0.63	1.11	1.27	1.42	3.48	9.18	24.05	38.61*	19.46
90-99	0	0	0.48	0.65	0.97	0.65	1.29	6.95	25.2	63.81*

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. For example, 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 continue 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 scenario, the transition matrix functions as a coarse trajectory model, where an increase in one decile category is extrapolated and assumed to continue 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 compare the reading 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 objectives 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 Reading 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 4.15 summarizes the growth descriptions.

Table 4.15. *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 4.16 illustrates these expected growth pathways within each BOY instructional group for ISIP Reading Overall scores. 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 4.16. *Expected Growth Pathways (Gains) for ISIP Reading 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**	**Grade 7**	**Grade 8**
<21st	Below Typical	<89	<84	<52	<41	<35	<33	<24	<16	<12	<10
	Typical	89-114	84-105	52-70	41-63	35-50	33-49	24-38	16-32	12-28	10-29
	Above Typical	115-143	106-131	71-98	64-85	51-70	50-68	39-58	33-52	29-51	30-53
	Accelerated	≥144	≥132	≥99	≥86	≥71	≥69	≥59	≥53	≥52	≥54
21st - 40th	Below Typical	<68	<77	<58	<44	<32	<29	<19	<13	<14	<11
	Typical	68-86	77-92	58-73	44-57	32-47	29-43	19-32	13-26	14-28	11-27
	Above Typical	87-104	93-111	74-95	58-73	48-63	44-59	33-49	27-43	29-47	28-49
	Accelerated	≥105	≥112	≥96	≥74	≥64	≥60	≥50	≥44	≥48	≥50
41st - 60th	Below Typical	<60	<70	<60	<39	<32	<27	<17	<13	<14	<10
	Typical	60-76	70-83	60-73	39-51	32-44	27-40	17-30	13-26	14-27	10-25
	Above Typical	77-91	84-102	74-91	52-69	45-63	41-56	31-47	27-43	28-45	26-47
	Accelerated	≥92	≥103	≥92	≥70	≥64	≥57	≥48	≥44	≥46	≥48
61st - 80th	Below Typical	<54	<64	<58	<38	<32	<26	<17	<11	<15	<12
	Typical	54-69	64-79	58-70	38-51	32-47	26-38	17-29	11-25	15-29	12-29
	Above Typical	70-89	80-98	71-88	52-69	48-66	39-54	30-44	26-41	30-47	30-55
	Accelerated	≥90	≥99	≥89	≥70	≥67	≥55	≥45	≥42	≥48	≥56
>80th	Below Typical	<48	<55	<57	<36	<23	<15	<15	<13	<12	<9
	Typical	48-60	55-70	57-69	36-50	23-37	15-28	15-28	13-26	12-26	9-23
	Above Typical	61-82	71-89	70-86	51-67	38-51	29-46	29-46	27-44	27-44	24-45
	Accelerated	≥83	≥90	≥87	≥68	≥52	≥47	≥47	≥45	≥45	≥46

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.

Chapter 5: Reliability and Validity

Reliability and validity are two important qualities of any assessment. Reliability is the consistency of items within a test event or a comparison of scores from multiple test events. Validity can be thought of as how accurate an assessment is: either the accuracy of the content or the constructs being measured.

During the development of the ISIP Reading assessment, extensive research was conducted to assess reliability and validity. Those interested in greater detail regarding the development of the assessment and the reliability and validity research that was conducted may consult the technical manuals of Mathes et al., (2016) and Mathes (2016). This chapter will consolidate some of the information from that original work; however, most of this chapter will focus on research conducted since the last renorming.

Reliability of measures refers to the accuracy, consistency, and stability of obtained assessment scores across conditions (Anastasi & Urbina, 1997). Reliability is a necessary but insufficient piece of evidence to support the validity of test score interpretation. From a classical test theory perspective, the observed assessment score can be conceptualized as a hypothesized true score that a student would receive if the assessment would be perfectly reliable. The difference between a hypothetical true score and a student's observed assessment score can be attributed to measurement error. Assessments are considered reliable if evidence is given that the assessment produces relatively small measurement errors in conjunction with consistent measurement results.

Reliability is a particularly important property to consider when interpreting students' assessment scores from multiple administration periods. There are various procedures to estimate assessment score reliability. The most appropriate procedure is dictated by the intended use of the assessment results. Therefore, Istation provides different types of reliability estimates that were designed to address various sources of measurement error: Test-retest reliability from Classical test theory (CTT), item response theory (IRT) Marginal reliability, IRT Reliability, and decision consistency.

Test-retest reliability examines how dependably students respond to the assessment over different administration intervals. In this context, the measurement errors of primary interest are the fluctuation of students' observed scores around the hypothetical true score due to temporary changes in the students' setting (Crocker & Algina, 2008). For marginal reliability, the sources of error of greatest concern arise

from the items and item sampling within the computer-adaptive environment. The theoretical treatment of this source of error is approached by internal consistency indices and measurement error (Green et al., 1984). Marginal reliability denotes the difference between the student's score and the variance with their estimated latent abilities. IRT reliability is somewhat different as it denotes the ratio of the true score variance to the total variance with respect to sum scores (Andersson & Xin, 2018).

Decision accuracy is another measure of the assessment quality as it relates to the precision of the decisions made on the basis of those observed assessment scores, which are distinct from the consistency of decisions made by repeated testing. To evaluate the degree of decision accuracy, indices such as probabilities of occurrences of false-positive and false-negative outcomes are examined. Decision consistency shows how accurate and stable an assessment is for determining classifications of students, and it is useful for making sure that an assessment consistently identifies students who are either struggling and may need intervention or are on track to meet grade-level expectations (LaFond, 2014).

Evidence of Reliability

Test-Retest Stability

This type of evidence allows one to examine how consistently students respond to the assessment over different occasions. In this situation the measurement errors of primary interest are the fluctuation of students' observed scores around the hypothetical true score due to temporary changes in the students' environment. To estimate the impact of such errors on assessment score reliability, evidence of test-retest stability was obtained for a subset of the normed sample. Students who tested twice within a time interval of 2-21 days in the middle of the year were selected in this study. This time interval was chosen to mitigate practice effects and/or maturational or historical changes in a student's true score.

Results for Test-Retest

Separate analyses were done by grades. See chapter 3 for a demographic breakdown of the norming sample. Sample sizes for these subsets ranged from 457 to 19,553 per grade. Test-retest reliability was estimated using Pearson's product-moment

correlations. The mean and standard deviations (SD) across each testing occasion are presented in Table 5.1.

The extent to which a sample of students performs consistently on the same assessment across multiple occasions is an indication of test-retest reliability. The data show strong evidence for test-retest reliability with coefficients ranging from .67 to .89.

Table 5.1. *Means and Standard Deviations (SD) and Reliability Estimates by Grade*

Grade	First Administration Mean	First Administration SD	Second Administration Mean	Second Administration SD	*r*
Prekindergarten	245	36	250	39	.67***
Kindergarten	305	43	311	43	.81***
1	359	48	365	50	.89***
2	420	58	425	58	.89***
3	465	58	468	58	.88***
4	506	57	511	58	.86***
5	535	61	538	63	.86***
6	553	60	555	64	.86***
7	590	67	591	68	.86***
8	624	73	626	76	.86***

*Note: *** correlation is significant at the 0.001 level (2-tailed)*

Marginal Reliability

Because ISIP Reading uses IRT as its method of implementation, reliability takes on a different meaning than it does in a CTT perspective. The biggest difference between the two approaches is their assumption about the measurement error related to the test scores obtained from the measure. CTT treats the error variance as being the same for all scores, whereas IRT views the level of error as dependent on the ability of the individual — as such the error variance is expressed as a function of the latent construct ability (θ).

In IRT it is possible to determine a generic estimate of reliability, known as marginal reliability (Sireci et al., 1991) with:

$$\bar{\rho} = \frac{\sigma_{\theta}^2 - \bar{\sigma}_{e*}^2}{\sigma_{\theta}^2},$$

where σ_{θ}^2 is the variance of performance score for the norming sample and $\bar{\sigma}_{e*}^2$ is the mean-squared error. More specifically the value of $\bar{\sigma}_{e*}^2$ is the average of the possible values of the error variance. For example, if several values of $\bar{\sigma}_{e*}^2$ were tabulated in a row for various levels of ability (θ), $\bar{\sigma}_{e*}^2$ would reflect the "marginal" average for that row. Hence the reliability that is derived from that marginal error variance is called the marginal reliability and denoted as $\bar{\rho}$ to indicate that it is an average. The construction of marginal reliability can be thought of as an analogue to internal consistency estimates of reliability for traditional test scores that are derived based on CTT. Similar to Cronbach's alpha, marginal reliability is a unitless measure confined by 0 and 1 and thus can be used as an index to directly compare the internal consistencies of classical test data to IRT-based test data.

Measurement Error

Istation used simulation studies that examined the entire CAT system — including the item pool, the item selection algorithm, and the item exposure control system — to evaluate the theoretical precision of CAT. The true values of θ are known for each simulated observation, and thus it is straightforward to compute the association between the known values of θ and those estimates that were produced by testing the CAT system. These coefficients can be thought of as associations between observed test scores with true scores as is referred to in CTT. Furthermore, Istation examined errors of estimation, their variance, and the corresponding values of the information function at each level of the θ level. Based on these results, CAT stopping criteria are based on minimizing the standard error of the ability estimate ($\sigma_{e*} = 0.3$) for each examinee. Because CAT has been constructed so that precision of all test scores is approximately equal, the lower limit of the marginal reliability of the data for any administration will always be around $\rho = 0.90$ which supports a high level of internal consistency.

This is due to the relationships between the variances, standard errors, and reliability which all manifest information regarding measurement precision. For example, the variance of measurement and the variance of estimation within the IRT context can be conceptualized as follows:

$$\sigma_e^2 = \sigma_{\theta}^2[1 - \rho]$$

and

$$\sigma_{e*}^2 = [1 - \rho].$$

Then it follows that if $\rho = 0.90$, then σ_{e*}^2=0.1 and $\sigma_{e*} \approx 0.3$. That is under the assumption that the IRT error variance (information function) is uniform over the range of θ (in this case it would correspond to $I(\theta) = 10$).

IRT Reliability

We also used our norming samples in pre-kindergarten through eighth grade to compute the IRT-based reliability on the Overall scores at the middle-of-the-year (MOY) benchmarking assessment month. While test-retest is based more on CTT, the IRT reliability provides information for a CAT.

We derived IRT-based reliability from the CTT to IRT using the formulas below (Andersson & Xin, 2018):

$$X = T + E$$

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2$$

$$\rho_{xx'} = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_X^2 - \sigma_E^2}{\sigma_X^2} = 1 - \frac{\sigma_E^2}{\sigma_X^2}$$

If X (i.e., θ in IRT) is standardized: $\rho_{xx'} = 1 - \sigma_E^2$

Since $\sigma_E^2 = SE(\theta)^2$ in IRT,

$$\rho_{xx'} = 1 - SE(\theta)^2 \qquad (1)$$

Results for IRT Reliability

Table 5.2 shows the IRT-based reliability results for prekindergarten through eighth grade derived from the formulas above. The results show strong reliability across grades, ranging from 0.90 to 0.98. The reliabilities are slightly higher in lower than in higher grades.

Table 5.2. *IRT Based Reliability at the Middle of the Year*

Grade	Sample	Reliability
Prekindergarten	9,000	0.96
Kindergarten	127,000	0.96
1	163,000	0.98
2	150,000	0.96
3	125,000	0.95
4	100,000	0.92
5	80,000	0.92
6	25,000	0.92
7	25,000	0.91
8	8,000	0.90

Decision Consistency

Decision consistency describes the degree to which test takers are re-classified into the same category over parallel replications. It is a form of reliability that measures the reliability of an assessment across multiple test events (Tomek, 2018). The purpose of this study is to conduct the decision consistency analyses for students in kindergarten to eighth grade who took ISIP Reading for fall, winter, and spring benchmarking assessment months.

Methodology for Decision Consistency

For this study, we used data from the extensive Istation database. Data from students who enrolled in the Istation Reading program in kindergarten to eighth grade in the 2018-2019 and 2021-2022 school years were selected. We eliminated students who did not have assessments in the beginning of the year (BOY), middle of the year (MOY), and end of the year (EOY). We used September for BOY, January for MOY, and May for EOY.

