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

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

Istation

Supporting Educators. Empowering Kids.
Changing Lives.

2000 Campbell Centre II
8150 North Central Expressway
Dallas, Texas 75206
866.883.7323

www.istation.com

Table of Contents

Chapter 1: Introduction

ISIP™, Istation's Indicators of Progress, Early Reading (ISIP Early Reading) is a sophisticated, web-delivered Computer Adaptive Testing (CAT) system that provides Continuous Progress Monitoring (CPM) by frequently assessing and reporting student ability in critical domains of reading throughout the academic years. ISIP Early Reading is the culmination of many years of work begun by Joseph K. Torgesen, Ph.D. and Patricia G. Mathes, Ph.D. on extending computerized CPM applications to beginning readers.

Designed for students in Pre-Kindergarten through Grade 3, ISIP Early Reading provides teachers and other school personnel with easy-to-interpret, web-based reports that detail student strengths and deficits and provide links to teaching resources. Use of this data allows teachers to more easily make informed decisions regarding each student's response to targeted reading instruction and intervention strategies.

ISIP Early Reading provides growth information in the five critical domains of early reading: phonemic awareness, alphabetic knowledge and skills, fluency, vocabulary, and comprehension. It is designed to (a) identify children at risk for reading difficulties, (b) provide automatic continuous progress monitoring of skills that are predictors of later reading success, and (c) provide immediate and automatic linkage of assessment data to student learning needs, which facilitates differentiated instruction.

ISIP Early Reading has been designed to automatically provide continuous measurement of Pre-Kindergarten through Grade 3 student progress throughout the school year in all the critical areas of early reading, including phonemic awareness, alphabetic knowledge and skills, fluency, vocabulary, and comprehension, as mandated by the Elementary and Secondary Education Act, No Child Left Behind

(NCLB). Importantly, there is no other continuous progress monitoring assessment tool that measures vocabulary and comprehension. This is accomplished through short tests, or "probes," administered at least monthly, that sample critical areas that predict later performance. Assessments are computer–based, and teachers can arrange for entire classrooms to take assessments as part of scheduled computer lab time or individually as part of a workstation rotation conducted in the classroom. The entire assessment battery for any assessment period requires 40 minutes or less. It is feasible to administer ISIP Early Reading assessments to an entire classroom, an entire school, and even an entire district in a single day - given adequate computer resources. Classroom and individual student results are immediately available to teachers, 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.

The Need to Link Early Reading Assessment to Instructional Planning

Perhaps the most important job of schools and teachers is to ensure that all children become competent readers, capable of fully processing the meaning of complicated texts from a variety of venues. Reading proficiency in our information-driven society largely determines a child's academic, social, occupational, and health trajectory for the rest of his or her life. In a society that requires increasingly higher literacy skills of its citizenry, it cannot be stated strongly enough that teaching every child to read well is not an option, but a necessity. Every child who can read benefits society by being healthier, better informed, and fully employed.

Sadly, teaching every child to read is a goal we are far from achieving. Large numbers of our children continue to struggle to become competent readers (National Reading Panel, 2000; Lyon, 2005). Without adequate reading skills to comprehend and apply information from text, students frequently experience school failure. In fact, many students drop out of school as soon as they are able (Alliance for Excellent Education, 2006). The solution is to intervene when these students are in the early grades (Bryant et al., 2000).

There is a wide consensus about what comprises the elements of effective reading instruction (e.g., National Reading Panel, 2000; Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001; Snow, Burns, & Griffin, 1998). These elements are the same, whether the focus is prevention or intervention, and they include: phonemic awareness, alphabetic knowledge and decoding skills, fluency in word recognition and text processing, vocabulary, and comprehension (Foorman & Torgesen, 2001). Likewise, consensus on the predictors of reading difficulties is emerging from longitudinal databases (e.g., Fletcher, Foorman, Boudousquie, Barnes, Schatschneider, & Francis, 2002; O'Connor & Jenkins, 1999; Scarsborough, 1998; Torgesen, 2002; Vellutino, Scanlon, & Lyon, 2000; Wood, Hill, & Meyer, 2001).

It is well established that assessment-driven instruction is effective. Teachers who monitor their students' progress and use this data to inform instructional planning and decision-making have higher student

outcomes than those who do not (Conte & Hintze, 2000; Fuchs, Fuchs, Hamlett, & Ferguson, 1992; Mathes, Fuchs, Roberts, 1998). These teachers also have a more realistic conception of the capabilities of their students than teachers who do not regularly use student data to inform their decisions (Fuchs, Deno, & Mirkin, 1984; Fuchs, Fuchs, Hamlett, & Stecker, 1991; Mathes et al., 1998).

However, before a teacher can identify students at risk of reading failure and differentiate their instruction, that teacher must first have information about the specific needs of his or her students. To link assessment with instruction effectively, early reading assessments need to (a) identify students at risk for reading difficulties; students that may need extra instruction or intensive intervention if they are to progress toward grade-level standards in reading by year end; (b) monitor student progress for skill growth on a frequent and ongoing basis, and identify students falling behind; (c) provide information about students who will be helpful in planning instruction to meet their needs; and (d) assess whether students have achieved grade-level reading standards by year end.

In any model of instruction, for assessment data to affect instruction and student outcomes, it must be relevant, reliable, and valid. To be relevant, data must be available on a timely basis and target important skills that are influenced by instruction. To be reliable, there must be a reasonable degree of confidence in the student score. To be valid, the skills assessed must provide information that is related to later reading ability. There are many reasons why a student score at a single point in time under one set of conditions may be inaccurate: confusion, shyness, illness, mood or temperament, communication or language barriers between student and examiner, scoring errors, and inconsistencies in examiner scoring. However, by gathering assessments across multiple time points, student performance is more likely to reflect actual ability. By using the computer, inaccuracies related to human administration errors are also reduced.

The collection of sufficient, reliable assessment data on a continuous basis can be an overwhelming and daunting task for schools and teachers. Screening and inventory tools such as the *Texas Primary Reading Inventory* (TPRI: Foorman et al, 2005) and *Dynamic Indicators of Basic Early Literacy Skills* (DIBELS: Good & Kaminski, 2002) use a benchmark or screen schema in which testers administer assessments three times a year. More frequent continuous progress monitoring is recommended for all low-performing students, but administration is at the discretion of already overburdened schools and teachers.

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

Early Reading Assessments

To link assessment with instruction effectively, early reading assessments need to be both formative and individualized. One such approach is diagnostic assessment, which is typically administered by a reading specialist rather than a classroom teacher given the time requirements for administration. Examples include the *Diagnostic Assessment of Reading* (Roswell & Chall, 1992), *Developmental Reading Assessment* (Beaver, 1999), *Fox in the Box* (CTB/McGraw-Hill, 2000), and the *Qualitative Reading Inventory-II* (Leslie & Caldwell, 1995). Another approach is to collect authentic assessments designed to "reflect the actual learning and instructional activities of the classroom and out-of-school worlds" (Hiebert, Valencia, & Afflerbach, 1994). Examples of authentic assessment systems are: the *Observation Survey* (Clay, 1993); South Brunswick, *New Jersey, Schools' Early Literacy Portfolio* (Salinger & Chittenden, 1994); *The Primary Language Record* (PLR; Barr, Ellis, Tester, & Thomas, 1988) and *The California Learning Record* (CLR; Barr, 1995); *The Primary Assessment of Language Arts and Mathematics* (PALM; Hoffman, Roser, & Worthy, 1998); *The Work Sampling System* (Meisels, 1997); and *Phonological Awareness and Literacy Screening* (PALS; Invernizzi & Meier,1999).

The problems with these assessment approaches are that (a) most lack adequate reliability and validity; and (b) all are labor intensive to administer, making them simply unfeasible for progress monitoring. A more feasible approach has been to create screening tools that allow teachers and schools to discriminate those children who are at risk for reading failure from those who are at low risk for reading difficulties. Only children who appear to have risk characteristics receive further assessment. One such assessment is the *Texas Primary Reading Inventory* (TPRI; Foorman et al., 2005). With this assessment, only students who are at risk receive the full inventory, which is administered 3 times per year in Grades K-3. Even so, this assessment is still labor intensive for the teacher.