To achieve a representative sample, we then applied post-stratification using the SI variable at the school level as described in chapter 3. A final sample consisted of 537,000 students in the 2018-2019 school year and 453,000 students in the 2021-2022 school year.

We used the Swaminathan-Hambleton-Algina Method to calculate the decision consistency across grades. This method was developed by Hambleton and Novick (1973) and Swaminathan et al. (1974). The decision consistency rates range from 0 to 1.0. The 0 indicates no consistency, and the 1.0 shows perfect consistency between the two test administrations.

Once a student takes ISIP Reading, the scale scores and the percentile rank are provided. A student is classified into one of the five instructional levels depending on their ability on ISIP Reading:

- Level 1: at or below the 20th percentile rank
- Level 2: from the 21st to the 40th percentile rank
- Level 3: from the 41st to the 60th percentile rank
- Level 4: from the 61st to the 80th percentile rank
- Level 5: from the 81st percentile rank and above

We established four cut points: P20, P40, P60, and P80. P20 is the 20th percentile rank cut point, and it classifies students into two categories: Level 1 students vs. Levels 2-5 students. P40 is the 40th percentile rank cut point that classifies students into the Levels 1 and 2 category vs. Levels 3-5 category. P60 is the 60th percentile rank cut point that classifies students into Levels 1-3 vs. Levels 4 and 5 categories. Finally, the P80 is the 80th percentile rank cut point that classifies students into Levels 1-4 vs. Level 5 categories.

Results from Decision Consistency Analysis

We first computed the Pearson product-moment correlations of overall scores between fall and winter benchmarking assessment months and winter and spring benchmarking assessment months. The results are in Tables 5.3 and 5.4. The correlations ranged from 0.709 to 0.899, indicating very high relationships of overall scores between the BOY to MOY and MOY to EOY for both school years.

We then computed the decision consistency for all cut points by grade level for both school years, and the results are in Table 5.4. The decision consistency rates range from 0.782 to 0.905, indicating that students remain in the same category between 78% and 91% of the time from BOY to MOY. For MOY to EOY the decision consistency rates range from 0.848 to 0.925, meaning that students remain in the same category between 86% to 92% of the time. On average, 86% of students stay in the same category from BOY to EOY, and 88% remain in the same category from MOY to EOY.

Table 5.3. *Pearson Product-Moment Correlations of Overall Scores between Fall and Winter Benchmarking Assessment Months and Winter and Spring Benchmarking Assessment Months*

Grade	N 2018-2019	Fall-Winter 2018-2019	Winter-Spring 2018-2019	N 2021-2022	Fall-Winter 2021-2022	Winter-Spring 2021-2022
Kindergarten	97,000	0.713**	0.783**	60,000	0.709**	0.798**
1	100,000	0.837**	0.871**	70,000	0.842**	0.875**
2	100,000	0.845**	0.850**	80,000	0.870**	0.881**
3	80,000	0.871**	0.873**	80,000	0.876**	0.880**
4	50,000	0.884**	0.888**	60,000	0.877**	0.886**
5	50,000	0.899**	0.891**	60,000	0.890**	0.884**
6	13,000	0.892**	0.871**	20,000	0.884**	0.872**
7	11,000	0.887**	0.857**	15,000	0.863**	0.848**
8	6,000	0.881**	0.843**	8,000	0.854**	0.842**

Note: ** correlation is significant at the 0.01 level (2-tailed)

Table 5.4. *Decision Consistency by Grade Level for the 2018-2019 and 2021-2022 School Years*

Grade	Cut Point	Fall-Winter 2018-2019	Winter-Spring 2018-2019	Fall-Winter 2021-2022	Winter-Spring 2021-2022
Kindergarten	P20	0.843	0.898	0.807	0.888
	P40	0.785	0.856	0.782	0.860
	P60	0.787	0.863	0.810	0.868
	P80	0.849	0.904	0.876	0.906
1	P20	0.879	0.905	0.856	0.886
	P40	0.824	0.866	0.825	0.860
	P60	0.815	0.866	0.830	0.873
	P80	0.851	0.907	0.873	0.913
2	P20	0.872	0.905	0.851	0.893
	P40	0.824	0.865	0.830	0.862
	P60	0.840	0.867	0.858	0.869
	P80	0.892	0.909	0.906	0.910
3	P20	0.901	0.909	0.884	0.885
	P40	0.860	0.861	0.861	0.866
	P60	0.857	0.865	0.871	0.875
	P80	0.904	0.908	0.917	0.908

4	P20	0.894	0.895	0.881	0.876
	P40	0.849	0.851	0.857	0.861
	P60	0.851	0.852	0.864	0.869
	P80	0.899	0.905	0.902	0.905
5	P20	0.906	0.924	0.899	0.873
	P40	0.846	0.877	0.863	0.862
	P60	0.853	0.864	0.868	0.880
	P80	0.902	0.887	0.902	0.923
6	P20	0.899	0.906	0.892	0.866
	P40	0.850	0.859	0.852	0.854
	P60	0.860	0.848	0.866	0.874
	P80	0.895	0.884	0.894	0.868
7	P20	0.903	0.921	0.904	0.915
	P40	0.858	0.870	0.864	0.872
	P60	0.864	0.868	0.863	0.862
	P80	0.896	0.893	0.890	0.878
8	P20	0.890	0.900	0.875	0.925
	P40	0.851	0.853	0.848	0.868
	P60	0.860	0.860	0.863	0.848
	P80	0.901	0.914	0.905	0.915

Evidence of Validity

The following paragraphs summarize the validity evidence from the prior technical manuals, and then we present updated information.

Construct Validity

Construct validity demonstrates that the assessment measures what it is supposed to measure; in this instance, ISIP Reading includes reading domains that meaningfully predict reading proficiency. ISIP Reading was built upon Dr. Torgesen's prior work in developing the Comprehensive Test of Phonological Processing (CTOPP: Wagner et al., 1999) and the Test of Word Reading Efficiency (TOWRE: Torgesen et al., 1999). Early development was also informed by the work from the National Reading Panel (2000) that recommended specific reading domains. The test authors consulted state test standards and the recommendations from the National Reading Panel to

compose all of the subtests of ISIP Reading. Full information regarding the assessment's development can be found in the technical manuals.

Concurrent Validity

Relationships between test scores and other measures give us information of convergent evidence that demonstrates the assessments measure similar constructs (APA, AERA, NCME Test Standards, 2014). Evidence of concurrent validity was established using several measures, including the Texas Primary Reading Inventory, the Letter Naming and Letter Sounds assessments, DIBELS®, the CTOPP, the TOWRE, Woodcock Language Proficiency Battery-Revised (WLPB-R),the Gray Oral Reading Test 4th Edition (GORT-4™), Woodcock-Johnson® III (WJ-III), Wechsler Individual Achievement Test® second edition (WIAT-II), the Iowa Test of Basic Skills™ (ITBS), and the Peabody Picture Vocabulary Test 4th Edition (PPVT™-IV). Cohen (1988) asserted that correlations at .30 are moderate, and those at .50 and greater are large. Those at the upper range of Cohen's scale at .70 are very large, and those at .90 are nearly perfect. For all grades, data were collected by the authors and their team of researchers and graduate students at Southern Methodist University.

Prekindergarten

Students in prekindergarten took the ISIP Reading and subtests from the English Language Skills Assessment (ELSA), Test of Preschool Early Literacy (TOPEL) Letter Name and Letter Sound subtests, and PPVT-4. Correlations for ISIP Letter Knowledge (LK) and the ELSA subtests ranged from .636 to .747. The correlation with ELSA Letter Names was .727, and ELSA Letter Sounds was .669. The correlation with ISIP LK and the TOPEL Print Knowledge score was .735. ISIP Vocabulary correlated with the PPVT-4 at .625, and the TOPEL Definitional Vocabulary at .520. ISIP Phonemic Awareness (PA) correlated with the ELSA subtests at .485 to .620. The correlation with the TOPEL Phonological Awareness score was .242. ISIP Overall reading correlated with the TOPEL total score at .677 (Mathes et al., 2016).

Kindergarten to Grade 3

The ISIP LK had correlations with Letter Names at .593, .693 with Letter Sounds, and .711 with the WLPB-R Letter Word ID. ISIP PA had correlations of .62 to .70 with subtests from the CTOPP. ISIP Alphabetic Decoding had correlations with the WLPB-R

Word Attack (.830), and TOWRE subtests at.811 and .838, and with WIAT-II Target Words at .589 (Mathes et al., 2016).

ISIP Spelling had correlations with the WJ-III Spelling at .890 and WIAT-II Spelling at .875. Text Fluency correlated with DIBELS ORF at .766, and ISIP Comprehension correlated with GORT-4 Comprehension at .621, WLPB-R Comprehension at .794, and WIAT-II Comprehension at .682. ISIP Vocabulary had a correlation with the PPVT-III Vocabulary at .814, and with WLPB-R Vocabulary at .836 (Mathes et al., 2016). ISIP ORF had correlations with the DIBELS 8 words correct per minute at .89 and with DIBELS 8 accuracy at .83 (Istation, 2020).

For the Overall score, there were correlations of .829 to .895 with the ITBS Reading scale score. Correlations with the Texas Assessment of Knowledge and Skills (TAKS) and Overall score was .695 to .741 (Mathes et al., 2016).

Grades 4 to 8

Correlations for grades 4 to 8 are similarly strong to very strong. ISIP Spelling correlations with WIAT-II Spelling ranged from .730–.835, and with WJ-III Spelling correlations ranged from .71–.849. ISIP Comprehension correlates with the GORT -4 Comprehension at .345–.471, with GORT-4 Fluency at .424–.640, and with WIAT-II Comprehension at .482–.565. ISIP Vocabulary correlates with the PPVT-IV at .52–.693. ISIP Text Fluency correlates with GORT-4 Fluency at .547–.631 (Mathes, 2016).

Evidence of Validity: Updated Research

Since the last norms update, the ISIP has been correlated with several other assessments, both formative and summative. The following section summarizes this research and gives evidence for correlational relationships, followed by linking studies that use a multinomial logistic regression model. Correlational studies demonstrate evidence of validity with other reading assessments, and the linking studies provide a projection of a student's proficiency level on state or other summative-type assessments.

We used cross-sectional Pearson product-moment correlations to establish relationships between ISIP Reading and other assessments in reading. Data were obtained from research partners throughout the United States, and we conducted studies for the Georgia Milestones (Patarapichayatham, 2016), Kansas Assessment Program (KAP) (Patarapichayatham, 2017), Virginia Standards of Learning (SOL)

(Patarapichayatham, 2018; Campbell et al., 2019), Colorado Measures of Academic Success (CMAS) (Patarapichayatham, 2019), California Smarter Balanced (Patarapichayatham & Wolf, 2022a), South Carolina Ready (SC Ready) (Cook & Ross, 2020a), Texas STAAR (Patarapichayatham et al., 2014; Patarapichayatham & Locke, 2020a; Patarapichayatham & Wolf, 2022), New Jersey Student Learning Assessment (NJSLA-ELA) (Wolf & Patarapichayatham, 2022), the Idaho Standards Achievement Test (ISAT) (Wolf et al., 2020a; Cook & Ross, 2022), the Renaissance STAR (Campbell et al., 2019; Sutter, et al., 2020), Ohio AIR (Patarapichayatham & Locke 2020b), and North West Education Association Measures of Academic Progress (NWEA MAP) (Cook & Ross, 2020c; Patarapichayatham & Wolf, 2022b).

We also have longitudinal Pearson product-moment correlations between ISIP Reading and Idaho ISAT (Wolf 2020b), Arkansas ACT Aspire (Patarapichayatham & Locke, 2020c), Texas STAAR (Patarapichyatham & Locke, 2020a), and New Mexico PARCC (Cook & Ross, 2020b) in the lower elementary grades. In the longitudinal correlational study, students had their ISIP scores in their kindergarten, first, second, or third grade, correlated with their state test scores in reading in their third grade. These correlations helped establish that ISIP can be used to identify students at risk for reading difficulties earlier in their elementary school years, giving teachers more time to help them meet expectations by third grade.

Tables 5.5 through 5.7 show the cross-sectional Pearson product-moment correlation coefficients between ISIP Reading and other reading measures from these studies. For full information including demographics and linking information, please consult the Istation website at www.istation.com/studies. Overall, the correlations between ISIP Reading and other measures show a strong relationship, ranging from 0.55 to 0.83. This indicates that if students do well on ISIP Reading, it is very likely that they will do well on these measures.

Table 5.8 shows the longitudinal Pearson product-moment correlation coefficients between ISIP Reading and Idaho ISAT, Arkansas ACT Aspire, Texas STAAR, and New Mexico PARCC Reading measures. We looked at the correlations among students' ISIP Reading scores in their kindergarten, first, second, and third grade and their reading state test scores (STAAR or PARCC) in their third grade. The correlations range from 0.49 to 0.79, indicating a relatively strong relationship between ISIP Reading and these measures. These results indicate that teachers and school districts could use ISIP Reading scores to monitor and predict their students' performance in first and second grade and determine whether they will do well when they take the summative assessment in third grade.

Table 5.5. *Pearson Product-Moment Correlation Coefficients for State Assessments 2017-2019*

Grade	Texas STAAR 2017	Georgia Milestones 2017	Kansas KAP 2017	Virginia SOL 2017	Colorado CMAS 2017	Virginia SOL 2019	California SBAC 2019	SC Ready 2019	Ohio AIR 2019	Idaho ISAT 2019
3	.74	0.77	0.76	0.72	0.75	0.75	0.75	0.71	0.66	.74
4	.74	0.80	0.74	0.76	0.79	0.73	0.72	0.74	0.63	
5	.71	0.70	0.77	0.74	0.82	0.69	0.60		0.62	
6	.75	0.78	0.75			0.67	0.69		0.68	
7	.55					0.78			0.61	
8	.63					0.47			0.70	

Table 5.6. *Pearson Product-Moment Correlation Coefficients for 2022 Assessments*

Grade	NJSLA 2022	Texas STAAR 2022	NWEA MAP 2022
K			0.59
1			0.74
2			0.78
3	.68	0.71	0.82
4	.68	0.71	0.83
5	.67	0.73	0.79
6	.69	0.67	0.78
7	.71	0.71	0.76
8	.69	0.62	0.76

Table 5.7. *Pearson Product-Moment Correlation Coefficients for ISIP Reading and STAR Reading*

Grade	Renaissance STAR
K–2	.83

Table 5.8. *Longitudinal Pearson Product-Moment Correlation Coefficients between ISIP Reading and Other Reading Measures*

Grade and Benchmark	Idaho ISAT 2019	Arkansas ACT 2019	Texas STAAR 2019	New Mexico PARCC 2019
K Winter			0.49	
K Spring			0.55	
1 Winter			0.67	
1 Spring			0.69	0.79
2 Fall	0.70			
2 Winter	0.71		0.73	
2 Spring	0.71	0.71	0.72	0.78
3 Fall	0.73			0.71
3 Winter	0.74	0.73	0.73	0.74
3 Spring	NA	NA	0.72	NA

Evidence for ISIP as a Dyslexia Screener

ISIP Reading assesses skills that are associated with a risk of dyslexia, and it is an approved dyslexia screener in several states including Washington, Indiana, Arkansas, Oklahoma, Texas, and Georgia. Dyslexia is a neurological variation that affects how a person processes language and sound, and it is often an inherited trait (International Dyslexia Association, 2019). Typically, people with dyslexia will have difficulty with the alphabet, phonics, spelling, and rapid naming, and these can result in difficulties in reading comprehension.

To determine evidence of validity with other dyslexia screeners, we collected data for ISIP Reading, ISIP RAN, and the Wechsler Individual Achievement Test | Fourth Edition (WIAT-4) dyslexia index (Wechsler, 2020), as well as the object naming fluency

(ONF) and letter naming fluency (LNF) subtests from the Kaufman Test of Educational Achievement Third Edition (KTEA™-3) (Kaufman & Kaufman, 2014). Data were collected on 199 students in the spring of 2022, 49% of whom were male and 51% female. Students were from all races/ethnicities, including approximately 16% African American or Black, 12% of Hispanic origin, 4% Asian/other race/ethnicity, and 68% white. Data were collected by a qualified clinical psychologist in 16 states. All students were unfamiliar with ISIP Reading, ISIP RAN, and the validity instruments. Results from ISIP Reading and the WIAT-4 are in table 5.9, and the results from the ISIP RAN and the WIAT-4 Dyslexia Index and KTEA-3 subtests are in table 5.10. Full technical details regarding the ISIP RAN can be found in the technical report for that assessment.

Table 5.9. *Correlations with ISIP Reading and the WIAT-4 Dyslexia Screener*

Grade	ISIP Reading Subtest	WIAT-4 Dyslexia Index	WIAT-4 Phonemic Proficiency	WIAT-4 Word Reading	WIAT-4 Pseudoword Decoding
K	**Overall Score**	**.82*** *N* = 50**	**.78*** *N* = 50**	**.71*** *N* = 50**	
	Letter Knowledge	.68*** *N* = 50	.52*** *N* = 50	.70*** *N* = 50	
	Phonemic Awareness	.72*** *N* = 50	.73*** *N* = 50	.59*** *N* = 50	
1	**Overall Score**	**.84*** *N* = 49**	**.69*** *N* = 49**	**.86*** *N* = 49**	**.70*** *N* = 49**
	Letter Knowledge	.49** *N* = 49	.43** *N* = 49	.49*** *N* = 49	.49*** *N* = 49
	Phonemic Awareness	.63** *N* = 49	.56** *N* = 49	.61*** *N* = 49	.55*** *N* = 49
	Alphabetic Decoding	.76*** *N* = 49	.61*** *N* = 49	.79*** *N* = 49	.65*** *N* = 49
	Spelling	.84*** *N* = 49	.67*** *N* = 49	.86*** *N* = 49	.72*** *N* = 49
	Reading Comprehension	.81*** *N* = 49	.62*** *N* = 49	.85*** *N* = 49	.59*** *N* = 49
2	**Overall Score**	**.82*** *N* = 49**	**.79*** *N* = 49**	**.86*** *N* =49**	**.83*** *N* = 49**
	Spelling	.82*** *N* = 49	.82*** *N* = 49	.87*** *N* = 49	.86*** *N* = 49
	Reading Comprehension	.73*** *N* = 49	.66*** *N* = 49	.77*** *N* = 49	.70*** *N* = 49
	Text Fluency	.77*** *N* = 49	.71*** *N* = 49	.75*** *N* = 49	.70*** *N* = 49
3	**Overall Score**	**.78*** *N* = 51**	**.60*** *N* = 51**	**.80*** *N* =51**	**.67*** *N* = 51**
	Spelling	.80*** *N* = 51	.58*** *N* = 51	.84*** *N* = 51	.75*** *N* = 51
	Reading Comprehension	.61*** *N* = 51	.45*** *N* = 51	.65*** *N* = 51	.50*** *N* = 51
	Text Fluency	.61*** *N* = 51	.46*** *N* = 51	.63*** *N* = 51	.54*** *N* = 51

*** $p < .001$, **$p < .01$, * $p < .05$

The ISIP Reading Overall score has a strong to very strong relationship with the WIAT-4 Dyslexia Index. The ISIP Phonemic Awareness subtest has a strong to very strong relationship with the WIAT-4 Phonemic Proficiency, and ISIP's Alphabetic Decoding has a strong relationship with WIAT-4 Phonemic Proficiency and Pseudoword Decoding. ISIP's Spelling and Reading Comprehension subtests also have a very strong relationship with the WIAT-4 Dyslexia Index and WIAT-4 Word Reading.

Table 5.10. *Correlations with the ISIP RAN and the KTEA RAN Subtests*

Grade	ISIP RAN	KTEA – LNF	KTEA – ONF	WIAT-4 Dyslexia Index
K	Letters	.22 $N = 44$	.55*** $N = 44$	.75*** $N = 44$
	Numbers	.25 $N = 47$	.60*** $N = 47$	.65*** $N = 47$
	Objects	.15 $N = 50$	.65*** $N = 50$	.48*** $N = 50$
	Composite	.21 $N = 44$	.70*** $N = 44$	.63*** $N = 44$
1	Letters	.71*** $N = 47$	.61*** $N = 47$	.70*** $N = 47$
	Numbers	.68*** $N = 49$	.57*** $N = 49$	.64*** $N = 49$
	Objects	.41** $N = 49$	.54*** $N = 49$	.29 $N = 49$
	Composite	.69*** $N = 47$	.66*** $N = 47$	.62*** $N = 47$
2	Letters	.75*** $N = 48$	.59*** $N = 49$	.58*** $N = 49$
	Numbers	.74*** $N = 47$	.61*** $N = 48$	.52*** $N = 49$
	Objects	.57*** $N = 47$	.61*** $N = 48$	.39** $N = 48$
	Composite	.77*** $N = 46$	.68*** $N = 47$	.55*** $N = 47$
3	Letters	.71*** $N = 51$	.46*** $N = 51$	.47*** $N = 51$
3	Numbers	.74*** $N = 50$	.52*** $N = 50$	.51** $N = 50$
3	Objects	.53*** $N = 51$	.61*** $N = 51$	.20 $N = 51$
3	Composite	.76*** $N = 50$	.59*** $N = 50$	.45** $N = 50$

*** $p < .001$, **$p < .01$, * $p < .05$

For ISIP RAN, the strongest correlations are with ISIP RAN letters and KTEA-3 LNF in grades 1-3 as well as with ISIP RAN numbers and the ISIP RAN composite score and the LNF in grades 1-3. ISIP RAN letters and numbers also correlate with the WIAT-4 Dyslexia Index, with stronger correlations in kindergarten and grade 1. ISIP RAN objects have the lowest correlations with the WIAT-4 Dyslexia Index.

We also conducted a study to determine if ISIP Reading and its subtests could be used to identify students at risk of dyslexia as early as kindergarten. The data for the study came from three school districts in two different states. Two of the school districts are classified as suburban and are located in a large metropolitan area. The other school district is classified as urban in a midsize city. We obtained information on the third-grade cohort of students in the 2018-2019 school year, including whether or not the students had been diagnosed with dyslexia by the end of third grade. We matched the students with their ISIP scores going back to kindergarten. The third-grade cohort consisted of 5,634 students at the middle of the year benchmark; 8.3% had been identified with dyslexia. The sample was approximately 56% Hispanic/Latino, 20% African American or Black, 17% White/non-Hispanic, and 7% Asian or other race/ethnicities. Sample sizes varied throughout the school years due to attrition; however, the demographic percentages were consistent. Mean differences were observed at each benchmark for all subtests, with the exception of Listening Comprehension in kindergarten, which is often a strength for students at risk of dyslexia. Results are available in Table 5.11. Full information about this study is available in a separate report from Istation.

Table 5.11. *ISIP Reading Means and Standard Deviations for Students Not at Risk and at Risk, by Overall and Subtest Scores*

Grade	Subtest	Benchmark	Students Not at Risk Mean and Standard Deviation	Students at Risk Mean and Standard Deviation	*F*	*p*
Kindergarten	Overall	Fall	259.73 (46.09)	243.04 (39.12)	24.30	< .001
	Listening Comprehension	Fall	249.69 (46.37)	245.41 (43.21)	1.48	.23
	Vocabulary	Fall	280.13 (56.46)	273.37 (54.88)	2.58	.11
	Phonemic Awareness	Fall	263.82 (50.60)	244.66 (39.94)	25.44	< .001
	Letter Knowledge	Fall	254.03 (60.93)	222.22 (56.19)	49.18	< .001
Kindergarten	Overall	Winter	314.64 (46.89)	290.34 (35.26)	82.43	< .001
	Listening Comprehension	Winter	301.51 (54.18)	299.47 (52.51)	.397	.53
	Vocabulary	Winter	318.31 (63.99)	301.25 (54.11)	21.37	< .001
	Phonemic Awareness	Winter	310.39 (49.26)	284.22 (42.55)	83.39	< .001
	Letter Knowledge	Winter	305.60 (53.53)	276.91 (40.53)	81.16	< .001