Perhaps the most visible approach to linking assessment data with instruction has been Continuous Progress Monitoring (CPM) using the model of Curriculum-Based Measurement (CBM: Fuchs, et al, 1984). Teachers use Curriculum-Based Measurement to index student progress over time. This is accomplished through the administration of short tests, or probes, administered at least once monthly, that sample critical areas that predict later performance. The relevant student performance information is the rate of change, displayed in graphic form, which illustrates each student's past, present, and probable future growth. More importantly, it alerts the teacher when a particular student is not making adequate progress that the instructional program can be modified.

The popular *Dynamic Indicators of Basic Early Literacy Skills* (DIBELS: Good & Kaminski, 2002) is built on of the Curriculum-Based Measurement model. The problem with current Curriculum-Based Measurement assessment is that it is very cumbersome for teachers to utilize (DiGangi, Jannasch-Pennell, Yu, Mudiam, 1999; Fuchs, Hamlet, & Fuchs, 1995). Presently, teachers have to physically administer probes to each child individually and either graph data by hand or enter data into a website (in the case of *DIBELS*) to access results. In order to reduce the burden on teachers, the authors of *DIBELS* have recently experimented with a hybrid model in which students are screened, and then only students not meeting

benchmark standards are assessed continuously. The remaining students are only assessed at benchmark points (beginning, middle, and end of year). Even with these concessions, teachers find *DIBELS* onerous. Also, *DIBELS* does not measure important constructs of vocabulary and comprehension.

Continuous Progress Monitoring

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

Computer Application

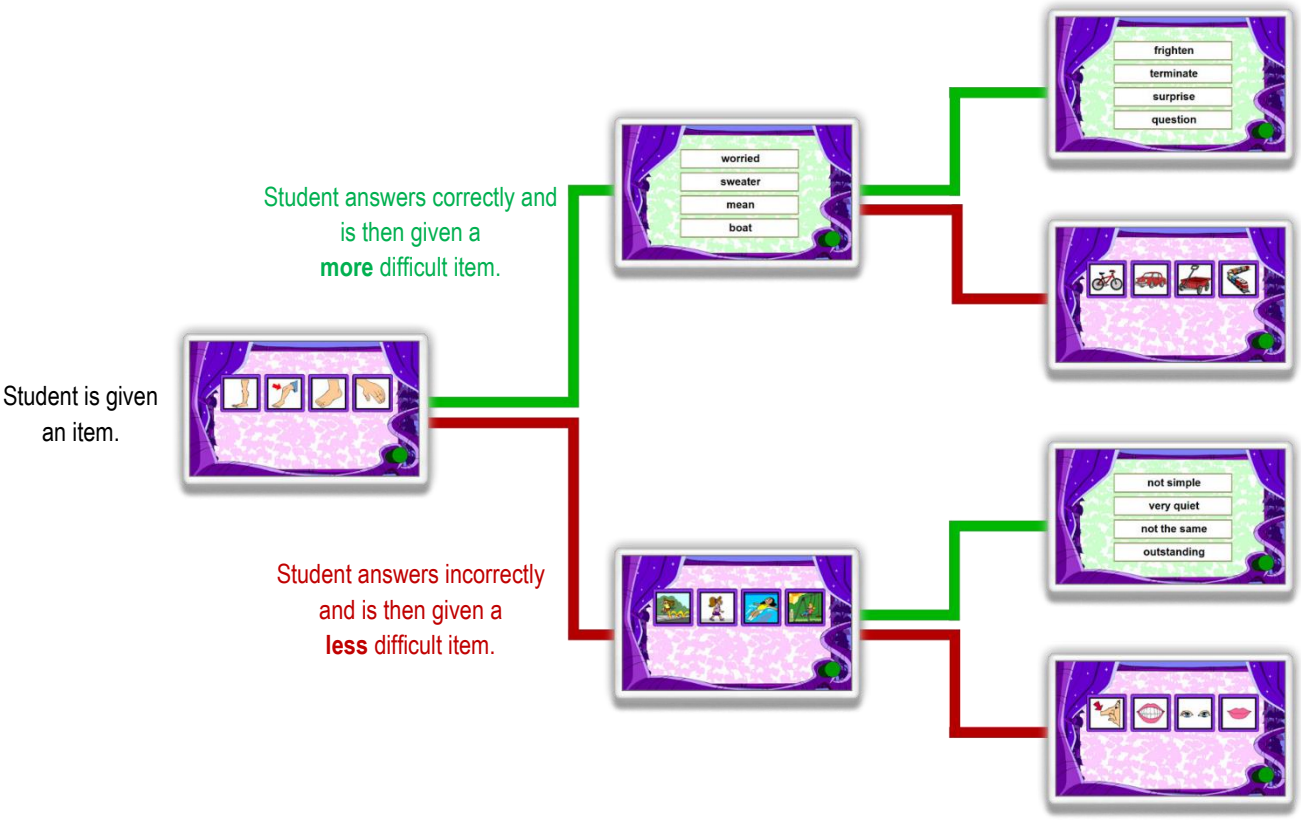
The problem with most CPM systems is that they have been cumbersome for teachers to utilize (Stecker & Whinnery, 1991). Teachers have to physically administer the tests to each child individually and then graph data by hand. The introduction of hand-held technology has allowed for graphing student results, but information in this format is often not available on a timely basis. Even so, many teachers find administering the assessments onerous. The result has been that CPM has not been as widely embraced as would be hoped, especially within general education. Computerized CPM applications are a logical step to increasing the likelihood that continuous progress monitoring occurs more frequently with monthly or even weekly assessments. Computerized CPM applications using parallel forms have been developed and used successfully in upper grades in reading, mathematics, and spelling (Fuchs et al., 1995). Computerized applications save time and money. They eliminate burdensome test administrations and scoring errors by calculating, compiling, and reporting scores. They provide immediate access to student results that can be used to affect instruction. They provide information organized in formats that automatically group children according to risk and recommended instructional levels. Student results are instantly plotted on progress charts with trend lines projecting year-end outcomes based upon growth patterns, eliminating the need for the teacher to manually create monitoring booklets or analyze results.

Computer Adaptive Testing

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

performance abilities of the students. This also means that students who are achieving significantly above or below grade expectations can be assessed to more accurately reflect their true abilities.

There are many advantages to using a CAT model rather than a more traditional parallel forms model, as is used in many early-reading instruments. For instance, it is virtually impossible to create alternate forms of any truly parallel assessment. The reliability from form to form will always be somewhat compromised. However, when using a CAT model, it is not necessary that each assessment be of identical difficulty to the previous and future assessments. Following a CAT model, 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) work. Once item parameters have been determined, the CAT algorithm can be programmed. Then, using this sophisticated computerized algorithm, the computer 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 computer is able to generate "probes" that have higher reliability than those typically associated with alternate formats and that better reflect each student's true ability.



ISIP Early Reading Assessment Domains

ISIP Early Reading uses a CAT algorithm that tailors each assessment to the performance abilities of individual children while measuring progress in the five critical early reading skill domains of (a) phonemic awareness, (b) alphabetic knowledge and skills, (c) connected text fluency, (d) vocabulary, and (e) comprehension.

Phonemic Awareness

Phonemic awareness refers to the understanding that spoken words are comprised of individual sounds called phonemes. This awareness is important because it underpins how sound-symbols in printed words map onto spoken words. Deficits in phonemic awareness characterize most poor readers, whether they are children, adolescents, or adults (at all levels of intelligence) and whether or not they are from economically disadvantaged or non-English speaking backgrounds (Share & Stanovich, 1995).

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. Today, it is understood that reading problems for most children occur at the level of the single word because of faulty or incomplete alphabetic knowledge and skills. In fact, 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).

Text Fluency

Beyond phonological and alphabetic knowledge, children must be able to read connected text with relative ease if the meaning of that text is to be accessed and the development of mature comprehension strategies are to prosper (Torgesen, Rashotte, & Alexander, 2002). When fluency-building activities are utilized during instruction, children's fluency does increase (Torgesen et al., in press, 2001). Teachers need to know which children are not making desired gains in fluency if they are to know that fluency strategies need to be incorporated.

Vocabulary and Comprehension

The ultimate goal of all reading is to ensure that children comprehend what they read. Thus, there is increasing understanding that it is not enough to only teach children to decode words. Increasingly, there is a greater focus on the need to ensure that children possess an adequate vocabulary and comprehension

strategies to allow them to process text for meaning. This is especially true for children from lower socioeconomic backgrounds and from households in which English in not the primary language. Teachers need to know (a) if children have vocabulary deficits that place them at risk for failing to comprehend what they read, (b) if instruction is having the desired effect of raising students' vocabulary knowledge, and (c) if students are making progress in comprehending increasingly challenging materials.