Kindergarten	Overall	Spring	344.04 (48.06)	320.37 (33.16)	76.88	< .001
	Listening Comprehension	Spring	325.50 (57.75)	327.60 (57.61)	.357	.551
	Vocabulary	Spring	353.45 (66.12)	334.01 (52.51)	26.84	< .001
	Phonemic Awareness	Spring	328.01 (48.68)	309.08 (40.34)	40.11	< .001
	Letter Knowledge	Spring	330.49 (55.82)	312.17 (44.96)	24.61	< .001
1	Overall	Fall	339.86 (46.19)	309.26 (28.65)	165.42	< .001
	Letter Knowledge	Fall	343.48 (56.50)	312.55 (46.32)	105.73	< .001
	Vocabulary	Fall	352.50 (56.70)	341.72 (51.08)	13.09	< .001
	Phonemic Awareness	Fall	342.51 (56.78)	314.32 (43.90)	86.89	< .001
	Spelling	Fall	339.42 (46.19)	309.26 (28.65)	218.09	< .001
	Reading Comprehension	Fall	331.55 (64.02)	281.91 (42.18)	220.43	< .001
	Alphabetic Decoding	Fall	338.87 (54.49)	300.65 (37.95)	177.59	< .001
1	Overall	Winter	376.46 (51.65)	333.35 (29.60)	275.39	< .001
	Vocabulary	Winter	386.75 (64.09)	364.50 (53.39)	46.12	< .001
	Spelling	Winter	374.12 (51.80)	328.61 (37.25)	300.00	< .001
	Reading Comprehension	Winter	371.18 (70.05)	301.97 (39.45)	386.42	< .001
	Alphabetic Decoding	Winter	377.26 (61.20)	326.84 (39.32)	263.68	< .001
1	Overall	Spring	405.52 (54.31)	355.15 (36.68)	342.72	< .001
	Vocabulary	Spring	411.86 (67.30)	385.03 (60.79)	61.29	< .001
	Spelling	Spring	403.24 (53.88)	356.05 (36.87)	305.32	< .001
	Reading Comprehension	Spring	409.08 (70.20)	334.73 (45.17)	448.81	< .001
	Alphabetic Decoding	Spring	407.02 (67.89)	346.19 (41.89)	320.55	< .001
2	Overall	Fall	408.91 (52.32)	359.79 (33.55)	347.09	< .001
	Vocabulary	Fall	410.51 (53.14)	388.03 (44.68)	68.52	< .001
	Spelling	Fall	400.99 (54.65)	347.69 (35.90)	129.49	< .001
	Reading Comprehension	Fall	418.12 (64.24)	349.00 (48.51)	436.32	< .001
	Text Fluency	Fall	27.42 (30.57)	2.99 (7.28)	219.74	< .001
2	Overall	Winter	434.86 (57.54)	375.74 (38.52)	464.05	< .001
	Vocabulary	Winter	443.47 (68.73)	410.92 (55.91)	96.75	< .001
	Spelling	Winter	425.41 (58.57)	360.95 (37.11)	218.40	< .001
	Reading Comprehension	Winter	447.39 (73.55)	365.74 (52.41)	538.80	< .001
	Text Fluency	Winter	44.69 (34.84)	8.04 (13.44)	428.91	< .001

2	Overall	Spring	453.55 (61.56)	393.94 (41.71)	421.17	< .001
	Vocabulary	Spring	463.79 (76.30)	426.82 (60.61)	103.92	< .001
	Spelling	Spring	443.52 (61.13)	375.17 (41.17)	223.86	< .001
	Reading Comprehension	Spring	468.17 (78.03)	391.80 (55.28)	428.69	< .001
	Text Fluency	Spring	57.19 (37.59)	19.59 (21.21)	381.95	< .001
3	Overall	Fall	452.80 (57.70)	401.73 (35.66)	318.61	< .001
	Vocabulary	Fall	453.94 (60.91)	427.50 (47.57)	74.93	< .001
	Spelling	Fall	443.28 (61.19)	385.45 (39.66)	360.88	< .001
	Reading Comprehension	Fall	466.25 (70.58)	404.45 (49.91)	308.23	< .001
	Text Fluency	Fall	56.92 (38.37)	22.41 (23.03)	309.34	< .001
3	Overall	Winter	468.98 (62.39)	410.52 (45.92)	389.00	< .001
	Vocabulary	Winter	485.04 (77.60)	445.34 (65.38)	114.38	< .001
	Spelling	Winter	457.37 (61.72)	392.42 (45.26)	490.02	< .001
	Reading Comprehension	Winter	484.88 (81.11)	415.09 (56.52)	329.78	< .001
	Text Fluency	Winter	59.38 (36.36)	22.61 (56.45)	369.57	< .001
3	Overall	Spring	481.97 (67.56)	424.13 (48.91)	288.31	< .001
	Vocabulary	Spring	504.98 (83.86)	462.79 (71.26)	98.05	< .001
	Spelling	Spring	469.02 (64.44)	406.37 (47.98)	371.01	< .001
	Reading Comprehension	Spring	498.32 (88.06)	428.41 (62.82)	248.25	< .001
	Text Fluency	Spring	65.85 (40.56)	32.42 (28.50)	260.06	< .001

We also established cut points to identify students at risk of dyslexia. Full information about how the classification accuracy was conducted and the sensitivity and specificity can be found in the special report for using ISIP as a dyslexia screener:

Istation (2022b). Istation's Indicators of Progress (ISIP) Reading and Rapid Auto Naming as a Dyslexia Screener. Dallas, TX: Istation.

Special Group Studies

Special group studies provide a different kind of validity information, primarily test criterion validity. We would expect students identified with different disabilities or disorders to have higher or lower ISIP scores, depending on their identification and grade level.

For students, a disability is defined as an emotional or intellectual condition requiring assistance to access the educational environment. A student with a disability has various learning challenges depending on the type of disability. Types of disabilities may include hearing loss, low vision or blindness, learning disabilities (such as dyslexia or dyscalculia), other health impairments such as mobility limitations or chronic health conditions (such as epilepsy, cancer, diabetes, migraine headaches, or multiple sclerosis), and psychological or psychiatric disabilities and neurodivergence (such as mood, anxiety, and depressive disorders; attention-deficit hyperactivity disorder; autism; and traumatic brain injury) (National Center for Education Statistics, 2022).

Methodology for Special Group Studies

Not all school districts provide information to Istation regarding whether a student has a disability. Because students with disabilities may need some assessment adaptations, accommodations, and modifications based on their needs, learning styles, and interests, we evaluated their learning progress and growth using 2018-2019 school year data, and we validated the results using the 2021-2022 school year data. The ultimate goal is to determine whether ISIP assessments are suitable for students with disabilities.

The data were pulled from the Istation database from three benchmarking assessment months — beginning of the year (BOY), middle of the year (MOY), and end of the year (EOY) — in the 2018-2019 and 2021-2022 school years. Millions of students enroll in the Istation reading program yearly, but not all schools provide demographic information. The sample consists of students who enrolled in the Istation reading program in the 2018-2019 and 2021-2022 school years who had disability information. There were over 10,000 students in each school year. We selected students with three complete data points of BOY, MOY, and EOY; the results are in Tables 5.12 and 5.13. This reduced the sample size to 6,039 students in 2018-2019 and 5,279 students in the 2021-2022 school year. We focused on nine disability categories: autism (AU), developmental disabilities (DD), emotional disturbances (ED), learning disabilities (LD), multiple disabilities (MD), intellectual disabilities (ID), other health impairments (OHI), speech impairment (SI), and specific learning disabilities (SLD). We computed the mean scores of Overall, Reading Comprehension (CMP), Vocabulary (VOC), and Spelling (SPL) for all three benchmarking assessment months by grade level for both school years. Disability types that had less than 30 students per grade are not reported. Students' growth from BOY to MOY and EOY are reported.

Results for Special Group Studies

The mean scores of Overall, Reading Comprehension (CMP), Vocabulary (VOC), and Spelling (SPL) by grade and by disability type at the BOY, MOY, and EOY of the 2018-2019 and 2021-2022 school years are in Tables 5.14 and 5.15. Students' growth from BOY to MOY and BOY to EOY are in Tables 5.16 and 5.17. Although some kindergarteners took the CMP and SPL subtests, these two subtests are designed for students in first through eighth grade, and results are not reported here. Overall, students with disabilities had lower scores at BOY, MOY, and EOY than students without disabilities. Also, students with each disability type performed differently from each other. In the 2018-2018 school year, students who were ID showed the lowest mean scores, whereas SI students showed the highest mean scores in lower grade levels. In the 2021-2022 school year, students with LD showed the lowest mean scores, whereas students with SI students showed the highest mean scores in lower grade levels.

Students showed slightly higher growth from BOY to MOY than MOY to EOY, and this is typical of students' growth in general. Some students did not show positive growth from BOY to MOY or BOY to EOY, especially in Spelling and Vocabulary subtests. These students may need specialized instruction.

These findings indicate that ISIP Reading is suitable for students with disabilities. Because students with disabilities have a different pattern of learning progression, the adaptive curriculum and assessments' accommodations and modifications may be considered in the future for students with disabilities.

Table 5.12. *Sample Size for each Type of Disability in the 2018-2019 School Year*

Grade	AU	DD	ED	LD	ID	MD	OHI	SI	SLD	Total
Kindergarten	74	182					38	385	63	742
1	87	154		105	31		95	596	74	1,142
2	90	135	30	222	33		119	480	100	1,209
3	77	49	36	287	39		121	313	78	1,000
4	82			300	39		144	167	40	772
5	38			263			85	86	47	519
6				83			35			118
7				101			36			137
8				36						36
Total	448	520	66	1397	142	0	673	2,027	402	6,039

Table 5.13. *Sample Size for Type of Disability in the 2021-2022 School Year*

Grade	AU	DD	ED	LD	ID	MD	OHI	SI	SLD	Total
Kindergarten	62	122					43	258	78	563
1	69	214		47			81	414	130	955
2	85	187		161			101	349	142	1,025
3	96	111		236		30	126	249	88	936
4	70			270		33	94	139	107	713
5	72			260		31	95	78	104	640
6				160			40	30		230
7				120			33			153
8				64						64
Total	454	634	0	1318	0	94	613	1517	649	5,279

Table 5.14. *Mean Scores of the 2018-2019 School Year*

Grade	Type	BOY Overall	MOY Overall	EOY Overall	BOY CMP	MOY CMP	EOY CMP	BOY VOC	MOY VOC	EOY VOC	BOY SPL	MOY SPL	EOY SPL
K	Mean	265	311	340				270	318	351			
	AU	231	269	301				230	252	291			
	DD	224	266	304				235	260	302			
	OHI	230	275	297				253	293	318			
	SI	236	282	310				251	288	319			
	SLD	247	290	323				269	301	335			
1	Mean	333	365	396	309	348	390	353	383	408	328	362	388
	AU	296	327	350	306	337	357	294	307	339	294	334	359
	DD	296	326	350	280	309	339	311	330	363	311	327	340
	LD	283	308	331	271	288	309	311	328	356	311	303	334
	ID	247	256	266	251	258	264	267	261	269	267	258	277
	OHI	289	317	332	279	307	323	309	333	355	309	319	337
	SI	309	339	365	293	325	358	329	352	376	329	340	364
	SLD	297	318	337	282	294	322	323	335	354	323	319	341
2	Mean	405	430	455	411	437	458	413	443	465	400	419	435
	AU	346	368	396	359	374	405	351	375	398	351	378	401
	DD	346	380	398	342	378	404	350	393	417	350	379	391
	ED	371	389	408	364	395	425	377	405	429	377	384	408
	LD	332	347	371	323	340	364	355	371	403	355	349	369
	ID	302	311	323	292	306	329	319	327	350	319	311	330
	OHI	341	358	369	336	352	367	361	387	397	361	356	369
	SI	376	401	421	381	410	432	386	418	442	386	392	413
	SLD	350	367	387	344	361	387	374	399	415	374	362	385
3	Mean	454	474	491	459	479	498	459	497	520	443	458	474
	AU	387	398	415	394	400	415	384	405	418	384	406	422
	DD	393	406	423	397	419	434	403	432	455	403	401	416
	ED	392	405	407	393	408	416	420	432	438	420	399	411
	LD	379	389	399	381	391	403	405	427	440	405	378	388
	ID	309	321	316	310	323	331	322	335	327	322	326	324
	OHI	381	390	404	383	394	410	402	419	436	402	389	398
	SI	424	444	458	439	460	475	430	461	481	430	429	443
	SLD	379	393	412	382	393	413	398	423	445	398	387	399

Table 5.14. *Mean Scores of the 2018-2019 School Year (continued)*

Grade	Type	BOY Overall	MOY Overall	EOY Overall	BOY CMP	MOY CMP	EOY CMP	BOY VOC	MOY VOC	EOY VOC	BOY SPL	MOY SPL	EOY SPL
4	Mean	500	516	533	527	551	560	456	478	792	514	535	549
	AU	427	443	449	470	495	501	411	418	424	415	425	448
	LD	413	433	442	448	477	479	405	416	424	402	423	438
	ID	366	372	357	428	452	451	371	366	337	330	336	338
	OHI	423	437	446	449	479	479	411	415	422	420	434	451
	SI	470	489	506	499	525	537	432	452	471	480	500	517
	SLD	396	413	432	442	468	474	401	398	416	369	391	420
5	Mean	535	547	559	563	581	589	496	511	523	549	562	572
	AU	483	502	510	505	522	529	459	471	483	495	523	527
	LD	445	458	462	468	486	487	442	444	449	441	452	461
	OHI	444	461	472	467	481	497	436	442	447	444	468	483
	SI	496	509	517	524	545	552	464	473	481	507	525	536
	SLD	455	475	482	480	504	505	443	451	463	457	481	483
6	Mean	553	568	583	580	594	608	523	535	548	567	579	590
	LD	467	475	480	485	493	486	466	454	466	471	480	489
	OHI	473	474	479	495	501	495	467	454	462	477	481	485
7	Mean	590	604	617	622	636	650	565	585	607	596	608	620
	LD	483	484	489	493	494	499	488	466	473	492	499	500
	OHI	475	484	469	483	479	475	478	468	445	486	500	494
8	Mean	624	636	649	648	666	683	603	636	659	626	638	650
	LD	481	492	509	489	508	527	487	478	490	495	491	515

Table 5.15. *Mean Scores of the 2021-2022 School Year*

Grade	Type	BOY Overall	MOY Overall	EOY Overall	BOY CMP	MOY CMP	EOY CMP	BOY VOC	MOY VOC	EOY VOC	BOY SPL	MOY SPL	EOY SPL
K	Mean	265	311	340				270	318	351			
	AU	235	272	294				229	275	295			
	DD	228	258	290				234	253	292			
	OHI	221	260	287				219	259	296			
	SI	238	274	306				252	284	318			
	SLD	237	281	298				255	300	315			
1	Mean	333	365	396	309	348	390	353	383	408	328	362	388
	AU	282	310	333	288	312	337	290	302	329	290	312	340
	DD	276	303	326	272	298	316	296	312	341	296	302	322
	LD	260	294	306	263	283	286	286	319	328	286	295	311
	OHI	287	314	340	284	310	341	305	326	356	305	313	339
	SI	296	327	353	281	305	337	329	355	385	329	319	345
	SLD	295	323	346	279	298	325	335	365	385	335	307	338
2	Mean	405	430	455	411	437	458	413	443	465	400	419	435
	AU	344	374	394	348	373	402	359	391	401	359	370	391
	DD	326	346	365	328	348	370	349	370	390	349	341	362
	LD	311	326	342	296	306	331	346	368	382	346	324	339
	OHI	335	353	370	327	355	369	363	391	410	363	348	368
	SI	355	379	403	352	376	410	384	415	439	384	369	387
	SLD	347	371	392	341	372	396	378	409	434	378	363	379
3	Mean	454	474	491	459	479	498	459	497	520	443	458	474
	AU	379	400	407	385	409	417	403	427	445	403	392	398
	DD	360	377	389	366	380	391	388	415	431	388	364	377
	LD	357	375	392	353	374	395	393	419	441	393	364	373
	MD	367	360	362	369	384	381	378	380	384	378	370	372
	OHI	376	398	418	377	398	421	410	445	465	410	382	395
	SI	415	435	453	425	451	468	439	472	496	439	413	430
	SLD	368	380	395	373	385	403	399	416	441	399	369	382

Table 5.15. *Mean Scores of the 2021-2022 School Year (continued)*

Grade	Type	BOY Overall	MOY Overall	EOY Overall	BOY CMP	MOY CMP	EOY CMP	BOY VOC	MOY VOC	EOY VOC	BOY SPL	MOY SPL	EOY SPL
4	Mean	500	516	533	527	551	560	456	478	792	514	535	549
	AU	396	416	434	444	476	477	382	389	403	381	411	449
	LD	392	412	428	440	476	479	390	400	418	367	386	406
	MD	389	398	407	440	450	455	387	386	377	369	386	413
	OHI	420	449	460	458	492	497	407	430	448	412	435	448
	SI	453	476	496	490	519	536	427	447	464	454	481	502
	SLD	400	414	412	450	471	466	393	398	393	365	390	411
5	Mean	535	547	559	563	581	589	496	511	523	549	562	572
	AU	444	457	473	465	492	501	440	437	454	444	456	472
	LD	438	454	466	470	493	498	429	439	452	434	444	459
	MD	448	440	469	479	503	513	444	413	439	446	451	475
	OHI	457	477	496	472	503	522	446	461	482	462	478	496
	SI	487	497	513	514	527	555	460	467	482	493	509	515
	SLD	423	441	446	466	484	487	407	421	421	417	435	449
6	Mean	553	568	583	580	594	608	523	535	548	567	579	590
	LD	445	449	448	467	483	482	442	426	416	449	454	465
	OH	467	486	488	487	514	510	454	453	450	484	502	512
	SI	534	550	556	559	586	595	511	523	523	540	547	569
7	Mean	590	604	617	622	636	650	565	585	607	596	608	620
	LD	475	481	492	489	496	497	469	453	460	488	497	511
	OH	475	459	480	484	470	472	474	438	454	485	479	500
8	Mean	624	636	649	648	666	683	603	636	659	626	638	650
	LD	482	483	482	477	492	485	479	456	457	506	508	509

Table 5.16. *Students' Growth in the 2018-2019 School Year*

Grade	Type	BOY-MOY Overall	BOY-EOY Overall	BOY-MOY CMP	BOY-EOY CMP	BOY-MOY VOC	BOY-EOY VOC	BOY-MOY SPL	BOY-EOY SPL
K	AU	38	70			22	61		
	DD	42	80			25	67		
	OHI	45	67			40	65		
	SI	46	74			37	68		
	SLD	43	76			32	66		
1	AU	31	54	31	51	13	45	40	65
	DD	30	54	29	59	19	52	16	29
	LD	25	48	17	38	17	45	−8	23
	ID	9	19	7	13	−6	2	−9	10
	OHI	28	43	28	44	24	46	10	28
	SI	30	56	32	65	23	47	11	35
	SLD	21	40	12	40	12	31	−4	18
2	AU	22	50	15	46	24	47	27	50
	DD	34	52	36	62	43	67	29	41
	ED	18	37	31	61	28	52	7	31
	LD	15	39	17	41	16	48	−6	14
	ID	9	21	14	37	8	31	−8	11
	OHI	17	28	16	31	26	36	−5	8
	SI	25	45	29	51	32	56	6	27
	SLD	17	37	17	43	25	41	−12	11
3	AU	11	28	6	21	21	34	22	38
	DD	13	30	22	37	29	52	−2	13
	ED	13	15	15	23	12	18	−21	−9
	LD	10	20	10	22	22	35	−27	−17
	ID	12	7	13	21	13	5	4	2
	OHI	9	23	11	27	17	34	−13	−4
	SI	20	34	21	36	31	51	−1	13
	SLD	14	33	11	31	25	47	−11	1
4	AU	16	22	25	31	7	13	10	33
	LD	20	29	29	31	11	19	21	36
	ID	6	−9	24	23	−5	−34	6	8
	OHI	14	23	30	30	4	11	14	31
	SI	19	36	26	38	20	39	20	37
	SLD	17	36	26	32	−3	15	22	51
5	AU	19	27	17	24	12	24	28	32
	LD	13	17	18	19	2	7	11	20
	OHI	17	28	14	30	6	11	24	39
	SI	13	21	21	28	9	17	18	29
	SLD	20	27	24	25	8	20	24	26
6	LD	8	13	8	1	−12	0	9	18
6	OHI	1	6	6	0	−13	−5	4	8
7	LD	1	6	1	6	−22	−15	7	8
	OH	9	−6	−4	−8	−10	−33	14	8
8	LD	11	28	19	38	−9	3	−4	20

Table 5.17. *Students' Growth in the 2021-2022 School Year*

Grade	Type	BOY-MOY Overall	BOY-EOY Overall	BOY-MOY CMP	BOY-EOY CMP	BOY-MOY VOC	BOY-EOY VOC	BOY-MOY SPL	BOY-EOY SPL
K	AU	37	59			46	66		
	DD	30	62			19	58		
	OHI	39	66			40	77		
	SI	36	68			32	66		
	SLD	44	61			45	60		
1	AU	28	51	24	49	12	39	22	50
	DD	27	50	26	44	16	45	6	26
	LD	34	46	20	23	33	42	9	25
	OHI	27	53	26	57	21	51	8	34
	SI	31	57	24	56	26	56	−10	16
	SLD	28	51	19	46	30	50	−28	3
2	AU	30	50	25	54	32	42	11	32
	DD	20	39	20	42	21	41	−8	13
	LD	15	31	10	35	22	36	−22	−7
	OHI	18	35	28	42	28	47	−15	5
	SI	24	48	24	58	31	55	−15	3
	SLD	24	45	31	55	31	56	−15	1
3	AU	21	28	24	32	24	42	−11	−5
	DD	17	29	14	25	27	43	−24	−11
	LD	18	35	21	42	26	48	−29	−20
	MD	−7	−5	15	12	2	6	−8	−6
	OHI	22	42	21	44	35	55	−28	−15
	SI	20	38	26	43	33	57	−26	−9
	SLD	12	27	12	30	17	42	−30	−17
4	AU	20	38	32	33	7	21	30	68
	LD	20	36	36	39	10	28	19	39
	MD	9	18	10	15	−1	−10	17	44
	OHI	29	40	34	39	23	41	23	36
	SI	23	43	29	46	20	37	27	48
	SLD	14	12	21	16	5	0	25	46
5	AU	13	29	27	36	−3	14	12	28
	LD	16	28	23	28	10	23	10	25
	MD	−8	21	24	34	−31	−5	5	29
	OHI	20	39	31	50	15	36	16	34
	SI	10	26	13	41	7	22	16	22
	SLD	18	23	18	21	14	14	18	32
6	LD	4	3	16	15	−16	−26	5	16
	OH	19	21	27	23	−1	−4	18	28
	SI	16	22	27	36	12	12	7	29
7	LD	6	17	7	8	−16	−9	9	23
	OH	−16	5	−14	−12	−36	−20	−6	15
8	LD	1	0	15	8	−23	−22	2	3

Item Reliability and Bias Analysis

Assessments must meet specific important psychometric properties, including test fairness, or item bias. Item bias is an essential issue in educational testing because different subgroups of examinees should have an equal probability of answering an item correctly, given the same ability level. Differential item functioning (DIF) analyses were developed to identify bias in assessments. After adjusting for a group's overall test ability, DIF analyses were first used in educational testing settings to investigate whether particular items in a test were fair to some subgroups, such as female students or a particular ethnic group. DIF analyses can be carried out using a wide range of statistical methods to explore the relationship between these three variables: group membership (e.g., male vs. female) associated with differential responses (correct vs. incorrect) to an item (x) for respondents at the same level of a matching criterion or matching variable (e.g., a latent variable such as an ability (θ) score or observed scale score). The matching criterion or matching variable is used to account for different levels of functioning or ability in each subgroup. DIF can also be used to assess item drift, or changes in difficulty over different periods of time. We investigated both item DIF and item drift during our renorming process.

Item DIF

During the development of ISIP Reading, experts analyzed items for bias, and we conducted DIF analysis. For this update, we conducted an updated DIF study to determine whether the items were performing consistently.

There are many different methodologies to detect DIF in a test. One widely used approach for detecting DIF is a logistic regression because it is simple, robust, and easy to implement. 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. Logistic regression is as robust as the Mantel-Haenszel procedure at detecting uniform DIF. We applied the logistic regression DIF detection method to explore potential DIF items in all ISIP Reading subtests with two different matching criteria (overall reading ability score and subscale reading score), two DIF factors of gender (male vs. female) and race/ethnicity (non-Hispanic

White vs. all other combined), and two DIF detection criteria: Zumbo & Thomas (ZT) and Jodoign & Gierl (JG) (Magis, et al., 2010).