ISIP Early Reading Items

The purpose of the ISIP Early Reading Item Bank is to support teachers' instructional decisions. Specifically, the item bank is designed to serve as a computerized adaptive universal screening and progress monitoring assessment system. By administering this assessment system, teachers and administrators can use the results to answer two questions: (1) are students in grades Pre-K through 3rd grade at risk of failing reading, and (2) what is the degree of intensity of instructional support students need to be successful readers? Because the assessment is designed to be administered, these decisions can be applied over the course of the school year.

Along with the authorship team, graduate students from the Institute for Evidence-Based Education at Southern Methodist University (SMU) were involved in item development by asking the following question: What are the best ways to assess the domains of reading students via computer administration? Knowing that students, depending on their grade, need to be assessed in Listening Comprehension, Phonemic Awareness, Letter Knowledge, Alphabetic Decoding, Spelling, Fluency, Vocabulary, and Reading Comprehension, a search of the literature was completed to locate studies that focused on how to best assess each of these dimensions of reading, as well as possible confounds to the design of these assessments. An extensive search of the literature base on how to best assess each of the areas was conducted to provide the team clarity about the then current understanding about assessment techniques for assessing these reading domains. Much time was spent defining models for each of the constructs and designing items to assess the models. It was further examined how each of the reading domains had been assessed in other widely accepted assessments. Armed with this information, the team met frequently to discuss the pros and cons of various formats and ideas for how best to assess each domain in order to reflect the model through computer administration of items.

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 early release of the Common Core State Standards and state standards for California, Florida, New York, Virginia, and Texas, were reviewed for Grades K-3 and Pre-K when available. The standards were listed by grade, reading domain, and 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 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 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 were 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 the end of Pre-K and as high as Grade 5/6. Additionally, items were developed within each domain to represent easy, moderate, and hard items for each grade. This wide range of items make ISIP Early Reading an appropriate measure for the full range of students, including students with special needs or who struggle and students who are high-achieving or gifted. While ultimately the IRT calibration work identified the difficulty of each item, the team was assured of having items representing the full continuum of achievement for each domain.

The use of CAT algorithms also creates efficiencies in test administration. The adaptive item algorithm allows the computer to adjust item difficulty while the child is taking the test, quickly zeroing in on ability level. Thus, the use of CAT algorithms reduces the amount of time necessary to accurately determine student ability.

Accuracy and Fluency

Within ISIP Early Reading, 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. Such fluency-based assessments have been proven to be efficient, reliable, and valid indicators of reading success (Fuchs et al. 2001; Good, Simmons, & Kame'enui, 2001). 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. In order to be fluent at higher-level processes of reading connected text, a student will also need to be fluent with foundational skills. *DIBELS* is the most widely used early reading assessment that incorporates a fluency component into each of its subtests.

Because each of the subtests has a fluency component, the tests are brief. This makes it feasible to administer the subtests on a large scale with minimal disruption of instructional time. Numerous items are available for each subtest, making the subtests repeatable throughout the school year with many alternative forms.

Teacher Friendly

ISIP Early 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 result reliability. In particular, teachers are alerted to observe specific

students identified by ISIP Early Reading as experiencing difficulties as they complete ISIP Early Reading. They subsequently review student results to validate outcomes. For students whose skills may be a concern, based upon performance level, teachers may easily validate student results by re-administering the entire ISIP Early Reading battery or individual skill assessments.

Child Friendly

ISIP Early Reading is also child friendly. Each assessment session feels to a childlike he or she is playing a fast-paced computer game called "Show What You Know." In the beginning of the session, an animated owl named Smart Owlex Treebeak enters the screen with his assistant, Batana White, a female bat. The owl announces to the child in a game show announcer voice, "It's time to play… Show What You Know!" A curtain pulls back to show the first game. The owl announces the game quickly, and the assessment begins. At the end of the assessment, the child sees an animated graph of progress. Each assessment proceeds in a similar fashion.

ISIP Early Reading Subtests

ISIP Early Reading measures progress in each critical component of reading instruction in a manner appropriate to the underlying domain. There are a total of 8 subtests that align to the 5 critical domains of reading, as shown in the table below. Of these subtests, 6 are built using a CAT algorithm, while 2 use parallel forms. Subtests that tailor items using CAT include Listening Comprehension, Phonemic Awareness, Letter Knowledge, Alphabetic Decoding, and Spelling, Vocabulary, and Reading Comprehension. Connected Text Fluency is designed as a parallel forms assessment that measures end of grade level expectations.

Domain	Subtest
Phonemic Awareness	Phonemic Awareness
Phonics	Letter Knowledge Alphabetic Decoding Spelling
Vocabulary	Vocabulary
Comprehension	Listening Comprehension Reading Comprehension
Fluency	Text Fluency

ISIP Early Reading Administration Format

ISIP Early Reading is presented to students using a game-like format. Students are never told that they are being given a test. Instead, they are told that they are playing a game called "Show What You Know."

The first time a student takes ISIP Early Reading, the computer will administer items that are defaulted based on the student's grade, unless the default setting is changed intentionally, as may be appropriate in special education settings. From the very first item, however, the CAT engine immediately begins to tailor the test to the individual student. As a result, students will only be administered subtests that are appropriate for their performance abilities. Within a classroom, students may have some variation in the exact subtest they are administered. However, scores reflect these differences (explained below). For example, students whose performance scores indicate that they are not yet reading words will not be asked to read connected text. Likewise, 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.

Listening Comprehension is administered only in PreK and Kindergarten. In Grade 1, Text Fluency is administered only after students obtain a high enough score on Alphabetic Decoding to suggest that they can handle the task. Connected Text Fluency is administered to all students, beginning in Grade 2

The table below presents the defaults for subtest administration for each grade level.

Grade	Subtest
Pre-Kindergarten	Listening Comprehension Phonemic Awareness Letter Knowledge Vocabulary
Kindergarten	Listening Comprehension Phonemic Awareness Letter Knowledge Vocabulary
1st Grade	Phonemic Awareness Letter Knowledge Vocabulary Alphabetic Decoding Reading Comprehension Spelling
2nd and 3rd Grade	Vocabulary Reading Comprehension Spelling Text Fluency

Rationale for Subtest Defaults by Grade

ISIP Early Reading follows a continuum of learning that, research indicates, is predictive of later reading success. Skills build upon skills, and the sequence of subtests builds upon prior subtests. As skills of lower-level difficulty are eliminated from the test battery, more difficult skills that rely on achievement of the prior skills are added.

Because ISIP Early Reading incorporates computer-adaptive algorithms, students are administered items of increasing difficulty until either an appropriate level of ability is established or it is determined through other higher-level subtests that skill mastery has been achieved. Thus, defaults are only a starting point. Once ISIP Early Reading calibrates to the performance ability of a particular student, each subsequent test relies on the previous calibrations to determine with which items to begin subsequent administrations.

PreK and Kindergarten

Kindergarten students require assessment of their growth in listening comprehension, phonemic awareness, alphabetic knowledge and skills, and vocabulary. Fluency in letter names and sounds facilitates spelling, but these skills are usually not developed sufficiently to assess spelling ability. Their reading skills are also rarely sufficiently developed to usefully assess reading fluency and reading comprehension. In general, research has shown that phonological awareness and letter sound knowledge in Kindergarten are predictive of Grade 1 outcomes. For children at risk of reading difficulty due to poverty or language background, vocabulary is critical to reading success (Foorman, Anthony, Seals, & Maouzaki, in press; Snow et al., 1998; Dickinson & Tabors, 2001). Vocabulary assessments for Kindergarten students are mostly "tier 1" words and items to assess children's knowledge of prepositions and verbs of varying tense, since these classes of words are particularly difficult for young children.

Grade 1

It is important to continue to monitor students' development of phonemic awareness and alphabetic knowledge and skill, because struggling students may continue to have difficulty in these areas. The development of accurate and fluent decoding skills should be monitored, since these foundational skills for reading accuracy undergo major development. Word recognition at the beginning of Grade 1 has been found to be predictive of Grade 1 outcomes. Spelling has also been found to be a predictor of oral reading fluency. Vocabulary growth is important in the development of reading comprehension. As soon as students can demonstrate the ability to read connected text with reasonable accuracy and understanding, reading fluency (timed reading with meaning) should be monitored. Continued growth in Vocabulary should be assessed, as well as Reading Comprehension.

Grade 2

In Grade 2, word reading continues to be a strong predictor of Grade 2 outcomes, with reading fluency and comprehension becoming increasing important predictors. Second graders need continued monitoring of their decoding abilities because struggling students may still have difficulty in this area. Reading fluency is critical through Grade2 since students must make strong growth in this skill to maintain grade level reading proficiency. The development of reading comprehension is dependent on fluency and vocabulary. Sight vocabulary must grow rapidly in second grade to keep pace with expected reading outcomes. Thus, continued growth in Spelling, Vocabulary, and Reading Comprehension should be measured.

Grade 3

In Grade 3, reading fluency and comprehension are strong predictors of Grade 3 outcomes. The primary dimensions of reading growth that should be measured in Grade 3are Reading Fluency, Reading Comprehension, Spelling, and Vocabulary.

Because reading fluency and comprehension are key predictors of later reading success, instructional recommendations are based upon consistency of risk levels across these subtests. Greater weight is placed on the higher-risk measure. Students with mixed results are typically recommended for strategic instruction.

Description of Each Subtest

Listening Comprehension

In this subtest, children 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 read aloud.

Matching Sentences and Pictures.

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 listens to the sentence and identifies the picture that best illustrates the orally read sentence's meaning.

Sentence and Story Completion

Sentence 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 four choices, the word that best completes the sentence or story.

Phonemic Awareness

The Phonemic Awareness subtest is comprised of 2 types of items: Beginning, Ending and Rhyming Sounds and Phonemic Blending.

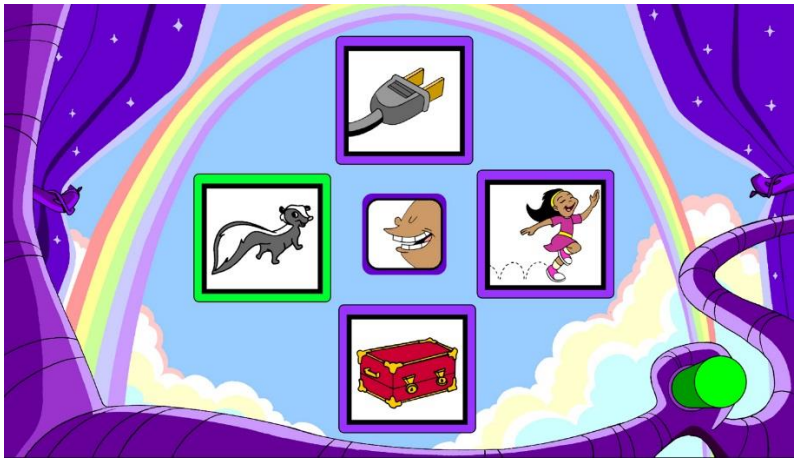
Beginning, Ending and Rhyming Sounds

Beginning Sound assesses a student's ability to recognize the initial, final or rhyming sound in an orally presented word. Four items appear on the screen at once. The narrator says the name of each picture as the box around it highlights. Then the student is asked to click on the picture that has the same beginning,

ending, or rhyming sound as the sound produced orally by the narrator. The student may move the mouse pointer over a picture to hear the picture name repeated.

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. 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 produces 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 in order to give the test a top range.



Letter Knowledge

Letter Knowledge represents the most basic level of phonics knowledge (i.e. whether students know the names and sounds represented by the letters of the alphabet). Letter knowledge is comprised of two types of items: recognition of letter names and recognition of letter-sound correspondences. It is

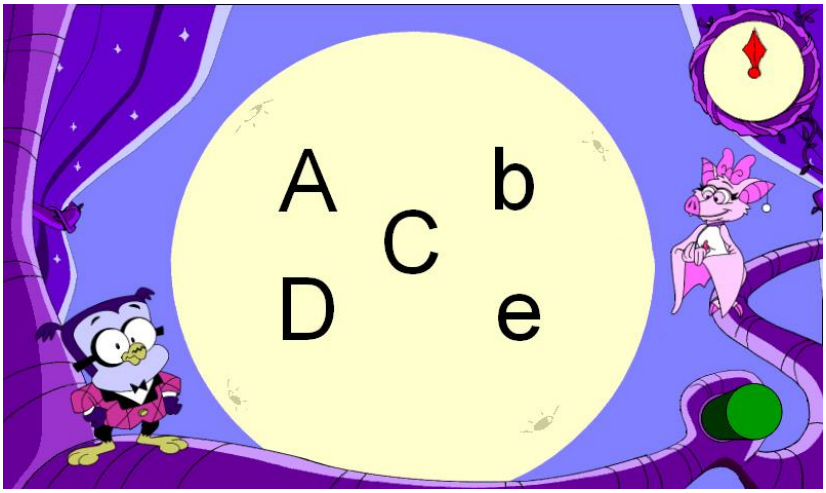
important to note that only the most frequent letter-sound correspondences are included in this subtest. More complex elements such as variant spellings, diphthongs, vowel teams, and r-controlled vowels are embedded in the Alphabetic Decoding and Spelling subtests.

Letter Recognition

Letter Recognition is a measure of alphabetic principle that assesses how many letters a student can correctly identify in a minute. Five items, in a combination of both uppercase and lowercase letters, appear on screen at once. The student is asked to identify the symbol for the letter name that is orally produced by the narrator.

Letter Sound

Letter Sound 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 on screen at once. The student is asked to identify the symbol for the letter sound that is orally produced by the narrator.

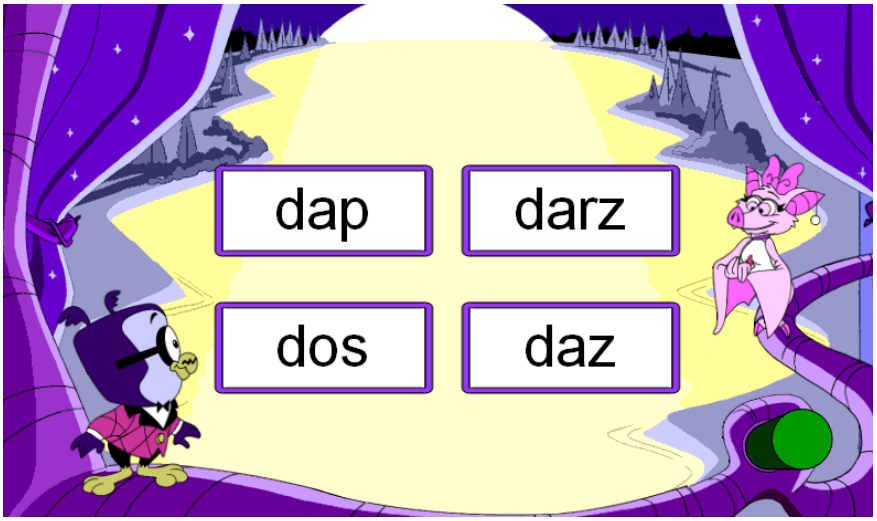


Alphabetic Decoding

Alphabetic Decoding

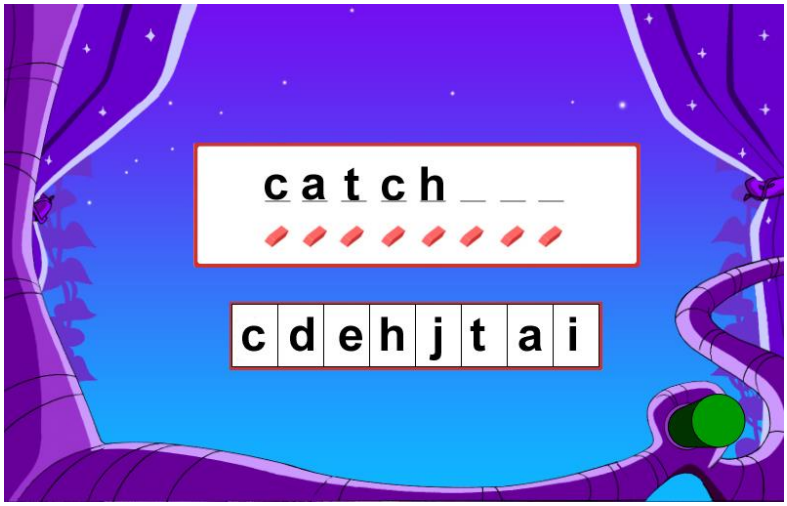
Alphabetic Decoding measures the ability to blend letters into nonsense words in which letters represent their most common sounds. Nonsense words are used because students differ in their sight word recognition skills. By using nonsense words, the test more accurately assesses the ability to match letters to sounds and the ability to decode an unknown word when it is presented. For this subtest, four items appear on the screen. 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.