Data and Methods for DIF Analysis

The sample consisted of 65,000 students from kindergarten to eighth grade from three large districts in three states. We limited the sample to three districts to keep the file size manageable, and the districts gave sufficient racial/ethnic diversity for Hispanic/Latino, African American/Black, and White students. The data were pulled by subtest from the Istation database from May of the 2018-2019 school year. The original data had over 3,000 items across all subtests. Items that had less than 100 responses were discarded. Thus, 2,598 items were included in this study. The analyses were completed using the difR package in R software. The DIF effect size under ZT is as follows. If the DIF magnitude (Nagelkerke's R^2) is less than or equal to 0.13, an item displays A DIF item (negligible or non-significant DIF effect). If the DIF magnitude is greater than 0.13 but less than or equal to 0.26, an item displays B DIF item (slightly to moderate DIF effect), and an item displays C DIF item (moderate to large DIF effect) if the magnitude is greater than 0.26. On the other hand, the JG criterion has a different cut point for identifying DIF effect size. If the DIF magnitude is less than or equal to 0.035, an item displays A DIF item. If the DIF magnitude is greater than 0.035 but less than or equal to 0.070, an item displays B DIF item, and an item displays C DIF item if the magnitude is greater than 0.070.

Results of DIF Analysis

Pearson product-moment correlation coefficients between overall scores and subscale scores were very high, ranging from 0.72 to 0.90 (R^2 ranges from 0.52 to 0.81), indicating that the overall reading ability scores and subscale reading scores are highly associated with each other. The DIF results are summarized in Tables 5.18 through 5.25. Tables 5.18 to 5.21 show the potential DIF items when the overall score and subscale score are the matching criteria, and the ZT is the DIF Detection criteria. Overall, only 1% of items display B and C DIF items combined. Gender DIF factor detects four B DIF items and one C DIF item, and the race/ethnicity DIF factor detects only three B DIF items across all subtests. Tables 5.22 to 5.25 show the potential DIF items when the overall score and subscale score are the matching criteria, and the JG is the DIF Detection criteria. Because the JG criterion is more sensitive than ZT, approximately 2% of items displayed B and C DIF items combined. Gender DIF factor detects 25 B DIF items and 13 C DIF items, and race/ethnicity DIF factor detects 23 B

DIF items and 13 C DIF items across all subtests. Curriculum experts reviewed all B and C DIF items and recommended whether to keep or remove these items from the item pool.

Table 5.18. *Potential Gender DIF Items with ZT DIF Detection Criteria and Overall Score as Matching Criteria*

Grades	Subtest	A item	B item	C item
4 to 8	Reading Comprehension	724	0	0
4 to 8	Spelling	203	1	0
4 to 8	Vocabulary	230	1	0
K to 1	Alphabetic Decoding	172	0	0
1 to 3	Reading Comprehension	143	0	0
Pre-K to K	Listening Comprehension	170	0	0
Pre-K to 1	Letter Knowledge	244	0	1
K to 1	Phonemic Awareness	409	0	0
1 to 3	Spelling	149	2	0
Pre-K to 3	Vocabulary	179	0	0

Table 5.19. *Potential Race/Ethnicity DIF Items with ZT DIF Detection Criteria and Overall Score as Matching Criteria*

Grades	Subtest	A item	B item	C item
4 to 8	Reading Comprehension	724	0	0
4 to 8	Spelling	203	1	0
4 to 8	Vocabulary	230	1	0
K to 1	Alphabetic Decoding	172	0	0
1 to 3	Reading Comprehension	143	0	0
Pre-K to K	Listening Comprehension	170	0	0
Pre-K to 1	Letter Knowledge	245	0	0
K to 1	Phonemic Awareness	409	0	0
1 to 3	Spelling	150	1	0
Pre-K to 3	Vocabulary	179	0	0

Table 5.20. *Potential Gender DIF Items with ZT DIF Detection Criteria and Subscale Score as Matching Criteria*

Grades	Subtest	A item	B item	C item
4 to 8	Reading Comprehension	724	0	0
4 to 8	Spelling	203	1	0
4 to 8	Vocabulary	230	1	0
K to 1	Alphabetic Decoding	172	0	0
1 to 3	Reading Comprehension	143	0	0
Pre-K to K	Listening Comprehension	170	0	0
Pre-K to 1	Letter Knowledge	244	0	1
K to 1	Phonemic Awareness	409	0	0
1 to 3	Spelling	150	1	0
Pre-K to 3	Vocabulary	179	0	0

Table 5.21. *Potential Race/Ethnicity DIF Items with ZT DIF Detection Criteria and Subscale Score as Matching Criteria*

Grades	Subtest	A item	B item	C item
4 to 8	Reading Comprehension	724	0	0
4 to 8	Spelling	203	1	0
4 to 8	Vocabulary	230	1	0
K to 1	Alphabetic Decoding	172	0	0
1 to 3	Reading Comprehension	143	0	0
Pre-K to K	Listening Comprehension	170	0	0
Pre-K to 1	Letter Knowledge	245	0	0
K to 1	Phonemic Awareness	409	0	0
1 to 3	Spelling	150	1	0
Pre-K to 3	Vocabulary	179	0	0

Table 5.22. *Potential Gender DIF Items with JG DIF Detection Criteria and Overall Score as Matching Criteria*

Grades	Subtest	A item	B item	C item
4 to 8	Reading Comprehension	724	0	0
4 to 8	Spelling	198	3	3
4 to 8	Vocabulary	227	2	2
K to 1	Alphabetic Decoding	170	1	1
1 to 3	Reading Comprehension	140	3	0
Pre-K to K	Listening Comprehension	139	1	0
Pre-K to 1	Letter Knowledge	238	6	1
K to 1	Phonemic Awareness	403	5	1
1 to 3	Spelling	146	3	2
Pre-K to 3	Vocabulary	176	1	2

Table 5.23. *Potential Race/Ethnicity DIF Items with JG DIF Detection Criteria and Overall Score as Matching Criteria*

Grades	Subtest	A item	B item	C item
4 to 8	Reading Comprehension	723	1	0
4 to 8	Spelling	203	0	1
4 to 8	Vocabulary	227	2	2
K to 1	Alphabetic Decoding	170	2	0
1 to 3	Reading Comprehension	143	0	0
Pre-K to K	Listening Comprehension	138	2	0
Pre-K to 1	Letter Knowledge	243	0	2
K to 1	Phonemic Awareness	405	3	1
1 to 3	Spelling	150	0	1
Pre-K to 3	Vocabulary	176	3	0

Table 5.24. *Potential Gender DIF Items with JG DIF Detection Criteria and Subscale Score as Matching Criteria*

Grades	Subtest	A item	B item	C item
4 to 8	Reading Comprehension	724	0	0
4 to 8	Spelling	198	3	3
4 to 8	Vocabulary	227	2	2
K to 1	Alphabetic Decoding	170	0	2
1 to 3	Reading Comprehension	140	3	0
Pre-K to K	Listening Comprehension	139	0	1
Pre-K to 1	Letter Knowledge	240	4	1
K to 1	Phonemic Awareness	402	5	2
1 to 3	Spelling	147	3	1
Pre-K to 3	Vocabulary	177	1	1

Table 5.25. *Potential Race/Ethnicity DIF Items with JG DIF Detection Criteria and Subscale Score as Matching Criteria*

Grades	Subtest	A item	B item	C item
4 to 8	Reading Comprehension	723	1	0
4 to 8	Spelling	202	1	1
4 to 8	Vocabulary	226	2	3
K to 1	Alphabetic Decoding	170	2	0
1 to 3	Reading Comprehension	143	0	0
Pre-K to K	Listening Comprehension	139	0	1
Pre-K to 1	Letter Knowledge	243	1	1
K to 1	Phonemic Awareness	404	3	2
1 to 3	Spelling	150	0	1
Pre-K to 3	Vocabulary	176	3	0

Item Parameter Drift

Item parameter drift (IPD) is a special case of differential item functioning (DIF). In both instances, an item does not perform the same across sub-groups of examinees. DIF examines differences in item performance in different sub-groups such as gender, described above. IPD refers to changes in item performance at different points in time. The drift will affect item difficulties much more strongly than item discriminations. Easier items tended to drift in a positive direction, becoming more difficult, and difficult items tended to drift in a negative direction, becoming easier. This relationship suggested that the significantly drifting items tended to drift toward more moderate difficulty levels rather than drifting to the extremes (Gaertner & Briggs, 2009; Jones & Smith, 2006; Linacre, 2013; Risk, 2016). This study investigated IPD in ISIP Reading using three different years of data (May assessment months of the 2014-2015, 2018-2019, and 2020-2021 school years).

Data and Methods for the Item Drift Study

The original data files had approximately 3,000 items in each school year. Items with less than 100 responses and items with less than two categories were removed. The analyses were completed by subtest and by school year. All analyses are constrained to be on the same scale and fitted by the 1PL IRT model using Mplus software. Several studies show that Rasch Model or 1PL IRT is suitable for IPD analyses because the drift will affect item difficulties much more strongly than item discriminations (Gaertner & Briggs, 2009; Jones & Smith, 2006; Linacre, 2013; Risk, 2016; Wright & Douglas, 1976). Samples consisted of kindergarten to eighth-grade students from two large districts in two different states. These students enrolled in the Istation reading program in the 2014-2015, 2018-2019, and 2020-2021 school years, and there were approximately 70,000 students each school year. In the 2014-2015 school year, 32% were female students, 35% were male, and 33% were unknown. Approximately 40% in this school year were non-Hispanic White. In 2018-2019, 48% were female students, and 52% were male. Approximately 37% of the 2018-2019 students were non-Hispanic White. In the 2020-2021 school year, 49% were female students, 51% were male, and 38% of the students in this year were non-Hispanic White.

The 0.6 logits approach is a simple method for identifying IPD using IRT-based parameter estimates and is recommended by Wright and Douglas (1975). If an item drifts more than 0.6 logits from one test occasion to another test occasion, it is considered a potential drift item. The 0.6 logits approach was modified to 0.7 logits if the item difficulty parameters in the pool range from −3.0 to 3.0 by Wright and Douglas (1976). The item difficulty parameters of the same item from two test events are estimated separately and then directly compared. Items for which the difficulty parameters have shifted greater than 0.7 logits or more from one testing occasion to the next are considered candidates for item drift. We implemented the 0.7 logits criteria in this study. The calibrated item difficulty parameters of the same items from 2015 are directly compared to those from 2019 and 2021. The calibrated item difficulty parameters from 2019 are also compared to the calibrated item difficulty parameters from 2021.

Item Drift Directions

Results show that some items get harder from year to year, others get easier over time, and some remain the same over the years. This is typical behavior for item drift

and has been found in many previous studies (Gaertner & Briggs, 2009; Jones & Smith, 2006; Linacre, 2013; Risk, 2016; Wright & Douglas, 1976).

Results are available in Table 5.26. We found 91 items with potential drift from 2015 to 2019. For grades 4 to 8, Reading Comprehension had 28 items, Spelling had 7 items, and Vocabulary had 11 items that met the criteria for drift. For prekindergarten through third grade, Alphabetic Decoding had 3 items identified with potential drift, Reading Comprehension had 1, Letter Knowledge had 21, Phonemic Awareness had 6, Spelling had 8, and Vocabulary had 6. From 2019 to 2021 there were 37 items with potential drift, and from 2015 to 2021 there were 110 items with potential drift. Overall, this study found approximately 3% of the pool had potential item drifts. Curriculum experts reviewed the items with potential drift and made recommendations for removal or retention in the pool.

Table 5.26. *Potential Item Drift by Year*

Grades	Subtest	2015-2019	2019-2021	2015-2021
4 to 8	Reading Comprehension	28	0	37
4 to 8	Spelling	7	7	8
4 to 8	Vocabulary	11	7	13
K to 1	Alphabetic Decoding	3	1	1
1 to 3	Reading Comprehension	1	1	1
Pre-K to 1	Letter Knowledge	21	8	13
K to 1	Phonemic Awareness	6	3	8
1 to 3	Spelling	8	1	7
1 to 3	Vocabulary	6	9	22
	Total	91	37	110

Scale Item Parameter Drift

We evaluated scale item parameter drifts using Kingsbury and Wise's (2011) method by computing the correlations between (a) the calibrated item difficulty parameters in 2015 and 2019, (b) the calibrated item difficulty parameters in 2019 and 2021, and (c) the calibrated item difficulty parameters in 2015 and 2021 by subtest, and results are in Table 2. We also evaluated an average of item parameter drift from (a)

2015 to 2019, (b) 2019 to 2021, and (c) 2015 to 2021, and the results are in Table 5.27. There are high correlations between the calibrated item difficulty parameters from year to year, ranging from 0.82 to 0.97, as displayed in Table 5.27. The average correlations across all subtests for 2015-2019, 2019-2021, and 2015-2021 are 0.92, showing the scale stability across the years. Table 5.28 shows that the average item parameter drifts from year to year. Results show that some subtests get more difficult, whereas some subtests get easier over time. The average item drifts from 2015-2019 were 0.15, meaning that the items were slightly easier in 2019. Similar to the average item drift from 2015-2021, those items were slightly easier in 2021. The average item drift from 2019-2021, on the other hand, was −0.07, meaning that the items were slightly harder in 2021. This result is not surprising, given that 2021 was during the COVID-19 pandemic. May 2019 data were pre-pandemic data, whereas May 2021 data were pandemic data. Because of students' learning lag during the COVID years, items turned out to be slightly harder in 2021. However, these item parameter drifts are minimal and will not affect any parameter estimations of students' abilities.