The sequence of difficulty moves in the following manner: (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 (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 2 syllable words (e.g., *cvc/cvc – webbet; cv/cvc – tebet*).



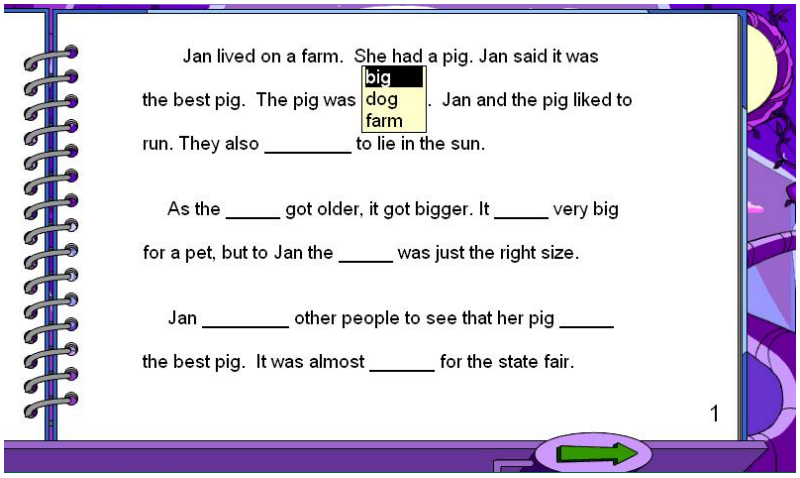
Spelling

Research has shown that learning to spell and learning to read rely on much of the same underlying knowledge, such the relationships between letters and sounds. Knowing the spelling of a word makes the representation of it sturdy and accessible for fluent reading (Ehri, 2000; Snow et al. 2005). The objective of the Spelling subtest is to determine if children are developing fully-specified orthographic representations of words. For each item, an array of letters appears on the screen, and the computer asks the child to spell a specific word using those letters. The child then spells the word by clicking on each letter. As each letter is selected, the word is formed on lines above the letter array. Items for this subtest have been carefully constructed to move from easier to harder, using the sequence of difficulty defined in Alphabetic Decoding. However, item parameters also include frequency of spelling patterns, with less frequent spelling patterns being considered more difficult. Two hundred spelling items spread across five levels of difficulty have been validated.



Text Fluency

Text Fluency measures a child's ability to read fluently with comprehension. This subtest is constructed in a very different manner than others. Rather than increasing text difficulty across time, the test assesses children on passages of equivalent difficulty to measure growth over time against a constant level of difficulty. Each of these passages was carefully written to conform to specific word level features, follow linear story grammar structure, and have readability according to a commonly accepted readability formula for end of grade level in each grade. In order to assess text reading on the computer, a maze task is utilized, in which every fifth or sixth word is left blank from the text. For each blank, the child is given three choices from which to choose the word that works in the sentence. It is the child's job to read the text, selecting 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, Deno, Maruyama, & Cohen, 1989; Fuchs & Fuchs, 1990; Jenkins, Pious, & Jewell 1990; Shinn, Good, Knurson, Tilly, Collins, 1992).



Vocabulary

The vocabulary subtest is designed to test a child's knowledge of "tier 2" vocabulary words, meaning words that are frequently encountered in text but are not typically used in daily conversation (Beck, McKeown, & Kucan, 2002). There are two formats: Pictures and Synonyms.

Picture Items

On picture items, pictures appear on the screen. The narrator asks the student to identify the picture that best illustrates the word that is orally produced by the narrator.

Synonym Items

To establish the upper bound of vocabulary development, an alternative synonym format is used. Four words appear on screen. The student is asked to identify the word that has the same or similar meaning as a target word pronounced by the narrator. The narrator says each of the four word choices as the box around it highlights.

Comprehension

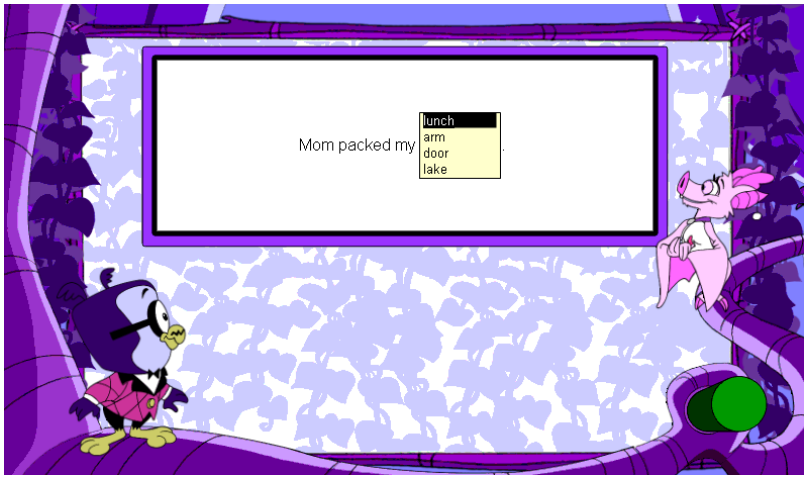
In this subtest, children are assessed on their ability to read and understand grade-level sentences and paragraphs. This is accomplished through matching sentences and pictures and sentence completion tasks.

Matching Sentences and Pictures.

Matching sentences and pictures assesses a student's knowledge of semantic and syntactic information where pictures can support their reading. In this task, a sentence and four pictures appear on the screen. The student reads the sentence and identifies the picture that best illustrates the sentence meaning.

Sentence Completion

Sentence completion measures a student's ability to use word meanings and word order to understand a sentence. In this task, a sentence, sentences, or a paragraph appears on screen. One word is missing from the text. The student reads the text and must choose, from four choices, the word that best completes the text.



The ISIP Early Reading Link to Instructional Planning

ISIP Early Reading provides continuous assessment results that can be used in recursive assessment instructional decision loops. First, ISIP Early Reading identifies students in need of support. Second, validation of student results and recommended instructional levels can be easily verified by re-administering assessments, which increases the reliability of scores. Teachers can assign assessments to individual students at the Istation website at *www.istation.com*. The student logs in to the assessment, and it is automatically administered.

Third, the delivery of student results facilitates the evaluation of curriculum and instructional plans. The technology underlying ISIP Early Reading delivers real-time evaluation of results and immediate availability of reports on student progress upon assessment completion. Assessment reports automatically group students according to the level of support needed as well as skill needs. Data is provided in both graphical and detailed numerical format on every measure and at every level of a district's reporting hierarchy. Reports provide summary and skill information for the current and prior assessment periods that can be used to evaluate curriculum, plan instruction and support, and manage resources.

At each assessment period, ISIP Early Reading automatically alerts teachers to children in need of instructional support through email notification and the "Priority Report." Students are grouped according to instructional level and skill need. Links are provided to teacher-directed plans of instruction for each instructional level and skill category. There are downloadable lessons and materials appropriate for each group. When student performance on assessments is below the goal for several consecutive assessment periods, teachers are further notified. This is done to raise teacher concern and signal the need to consider additional or different forms of instruction.

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

In addition to the recommended activities, Reading Coaches and Teachers have access to an entire library of teacher-directed lessons and support materials at *www.istation.com*. Districts and schools may also elect to enroll students in Istation's computer-based reading and intervention program, The Imagination Station. This program provides individualized instruction based upon ISIP Early Reading results. Student results from The Imagination Station are combined with ISIP Early Reading results to provide a deeper student profile of strengths and weaknesses that can enhance teacher planning.

All student information is automatically available by demographic classification and by specially designated subgroups of students who need to be monitored.

A year-to-year history of ISIP Early Reading results is available. Administrators, principals, and teachers may use their reports to evaluate and modify curriculum, interventions, AYP progress, the effectiveness of professional development, and personnel performance.

Chapter 2: IRT Calibration and the CAT Algorithm

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

During the 2007-08 school year, data were collected from two large north Texas independent school districts (ISD), labeled AISD and BISD henceforth. Five elementary schools from each district were recruited for the study. At each school, all Kindergarten through Grade 3 students in general education classrooms were asked to bring home introductory letters and study consent forms, which had prior approval by both the school districts and Southern Methodist University's institutional review board. Table 2-1 shows the number of students at each school and the number of students with signed consent forms who participated.