Table 5.27. *Scale Item Parameter Drifts*

Grades	Subtest	2015-2019	2019-2021	2015-2021
4 to 8	Reading Comprehension	0.91	0.96	0.90
4 to 8	Spelling	0.96	0.97	0.95
4 to 8	Vocabulary	0.96	0.93	0.97
K to 1	Alphabetic Decoding	0.92	0.94	0.95
1 to 3	Reading Comprehension	0.96	0.92	0.95
Pre-K to 1	Letter Knowledge	0.90	0.91	0.92
K to 1	Phonemic Awareness	0.83	0.82	0.86
1 to 3	Spelling	0.89	0.89	0.85
1 to 3	Vocabulary	0.97	0.96	0.96
	Average	0.92	0.92	0.92

Table 5.28. *Average Item Parameter Drift*

Grades	Subtest	2015-2019	2019-2021	2015-2021
4 to 8	Reading Comprehension	0.29	0.02	0.31
4 to 8	Spelling	0.10	−0.05	0.00
4 to 8	Vocabulary	0.26	−0.17	0.11
K to 1	Alphabetic Decoding	0.08	−0.13	−0.03
1 to 3	Reading Comprehension	0.12	−0.06	0.07
Pre-K to 1	Letter Knowledge	0.25	−0.09	0.17
K to 1	Phonemic Awareness	−0.26	−0.08	−0.33
1 to 3	Spelling	0.20	−0.12	0.07
1 to 3	Vocabulary	0.34	0.04	0.41
Average	Average	0.15	−0.07	0.09

REFERENCES

Anastasi, A., & Urbina, S. (1997). *Psychological Testing (7th ed.). Prentice Hall/Pearson Education.*

Andersson, B., & Xin, T. (2018). Large Sample Confidence Intervals for Item Response Theory Reliability Coefficients. *Educational and Psychological Measurement*, *78*(1), 32–45. https://doi.org/10.1177/0013164417713570

Beck, I., & McKeown, M. (2002). *Bringing Words to Life: Robust Vocabulary Instruction*. New York: Guilford.

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

Bourassa, D.C., & Treiman, R. (2001). Spelling development and disabilities: The importance of linguistic factors. *Language, Speech, and Hearing Services in Schools, 32,* 172-181.

Bravo, M.A., & Cervett, G .N. (2008). *Teaching vocabulary through text and experience in content areas*. In A.E. Farstrup, & S. Samuels (Eds.), *What research has to say about reading instruction*. Newark, DE: International Reading Association.

Cain, K. (2006). Individual differences in children's memory and reading comprehension: An investigation of semantic and inhibitory deficits. *Memory, 14*(5), 553-569.

Cain, K., & Oakhill, J. V. (2007). Children's comprehension problems in oral and written language: A cognitive perspective. New York: The Guilford Press.

Campbell, L. O., Sutter, C. C, & Lambie, G. W. (2019). Predictability of Istation's Indicators of Progress Scores on Students' Virginia Standard of Learning Scores: Grades 3 through 8. Orlando, FL: University of Florida Morgridge International Reading Center. Available at www.istation.com/studies

Campbell, L. O., Sutter, C. C., Lambie, G. W., & Tinstman Jones, J. (2019). Measuring the predictability of Istation Indicators of Progress Early Reading (ISIP-ER) scores on Renaissance STAR Reading® scores. University of Central Florida. www.ucf.edu/mirc

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

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.

Chan, L. K. (1991). Promoting strategy generalization through self-instructional training in students with reading disabilities. Journal of Learning Disabilities, 24(7), 427-433.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge.

Cook, M. A. & Ross, S. M. (2020a). SC-Ready Predictability Study. Baltimore, MD: Johns Hopkins University Center for Research and Reform in Education. Available at www.istation.com/studies

Cook, M. A. & Ross, S. M. (2020c). NWEA MAP Predictability Study. Baltimore, MD: Johns Hopkins University Center for Research and Reform in Education. Available at www.istation.com/studies

Cook, M. A. & Ross, S. M. (2020b). PARCC Predictability Study – 3rd grade. Baltimore, MD: Johns Hopkins University Center for Research and Reform in Education. Available at www.istation.com/studies

Crocker, Linda, Algina, James. (2008). *Introduction to Classical & Modern Test Theory*. United States: Cengage Learning.

CTB/McGraw-Hill. (1996). *CAT/5 technical report*. CTB/McGraw-Hill.

Cutting, L.E., & Scarborough, H.S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading*, 10(3), 277-299.

Dadey, N., Lyons, S., & DePascale, C. (2018). The comparability of scores from different digital devices: A literature review and synthesis with recommendations for practice. Applied Measurement in Education, 31, 1, 30-50. Retrieved from https://doi.org/10.1080/08957347.2017.1391262

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin and Review, 3,* 422-433.

Deane P., Sheehan, K.M., Sabatini J., Futagi Y., & Kostin, I. (2006). Differences in Text Structure and Its Implications for Assessment of Struggling Readers, *Scientific Studies of Reading*, 10(3), 257-275, Retrieved from https://doi.org/10.1207/s1532799xssr1003_4

Deshler, D. D., Schumaker, J. B., Lenz, B. K., Bulgren, J. A., Hock, M. F., Knight, J., & Ehren, B. J. (2001). Ensuring content-area learning by secondary students with learning disabilities. *Learning Disabilities Research & Practice*, *16*(2), 96-108.

Ehri, L. C., & Wilce, L. S. (1987). Does learning to spell help beginners learn to read words?. *Reading research quarterly*, 47-65.

Ehri, L. C. (2000). Learning to read and learning to spell: Two sides of a coin. *Topics in language disorders*.

Englert, C. S., & Mariage, T. V. (1991). Making students partners in the comprehension process: Organizing the reading "POSSE". *Learning Disability Quarterly*, *14*(2), 123-138.

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.

Fletcher, J. (2006). Measuring reading comprehension. *Scientific Studies of Reading, 10*(3), 323-330.

Fletcher, J. M., Lyon, G. R., Fuchs, L. S., & Barnes, M. A. (2007). Learning disabilities. New York: The Guilford Press.

Francis, D.J., Snow, C.E., August, D., Carlson, C.D., Miller, J., & Iglesias, A. (2006). Measures of Reading Comprehension: A Latent Variable Analysis of the Diagnostic Assessment of Reading Comprehension, Scientific Studies of Reading, 10(3), 301-322. Retrieved from https://doi.org/10.1207/s1532799xssr1003_6

Fuchs, D., & Fuchs, L. S. (1990). Making educational research more important. *Exceptional Children*, *57*(2) , 102 -108.

Fuchs, D., Fuchs, L. S., Thompson, A., Al Otaiba, S., Yen, L., Yang, N. J., et al. (2001). Is reading important in reading-readiness programs? A randomized field trial with teachers as program implementers. *Journal of Educational Psychology, 93*, 251–267.

Gaertner, M. N. & Briggs, D. C. (2009). Detecting and Addressing Item Parameter Drift in IRT Test Equating Contexts.

Gersten, R., Fuchs, L. S., Williams, J. P., & Baker, S. (2001). Teaching reading comprehension strategies to students with learning disabilities: A review of research. *Review of educational research*, *71*(2), 279-320.

Good, R. H., Simmons, D., & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading*, 5, 257-288.

Graham, J. R. (2000). How big are the tax benefits of debt?. *The journal of finance*, *55*(5), 1901-1941.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*(4), 347–360. https://doi.org/10.1111/j.1745-3984.1984.tb01039.x

Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and multivariate applications*. Routledge.

Grossmann, M., Reckhow, S., Strunk, K. O., & Turner, M. (2021). All states close but red districts reopen: The politics of in-person schooling during the COVID-19 pandemic. *Educational Researcher, 50*(9), 637-648. https://doi.org/10.3102/0013189X211048840

Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement, 10*(3), 159–170. https://doi.org/10.1111/j.1745-3984.1973.tb00793.x

Hemphill, L., & Tivnan, T. (2008). The importance of early vocabulary for literacy achievement in high-poverty schools. *Journal of Education for Students Placed at Risk*, *13*(4), 426-451.

Hasbrouck, J., & Tindal, G. (2017). An update to compiled ORF norms. *Behavioral Research and Teaching*.

International Dyslexia Association. (2019). *Dyslexia Basics*. Retrieved from https://dyslexiaida.org/dyslexia-basics/

Irwin, D. E. (1991). Information integration across saccadic eye movements. *Cognitive psychology*, *23*(3), 420-456.

Istation (2020). Istation's Indicators of Progress (ISIP) Oral Reading Fluency Technical Report. Dallas, TX: Istation.

Istation (2022a). Istation's Indicators of Progress (ISIP) Rapid Auto Naming (ISIP RAN Technical Report. Dallas, TX: Istation.

Istation (2022b). Istation's Indicators of Progress (ISIP) Reading and Rapid Auto Naming as a Dyslexia Screener. Dallas, TX: Istation. www.istation.com/studies.

Jagers, P. (1986). Post-stratification against bias in sampling. *International Statistical Review / Revue Internationale de Statistique, 54*(2), 159-167. https://doi.org/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.

Jitendra, A. K., Kay Hoppes, M., & Xin, Y. P. (2000). Enhancing main idea comprehension for students with learning problems: The role of a summarization strategy and self-monitoring instruction. The Journal of Special Education, 34(3), 127-139.

Jones, P. & Smith, R. (2006). Item Parameter Drift in Certification Exams and Its Impact on Pass-Fail Decision Making, Paper presented NCME, San Francisco.

Kaufman A.S., & Kaufman N.L. (2014). Technical and Interpretive *Kaufman Test of Educational Achievement–Third Edition (KTEA-3)*. Bloomington, MN: NCS Pearson

Kelly, D. P. (2021). Pandemic pedagogy: K-12 technology and engineering education under COVID-19. *The Journal of Technology Studies, 47*(1), 2-11. Retrieved from https://www.jstor.org/stable/48657931

Kilpatrick, D. A. (2015). *Essentials of assessing, preventing, and overcoming reading difficulties*. John Wiley & Sons.

Kingsbury, G. G., & Wise, S. L. (2011). Creating a K-12 Adaptive Test: Examining the Stability of Item Parameter Estimates and Measurement Scales. *Journal of Applied Testing*

Kirby, J. R., Parrila, R. K., & Pfeiffer, S. L. (2003). Naming Speed and Phonological Awareness as Predictors of Reading Development. *Journal of Educational Psychology, 95*(3), 453-464. https://doi.org/10.1037/0022-0663.95.3.453

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.

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.

Kuhfeld, M., Soland, J., & Lewis, K. (2022). Test score patterns across three COVID-19-impacted school years. *(EdWorkingPaper: 22-521)*. Retrieved from https://edworkingpapers.com/sites/default/files/ai22-521.pdf

LaFond, Lee James. (2014). Decision consistency and accuracy indices for the bifactor and testlet response theory models. PhD (Doctor of Philosophy) thesis, University of Iowa, 2014. https://doi.org/10.17077/etd.ytivc04x

Lenhard, A., Lenhard, W., & Gary, S. (2018). *Continuous norming (cNORM)* (3.01). The Comprehensive R Network. https://CRAN.R-project.org/package=cNORM

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

Lenz, B. K., & Hughes, C. A. (1990). A word identification strategy for adolescents with learning disabilities. *Journal of Learning Disabilities*, *23*(3), 149-158.

Linacre, J. M. (2013). Winsteps® Rasch measurement computer program User's Guide. Beaverton, Oregon: Winsteps.com.

Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. Behavior Research Methods 2010, 42 (3), 847-862. https://doi.org/10.3758/BRM.42.3.847

Malone, L. D., & Mastropieri, M. A. (1992). Reading comprehension instruction: Summarization and selfmonitoring training for students with learning disabilities. Exceptional Children, 58, 270-279.