Table 2-1: Number of Students in Study

School District	Signed Consent Forms	Total Students	Percent of Students with Signed Consent Forms
AISD	615	999	61.56
A.1	108	210	51.43
A.2	212	274	77.37
A.3	107	205	52.20
A.4	70	180	38.89
A.5	118	130	90.77
BISD	1,002	1,301	77.02
B.1	79	165	47.88
B.2	306	362	84.53
B.3	158	222	71.17
B.4	227	304	74.67
B.5	232	248	93.55
TOTAL	1,617	2,300	70.30

Both districts represented socially and ethnically diverse populations. Table 2-2 shows the demographics of participating students from each district.

Table 2-2: Demographics of Participating Students

	AISD		BISD		Study	
	Number in Category	Percent of Students	Number in Category	Percent of Students	Number in Category	Percent of Students
Total	615		1,002		1,617	
Kindergarten	130	21.14	238	23.75	368	22.76
1st Grade	164	26.67	257	25.65	421	26.04
2nd Grade	143	23.25	287	28.64	430	26.59
3rd Grade	178	28.94	220	21.96	398	24.61
Gender						
Male	271	44.07	533	53.19	804	49.72
Female	344	55.93	469	46.81	813	50.28
Ethnicity						
White	39	6.34	372	37.13	411	25.42
Hispanic	273	44.39	227	22.65	500	30.92
African American	288	46.83	230	22.95	518	32.03
Asian	11	1.79	162	16.17	173	10.70
American Indian	2	0.33	7	0.70	9	0.56
Unknown	2	0.33	4	0.40	6	0.37
Receiving ESL Services	122	19.84	305	30.44	427	26.41
Receiving Free/ Reduced Lunch	547	88.94	421	42.02	968	59.86
Receiving Special Ed Services	49	7.97	60	5.99	109	6.74

Students were escorted by trained SMU data collectors, typically graduate students, in convenience groupings to the school's computer lab for 30-minutes sessions on the ISIP Early Reading.

It was unrealistic to administer all the items to each student participating in the study. Therefore, items were divided into a relatively lower difficulty subpool and a higher difficulty subpool by content experts. Students in Kindergarten and 1st Grade (K-1) were given 970 ISIP items from 8 skill groups. Students in 2nd and 3rd Grades (2-3) were given 750 items. Included in each total are 148 overlapping items that were given to all students, Kindergarten through 3rd Grade (K-3), and used for comparison and vertical scaling. Table 2-3 shows the numbers of items given to the students in the study.

Table 2-3: Items Used in the Study

Skill	K-1	Overlap (K-3)	2-3
Beginning Sound	112	11	0
Phonemic Blending	83	19	87
Vocabulary	90	27	151
Comprehension	88	18	138
Alphabetic Decoding	102	23	105
Spelling	79	22	121
Letter Sound	110	12	0
Letter Recognition	158	16	0
TOTAL	822	148	602

The items in each grade group were divided into 12 blocks, each taking approximately 30 minutes to complete. The blocks were divided into 4 treatments using a cyclic Latin squares design in order to control for order main effects. Participating students were randomly assigned to one of the 4 treatments by Istation staff creating the student login accounts. ISIP Early Reading was programmed to automatically follow the treatment order based on the assigned treatment group.

Testing at AISD took place between January 2008 and May 2008. Testing at BISD took place between November 2007 and February 2008. Ideally, students were tested twice weekly for 6 consecutive weeks. However, circumstances occasionally arose which precluded testing for a given student or for groups of students, including absences, assemblies, and holidays. When testing did not occur for a group of students, additional testing sessions were added to the end of the schedule. As a rule, when 95% of the students at a school completed all 12 sessions, testing stopped at that school. After testing was completed, on average there were approximately 800 responses per item.

Data Analysis and Results

Due to the sample size for each item, a 2-parameter logistic item response model (2PL-IRT) was posited. We define the binary response data, x_{ij}, with index i=1,...*n* for persons, and index *j*=1,...*J* for items. The binary variable $x_{ij} = 1$ if the response from student *i* to item *j* was correct and $x_{ij} = 0$ if the response was wrong. In the 2PL-IRT model, the probability of a correct response from examinee *i* to item *j* is defined as

$$P(x_{ij} = 1) = \frac{\exp[\lambda_j(\theta_i - \delta_j),}{1 + \exp[\lambda_j(\theta_i - \delta_j)]}$$

where θ_i is examinee i's ability parameter, δ_j is item *j*'s difficulty parameter, and λ_j is item *j*'s discrimination parameter.

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

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

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

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

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

Variance parameters τ_θ and τ_δ each follow a conjugate inverse gamma prior to introduce more flexibility,

$$\tau_\theta \sim \text{IG}(a_\theta, b_\theta),$$

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

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

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

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

where the hyperparameters were defined as $a_\lambda = b_\lambda = 1$.

The Gibbs sampler, a Bayesian parameter estimation technique, was employed to obtain item parameter estimates by way of a Fortran program. Several items did not have a sufficient sample size to produce reliable estimates and were subsequently removed from future analyses. The resulting analysis produced two parameter estimates for each of the 1,550 items, a difficulty parameter as well as a discriminability parameter, which indicates how well an item discriminates between students with low reading ability and students with high ability.

In the study, we implemented the common-item nonequivalent groups design for the 1,550 items that had reliable parameter estimates. The parameter estimates for the 2-3 item group were transformed to the scale for the K-1 item group by using results from the 148 overlapping K-3 items using the mean/mean procedure (Kolen & Brennan, 2004). Equations above show the ranges of estimates for each parameter for the subtests developed for calibration: Beginning Sound, Comprehension, Letter Recognition, Letter Sound, Phoneme Blending, Spelling, Vocabulary Level 1, Vocabulary Level 2, and Alphabetic Decoding.

The Pearson product moment correlation coefficient between the difficulty and discriminability parameters was effectively zero (r = -0.0029).

Distributions of each parameter by skill were approximately normal. Subsequently, 95% confidence intervals (95CI) around each mean were computed. Items with parameters outside of the 95CI were examined by a panel of content experts, and all were determined to be valid items testing at the appropriate level. Therefore, 1,550 items were used for the ISIP Early Reading item pool.

Overall most items are in good quality in terms of item discriminations and item difficulties. The reliability is computed from IRT perspective by using this formula: $\rho^2 = 1 - [\mathrm{SE}(\theta)]^2$, where θ is the student ability. It is 0.891, indicating that ISIP Early Reading is very reliable. The standard error of measurement (SEM) is also computed from IRT point of view. Since the ISIP Early Reading scale score is $(20 * \theta) + 200$, $SEM(\theta) = 20 * \mathrm{SE}(\theta)$. It is 6.593.

CAT Algorithm

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

The algorithm is as follows:

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

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

Though the CAT algorithm is simple, it allows for endless variations on item selection criteria, stopping criteria and ability estimation methods. All of these elements play into the predictive accuracy of a given implementation and the best combination is dependent on the specific characteristics of the test and the test takers. In developing Istation's CAT implementation, we explored many approaches. To assess the various approaches, we ran CAT simulations using each approach on a large set of real student responses to our items (1,000 students, 700 item responses each). To compute the "true" ability of each student, we used Bayes expected a posteriori (EAP) estimation on all 700 item responses for each student. We then compared the results of our CAT simulations against these "true" scores to determine which approach was most accurate, among other criteria.

Ability Estimation

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

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

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

Item Selection

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

Stopping Criteria

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

Production Assessment

Item types were grouped according to key reading domains for the productions assessment. Beginning sound and phoneme blending were combined in to the Phonemic Awareness (PA) domain. Letter recognition and sounds were combined in to the Letter Knowledge (LK) domain. All vocabulary items were combined in to a single Vocabulary (VOC) domain.

Each grade-level (Kindergarten, 1st, 2nd, etc…) was given a different set of subtests depending on the domains expected by grade:

- **K**: Phonemic Awareness, Letter Knowledge, and Vocabulary
- **1st**: Phonemic Awareness, Letter Knowledge, Alphabetic Decoding, Vocabulary, Spelling, and Comprehension
- **2nd**: Alphabetic Decoding, Vocabulary, Spelling, and Comprehension
- **3rd**: Alphabetic Decoding, Vocabulary, Spelling, and Comprehension

These subtests were administered sequentially and treated as independent CAT tests. Items were selected from the full, non-truncated, item pool for each subtest, so students were allowed to demonstrate their ability regardless of their grade level. Each subtest has its own ability estimate and standard error, with no crossing between the subtests. After all subtests were complete, an overall ability score was computed by running EAP on the entire response set from all subtests. Each subtest used its own previous ability score to offset the standard normal prior used in EAP.

Scale scores used in the reporting of assessment results were constructed by a linear transformation of the raw ability scores (logits). The study resulted in a pool of 1,550 Kindergarten through Grade 3 items with

reliable parameter estimates aligned on a common scale with the majority of items ranging from 140 to 289 in difficulty. See Figure 2-A for sample items at various scale bands.

Figure 2-A: Sample Items from ISIP Early Reading

	below 140	140-169	170-199	200-229	230-259	260-289	above 289
Vocabulary Knowing high frequency words and synonyms	brushing (picture)	car (picture)	saddle (picture)	grateful (synonym)	admire (synonym)	dwell (synonym)	protrude (synonym)
Letter Knowledge Recogniz ing letter names and sounds	x (name)	h (name)	q (name) f (sound)	 E (sound)			
Phonemic Awareness Recognizing initial sounds and blending phonemes		r u g c _ a t	nest b_oo_k	boat a_n_i_m_a_l			
Alphabetic Decoding Recognizing phonemes from non-words			nol	fom	brimert	bripfuscate	fornalibe
Spelling Constructing words from letters and punctuation			love	some	I'll	rifle	they're

	below 140	140-169	170-199	200-229	230-259	260-289	above 289
Comprehension Reading and deriving meaning from words and sentences			The girl is jump- ing on the bed. (select from a series of pictures)	Beth earned washing dishes and cleaning her room. (select from a list of words) All of Ann's friends were busy. Nick was playing ball.... Jo was buying new shoes. Ann felt	All of Ann's friends were busy. Nick was playing ball.... Jo was buy- ing new shoes. Ann felt (select from a list of words)	A weath- ered old fisherman and his lively and jolly wife lived in a small cot- tage by the sea.... But lately his luck had not been as good. His wife's heart was sad for her husband.... She was hoping that he might have had better (select from a list of words)	Scotland is un- doubtedly one of the most beautiful countries in the world.... Perhaps Scotland is best known for its many lakes, called lochs, which reflect the turquoise and azure blue of the skies. Scot- land's country- side has a great deal of (select from a list of words)

After completing this study, which included determining an appropriate IRT model, calibrating the items, and constructing the CAT algorithm, the ISIP Early Reading assessment went into full production starting in the 2008-2009 school year.

Chapter 3: Assessing the Technical Adequacy for Pre-Kindergarten

Data from ISIP Early Reading have been shown to be valid and reliable for students in Kindergarten through Grade 3 (Istation, 2009). Although the initial set of items was targeted for students in Kindergarten through Grade 3, the items were developed for a wide range of abilities, including older students performing below grade level and younger students such as those in Pre-Kindergarten (Pre-K). To establish validity evidence for the younger population, data were collected during the 2009-2010 school year from eleven Pre-K classes at five elementary schools (A-E) in a large North Texas school district, which was different from what the district used in the Item Response Theory (IRT) calibration study or in the previous validity study. Demographics of the study participants are found in Table 3-1.

Table 3-1: Student Demographics

	Pre-K	
Students	179	
By School		
A	27	(15.1%)
B	33	(18.4%)
C	37	(20.7%)
D	28	(15.6%)
E	54	(30.2%)
By Gender		
Male	91	(50.8%)
Female	88	(49.2%)
By Race/Ethnicity		
African American	35	(19.6%)
Asian	26	(14.5%)
Hispanic	35	(19.6%)
Other	4	(2.2%)
Pacific Islander	1	(0.6%)
White	78	(43.6%)
Other		
Qualifying for Free/Reduced Lunch	140	(78.2%)
Receiving ESL Services	14	(7.8%)
In a Bilingual Classroom	2	(1.1%)
English Language Leaner (ELL)	17	(9.5%)

	Pre-K	
Having a disability	2	(1.1%)
Receiving Special Ed Services	2	(1.1%)

NOTE: Percentages may not add up to 100% for a given category, due to rounding.

The schools included in the study used ISIP™,Istation's Indicators of Progress, throughout the 2009-2010 school year. At the beginning of each month, ISIP assessments were automatically administered to students during regularly scheduled computer lab time. Research assistants from the Institute for Evidence-Based Education at Southern Methodist University (SMU) assisted teachers in proctoring ISIP. In addition to ISIP, SMU school coordinators administered external measures to participating students in each school over the course of a week in November. Prior to administering any external measures, the SMU research assistants underwent training on each instrument to increase inter-rater reliability. A four-group Latin square design was utilized to reduce ordering effects. The external measures were selected based on the reading skills being measured, as well as its suitability for Pre-K students, as indicated in Table 3-2.

Table 3-2. Assessments Administered by Skill

Assessment	Letter Knowledge	Vocabulary	Phonemic Awareness	Comprehensive Ability
ISIP Early Reading	Sep–Dec	Sep–Dec	Nov–Dec	Sep–Dec
ELSA	Nov		Nov	
Letter Names	Nov			
Letter Sounds	Nov			
PPVT-4		Nov		
TOPEL	Nov	Nov	Nov	Nov

The ISIP Early Reading assessment measures abilities in the domains of phonemic awareness, alphabetic knowledge, fluency with text, vocabulary, and comprehension. However, only the subtests Letter Knowledge (through alphabet letter recognition and letter-sound correspondence items), Vocabulary (through oral-picture correspondence items), and Phonemic Awareness (through initial sound and blending items) are appropriate for emergent readers enrolled in Pre-K. At the end of each session, responses from all subtests are combined, and a comprehensive reading ability measure, called Overall Reading, is estimated using IRT.

Regarding the external measures used in the current study, the Early Literacy Skills Assessment (ELSA; DeBruin-Parecki, 2005) is unique in that the assessment is presented to students in the form of a children's storybook. ELSA measures Comprehension (through prediction, retelling, and connection to real-life items), Phonological Awareness (through rhyming, segmentation, and phonemic awareness items), Alphabetic Principle (through sense of word, alphabet letter recognition, and letter-sound correspondence items), and Concepts about Print (through orientation, story beginning, direction of text, and book part items). ELSA is not norm-referenced. Instead, ELSA identifies children in one of three developmental levels for each subtest: Level 1, Early Emergent; Level 2, Emergent; and Level 3, Competent Emergent. Letter Names and

Letter Sounds measure a student's ability to recognize each of the 26 letters, randomly presented, by name and by sound. The Peabody Picture Vocabulary Test, Fourth Edition (PPVT-4; Dunn and Dunn, 2007) was designed to measure the oral vocabulary of children and adults. The Test of Preschool Early Literacy (TOPEL; Lonigan, Wagner, Torgesen, and Rashotte, 2007) was designed to identify students in Pre-Kwho might be at risk for literacy problems that affect reading and writing. TOPEL consists of four subtests: Print Knowledge (through written language conventions and alphabetic knowledge items), Definitional Vocabulary (through oral vocabulary and word meaning items), Phonological Awareness (through elision and blending items), and a composite score known as the Early Literacy Index. Both PPVT-4 and TOPEL are norm-referenced tests.

Reliability Evidence

Cronbach's (1951) coefficient alpha is often used as an indicator of reliability across test items within a testing instance. However, alpha assumes that all students in the testing instance respond to a common set of items. Due to its very nature, a CAT-based assessment such as ISIP Early Reading will present students with a custom set of items based on initial estimates of ability and response patterns. The IRT analogue to classical internal consistency is marginal reliability (Bock and Mislevy, 1982), and it can be used with Cronbach's alpha to directly compare the internal consistencies of classical test data to IRT-based test data. ISIP Early Reading has stopping criteria based on minimizing the standard error of the ability estimate. Therefore, the lower limit of the marginal reliability of the data for any testing instance of ISIP will always be approximately 0.90.

To establish test-retest reliability evidence, Pearson product moment correlation coefficients between ISIP Early Reading administrations were computed. Results for ISIP Letter Knowledge, Vocabulary, and Overall Reading ability range from 0.532 to 0.735 across four months of testing sessions, September to December, as indicated in Tables 3-3 through 3-5. Students had to demonstrate minimal ability before being presented the ISIP Phonemic Awareness subtest, unlike the ISIP Letter Knowledge and Vocabulary subtests, both of which all students were given every month. In November only four students met the criteria, and in December only 23 students met the criteria. Therefore, there was insufficient power to perform statistical analysis for Phonemic Awareness reliability.