Mastropieri, M. A., Scruggs, T. E., & Graetz, J. E. (2003). Reading comprehension instruction for secondary students: Challenges for struggling students and teachers. *Learning disability quarterly*, *26*(2), 103-116.

Mathes, P., Torgesen, J. & Herron, J. (2016). Istation's Indicators of Progress (ISIP) Early Reading Technical Report: Computer Adaptive Testing System for continuous Progress Monitoring of Reading Growth for Students Pre-K through Grade 3. Dallas, TX: Istation.

Mathes, P. (2016). Istation's Indicators of Progress (ISIP) Advanced Reading Technical Report: Computer Adaptive Testing System for Continuous Progress Monitoring of Reading Growth for Students Grade 4 through Grade 8. Dallas, TX: Istation.

McNamara, D. S., & Magliano, J. (2009). Towards a comprehensive model of comprehension. *Psychology of Learning and Motivation*, 51, 297-384.

Millis, K., Magliano, J., & Todaro, S. (2006). Measuring discourse-level processes with verbal protocols and latent semantic analysis. *Scientific Studies of Reading*, 10(3), 225–240. Retrieved from https://doi.org/10.1207/s1532799xssr1003_2

Nation, K. (1999). Reading skills in hyperlexia: A developmental perspective. *Psychological Bulletin, 125*, 338-355.

Nation, K., Adams, J. W., Bowyer-Crane, C. A., & Snowling, M. J. (1999). Working memory deficits in poor comprehenders reflect underlying language impairments. *Journal of Memory and Language, 73,* 139-158.

National Center for Education Statistics. (2022). Students With Disabilities. *Condition of Education*. U.S. Department of Education, Institute of Education Sciences. Retrieved May 31, 2023, from https://nces.ed.gov/programs/coe/indicator/cgg.

National Reading Panel. (2000). *Teaching children to read: An evidence based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*. Bethesda, MD: National Institute of Child Health and Human Development.

Nierenberg, A. (2022, February 23, 2022). The National Guard deploys to classrooms. *The New York Times*. Retrieved from https://www.nytimes.com/2022/02/23/us/national-guard-teaching.html

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

Patarapichayatham, C. (2019). Linking the Colorado Measures of Academic Success English Language Arts (CMAS ELA) Assessments to ISIP™ Reading Assessments Grades 3 through 5. Dallas, TX: Istation. Available at www.istation.com/studies.

Patarapichayatham, C. (2018). Predictability Study of ISIP Reading and Virginia Standards of Learning (SOL) for English Reading: 3rd – 5th Grade Students. Dallas, TX: Istation. Available at www.istation.com/studies.

Patarapichayatham, C. (2017). Predictability Study of ISIP Reading and Kansas Assessment Program: 3rd – 6th Grade Students. Dallas, TX: Istation. Available at www.istation.com/studies.

Patarapichayatham, C. (2016). Predictability Study of ISIP Reading and Georgia Milestones Assessment System: 3rd – 6th Grade Students. Dallas, TX: Istation. Available at www.istation.com/studies.

Patarapichayatham, C. (2014). Predictability Study of ISIP Reading and STAAR Reading: Prediction Bands. Dallas, TX: Istation. Available at www.istation.com/studies.

Patarapichayatham, C., & Locke, V. N. (2022a). Comparability of ISIP Reading Scores Across Alternate Backgrounds: 2022 Update. Dallas, TX: Istation. Available from Istation.

Patarapichayatham, C., & Locke, V. N. (2022b). From disruption to recovery. Paper presented at the National Council on Measurement In Education, San Diego, CA.

Patarapichaytham, C., Locke, V., & Lewis, S. (2021a). Comparability of ISIP Reading Scores Across Alternate Backgrounds. Dallas, TX: Istation. Available at www.istation.com/studies.

Patarapichayatham, C., Locke, V. N., & Lewis, S. (2021b). Summer slide is bad, COVID-19 slide is even worse: Online assessment perspective. Paper presented at the National Council on Measurement in Education Virtual Annual Meeting. www.istation.com/studies.

Patarapichayatham, C., & Locke, V. N. (2020c). Linking the ACT Aspire Assessments to ISIP Reading and Math. Dallas, TX: Istation. Available at www.istation.com/studies.

Patarapichayatham, C. & Locke, V. N. (2020b). Linking the Ohio AIR to ISIP. Dallas, TX: Istation. Available at www.istation.com/studies.

Patarapichayatham, C. & Locke, V. N. (2020a). Linking Study Between STAAR Reading and ISIP ER Assessment for Second and Third Grade Students. Dallas, TX: Istation. Available at www.istation.com/studies

Patarapichayatham, C. & Wolf, R. (2022a). Linking Study Between the California Smarter Balanced Assessment and ISIP Reading. Dallas, TX: Istation. Available at www.istation.com/studies.

Patarapichayatham C. & Wolf, R. (2022b). Linking Study Between North West Education Association Measures of Academic Progress and ISIP Reading. Dallas, TX: Istation. Available October 31, 2022 at www.istation.com/studies.

Patz, R. J., & Yao, L. (2006). Vertical scaling: Statistical models for measuring growth and achievement. In C. R. Rao & S. Sinharay (Eds.), *Psychometrics* (Vol. 26). Elsevier.

Pearson, P. D., Hiebert, E. H., & Kamil, M. L. (2007). Vocabulary assessment: What we know and what we need to learn. *Reading research quarterly*, *42*(2), 282-296.

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

Perfetti, C. A. (1986). Continuities in reading acquisition, reading skill, and reading disability. *Remedial and special education, 7*(1), 11-21.

Perfetti, C. A. (1997). The psycholinguistics of spelling and reading.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye Movements as Reflections of Comprehension Processes in Reading. Scientific Studies of Reading, 10(3), 241–255. Retrieved from https://doi.org/10.1207/s1532799xssr1003_3

Risk, N. M. (2016). The Impact of Item Parameter Drift in Computer Adaptive Testing (CAT). Dissertation University of Illinois at Chicago.

Ruggles, S., Flood, S., Goeken, R., Schouweiler, M., & Sobek, M. (2022). *IPUMS USA:* Version 12.0.Minneapolis, MN: IPUMS, 2022. https://doi.org/10.18128/D010.V12.0

SAS Institute Inc. (2019). *SAS 9.4 language reference: Concepts, sixth edition*. SAS Institute Inc.

Sesma, H. W., Mahone, E. M., Levine, T., Eason, S. H., & Cutting, L. E. (2009). The contribution of executive skills to reading comprehension. *Child Neuropsychology, 15*, 232-246.

Scammacca, N., Roberts, G., Vaughn, S., Edmonds, M., Wexler, J., Reutebuch, C. K., & Torgesen, J. K. (2007). Interventions for adolescent struggling readers: A meta-analysis with implications for practice. *Center on Instruction*.

Share, D. L., & Stanovich, K.E. (1995). Cognitive processes in early-reading development: Accommodating individual differences into a model of acquisition. *Issues in Education*, 1, 1-57.

Shaywitz, S. E. (1996, November). Dyslexia. *Scientific American,* 98-104.

Shinn, M. R., Good, R. H., Knutson, N., Tilly, W. D., & Collins, V.L. (1992). Curriculum-based measurement reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review*, 21, 459-479.

Schumaker, J. B., Deshler, D. D., Alley, G. R., Warner, M. M., & Denton, P. H. (1982). Multipass: A learning strategy for improving reading comprehension. *Learning disability quarterly*, *5*(3), 295-304.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*(3), 237–247. https://doi.org/10.1111/j.1745-3984.1991.tb00356.x

Snow, C. E., Griffin, P. E., & Burns, M. (2005). *Knowledge to support the teaching of reading: Preparing teachers for a changing world*. Jossey-Bass.

Stanovich, K. E. (1991). Word recognition: Changing perspectives. In R. Barr, M. L. Kamil, P. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (Vol. 2, pp. 418-452). New York: Longman.

Swanson, H. L., Howard, C. B., & Saez, L. (2007). Reading comprehension and working memory in children with learning disabilities in reading. In K. Cain & J. Oakhill (Eds.), *Children's comprehension problems in oral and written language: A cognitive perspective* (pp. 157-189). New York: The Guilford Press.Sutter, C. C.,

Campbell, L. O., & Lambie, G. W. (2020). Predicting second-grade students' yearly standardized reading achievement using a computer-adaptive assessment. *Computers in the Schools, 37,* 1 40-54. https://doi.org/10.1080/07380569.2020.1720611

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement, 11*(4), 263–267. https://doi.org/10.1111/j.1745-3984.1974.tb00998.x

Thum, Y. M., & Hauser, C. H. (2015). *NWEA 2015 MAP norms for student and school achievement status and growth.* Retrieved from Portland, OR: NWEA

Tomek, S. (2018). Decision Consistency. In Frey, B. (Ed.) (2018). *The SAGE encyclopedia of educational research, measurement, and evaluation*. (Vols. 1-4). SAGE Publications, Inc., https://dx.doi.org/10.4135/9781506326139

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

Torgesen, J. K., Alexander, A. W., Wagner, R. K., Rashotte, C. A., Voeller, K. K., & Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. *Journal of learning disabilities*, *34*(1), 33-58.

Torgesen, J. K., Rashotte, C A., & Alexander, A. W. (2002). Principles of fluency instruction in reading: Relationships with established empirical outcomes. In M. Wolf (Ed.) *Time, Fluency, and Dyslexia*. Parkton, MD: York Press.

Torgesen, J. K., Rashotte, C. A., Wagner, R. K. (1999). *Test of Word Reading Efficiency*. Austin, TX: Pro-ED.

Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Burgess, S., & Hecht, S. (1997). Contributions of Phonological Awareness and Rapid Automatic Naming Ability to the Growth of Word-Reading Skills in Second-to Fifth-Grade Children. *Scientific Studies of Reading, 1*(2), 161-185. https://doi.org/10.1207/s1532799xssr0102_4

US Census Bureau. (2021). *American community survey, 2015-2019 5-year period estimates*. Retrieved from: www.census.gov

Van Dijk, T. A., & Kintsch, W. (1983). Strategies of discourse comprehension. New York: Academic Press.

Vellutino, F. R. (1991). Introduction to three studies on reading acquisition: Convergent findings on theoretical foundations of code-oriented versus whole-language approaches to reading instruction. *Journal of Educational Psychology,* 83, 437-443.

Wechsler, D. (2020). Wechsler Individual Achievement Test – Fourth Edition. San Antonio, TX: Pearson.

Wickham, H. (2021). *tidyverse: Easily install and load the "Tidyverse"* (1.3.1). The Comprehensive R Network.

Wolf, B. (2020a). Linking Istation ISIP Early Reading with the Idaho ISAT. Baltimore, MD: Johns Hopkins University Center for Research and Reform in Education. Available at http://jhir.library.jhu.edu/handle/1774.2/62380

Wolf, B. (2020b). Linking 2nd Grade Istation ISIP Reading with 3rd Grade ISAT in English Language Arts. Baltimore, MD: Johns Hopkins University Center for Research and Reform in Education. Available at http://jhir.library.jhu.edu/handle/1774.2/63123

Wolf M, Denckla MB. 2005. RAN/RAS: Rapid Automatized Naming and Rapid Alternating Stimulus Tests. Austin, TX: Pro-Ed.

Wolf, R. & Patarapichayatham, C. (2022). Linking the ISIP Reading and the New Jersey Student Learning Assessment. Dallas, TX: Istation. Available at www.istation.com/studies

World Programming Limited. (2022). *WPS workbench* (4.4.2). World Programming Limited.

Wright, B. D., & Douglas, G. A. (1975). Best test design and self-tailored testing (Research Memorandum No. 19). Chicago: University of Chicago, Department of Education, Statistical Laboratory.

Wright, B.D. & Douglas, G. A. (1976). Rasch item analysis by hand. Research Memorandum No. 21, Statistical Laboratory, Department of Education, University of Chicago.

Yen, W. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*, *23*, 299–325.

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.

Yuill, N., & Oakhill, J. (1991). Children's problems in text comprehension: An experimental investigation. Cambridge, UK: Cambridge University Press.

Zwaan, R. A., & Singer, M. (2003). Text comprehension. In A. C. Graesser, M. A. Gernsbacher, & S. R. Goldman (Eds.), *Handbook of Discourse Processes* (pp. 83–121). Lawrence Erlbaum Associates.