Table 3-3: ISIP Early Reading Letter Knowledge Test-Retest Reliability[a] between Testing Sessions

	Sep	Oct	Nov	Dec
Sep	---			
Oct	0.632** (171)	---		
Nov	0.650** (165)	0.699** (172)	---	
Dec	0.538** (163)	0.532** (170)	0.735** (167)	---

[a]Pearson product moment correlations (*r*).

**Statistically significant (H_0: r=0) at p<.01.

NOTE: Sessions occurred at the start of the month indicated. *N* for each correlation is within parentheses.

Table 3-4: ISIP Early Reading Vocabulary Test-Retest Reliability[a] between Testing Sessions

	Sep	Oct	Nov	Dec
Sep	---			
Oct	0.683** (171)	---		
Nov	0.577** (168)	0.658** (175)	---	
Dec	0.571** (169)	0.691** (176)	0.644** (173)	---

[a]Pearson product moment correlations (*r*).

**Statistically significant (H_0: r=0) at p<.01.

NOTE: Sessions occurred at the start of the month indicated. *N* for each correlation is within parentheses.

Table 3-5: ISIP Early Reading Overall Reading Test-Retest Reliability[a] between Testing Sessions

	Sep	Oct	Nov	Dec
Sep	---			
Oct	0.687** (171)	---		
Nov	0.706** (168)	0.701** (175)	---	
Dec	0.669** (169)	0.652** (176)	0.707** (173)	---

[a]Pearson product moment correlations (*r*).

**Statistically significant (H_0: r=0) at p<.01.

NOTE: Sessions occurred at the start of the month indicated. *N* for each correlation is within parentheses.

Validity Evidence

Content validity was established through a series of steps to substantiate the test development process. First, early reading content experts Patricia Mathes and Joe Torgesen created ISIP Early Reading assessment items in key developmental areas, as suggested by the National Reading Panel (National Institute of Child Health and Human Development, 2000). Next, the items underwent review by a panel of reading specialists. The items were piloted and then operationally used in a previous version of ISIP and revised as necessary. For ISIP Early Reading, the items were calibrated under a 2PL-IRT model. Finally, item parameters were examined, and those items with unacceptable fit statistics in regard to the subtest they measured were removed from the pool. Based on the combined processes used to establish content validity, the items in the operational pool, grouped by subtest, are believed to be accurate representations of the domains they intend to measure.

Concurrent validity evidence was established by computing Pearson product moment correlation coefficients between ISIP Early Reading subtests and appropriate external measures, as illustrated in Table 3-6. Because students had to demonstrate minimal ability before being presented the ISIP Phonemic

Awareness subtest, only four students met the criteria in November. Therefore, December ISIP Phonemic Awareness scores were used for validity analysis.

Table 3-6: Correlations[a] between External Measures and ISIP Early Reading Scores

ISIP Subtest		
	External Measure	r (N)
ISIP Letter Knowledge (November)		
	ELSA Alphabetic Principle Level	0.747** (172)
	ELSA Upper Case Subtest Score	0.726** (172)
	ELSA Lower Case Subtest Score	0.692** (172)
	ELSA Letter Sounds Subtest Score	0.636** (172)
	Letter Name Score	0.727** (172)
	Letter Sound Score	0.669** (172)
	TOPEL Print Knowledge Std Score	0.735** (170)
ISIP Vocabulary (November)		
	PPVT-4 Std Score	0.625** (173)
	TOPEL Definitional Vocabulary Std Score	0.520** (173)
ISIP Phonemic Awareness (December)		
	ELSA Phonological Awareness Total Score	0.549** (23)
	ELSA Rhyming Subtest Score	0.485* (23)
	ELSA Phonemic Awareness Subtest Score	0.620** (23)
	TOPEL Phonological Awareness Std Score	0.242 (23)
ISIP Overall Reading (November)		
	TOPEL Total Std Score	0.677** (173)
	TOPEL Early Literacy Index	0.676** (173)

[a]Pearson product moment correlations (*r*).

*Statistically significant (H_0: r=0) at p<.05. **Statistically significant (H_0: r=0) at p<.01.

Note. Sessions occurred at the start of the month indicated. *N* for each correlation is within parentheses.

Discussion

Regarding measures of reliability in the current study for Pre-K students, ISIP Early Reading produced stable scores over time, even between testing instances four months apart (see Tables 3-3 – 3-5). These test-retest reliability results could stem from a number of converging reasons. First, the exit criteria of the adaptive algorithm used in ISIP produces consistently strong levels of internal consistency, at approximately 0.90, both in the subtest ability scores and in the overall reading ability scores. Second, the authors, reading experts Patricia Mathes and Joe Torgesen, took great care in constructing the ISIP Early Reading item pool, basing the item types and content on contemporary findings in early reading research.

Furthermore, the ISIP Early Reading items have been operational for several years in previous versions of the program. Inconsistent items have been culled over time, resulting in a very stable item pool. Finally, ISIP Early Reading is an engaging and adaptive computer-based assessment program. Items are presented to students at their ability level and using high-quality computer animation. Students feel like they are "playing a game" rather than "taking another test," which probably results in less off-task behavior during assessments, producing more consistent results.

Evidence of concurrent validity can be found in the numerous strong, positive relationships to external measures of reading constructs. Cohen (1988) suggested that correlations around 0.3 could be considered moderate and those around 0.5 could be considered large. Hopkins (2010) expanded the upper end of Cohen's scale to include correlations around 0.7 as very large and those around 0.9 as nearly perfect. Given those criteria, the data from the current study show mostly large to very large criterion validity with scores from well-known, norm-referenced measures such as TOPEL and PPVT-4, as well as the authentic assessment, ELSA.

Specifically for letter knowledge, scores from the ISIP Letter Knowledge (LK) subtest showed strong, positive correlations to scores from comparable ELSA subtests, such as the Upper Case (r = 0.726), Lower Case (r = 0.692), and Letter Sounds (r = 0.636) subtests. In addition, ISIP LK scores correlated very well with Letter Names (r = 0.727) and Letter Sounds (r = 0.669), as well as TOPEL Print Knowledge (r = 0.735). These results suggest that the ISIP Letter Knowledge subtest measures the same construct as other early reading assessments.

Regarding vocabulary, PPVT-4 is most similar to the item format used in ISIP Vocabulary for students with early-emergent reading abilities, namely oral-picture correspondence. Therefore, it is not surprising that the correlation between the two sets of scores was large (r = 0.625). TOPEL Definitional Vocabulary (DV) also uses the oral-picture correspondence item format, but it adds a task in which participants state the meaning of the target word. Appropriately, the correlation between ISIP Vocabulary and TOPEL DV scores (r = 0.520) was somewhat less than that between ISIP and PPVT-4 scores, but it is still considered large.

Participants had to demonstrate repeated minimal ability in ISIP Early Reading to be offered the ISIP Phonemic Awareness (PA) subtest. Because students first took ISIP in September, the first opportunity to take ISIP PA as a Pre-K student was in November, when four students met the criteria. With insufficient power to compute correlations to external measures, it was decided that ISIP PA scores from December (N = 23) would be used for validity analyses, even though the collection of external measures data occurred in November. Both ELSA and TOPEL assess the broader concept of phonological awareness, including onset, rhyme, and segmentation, whereas ISIP PA assesses phonemic awareness concepts such as initial sound and phoneme blending. The correlation between ISIP PA and ELSA Phonemic Awareness subtest scores (r = 0.620) was large. However, even the phonological concept of rhyming (as measured by the ELSA Rhyming subtest) correlated well with ISIP PA scores (r = 0.485). The overall correlation between ELSA Phonological Awareness and ISIP Phonemic Awareness scores was large (r = 0.549). ISIP PA scores did not show any meaningful correlation to TOPEL Phonological Awareness standard scores (r = 0.242). However, the correlation between TOPEL Phonological Awareness standard scores and ELSA

Phonological Awareness total scores was equally insignificant (r = 0.278). This suggests that the ISIP Phonemic Awareness subtest and the ELSA phonological/phonemic subtests were measuring the same construct, but this construct was very different from the construct measured by the TOPEL Phonological Awareness subtest.

Finally, ISIP Early Reading computes a comprehensive measure of reading ability, called Overall Reading, through IRT modeling that utilizes the response pattern from all subtests in a testing session. Scores from ISIP Overall Reading correlated highly with the total standard scores from the TOPEL (r = 0.677) and with the TOPEL Early Literacy Index (r = 0.676), which is a seven-level interpretation of performance, ranging from Very Poor to Very Superior.

Taken together, the evidence supports the claim that ISIP Early Reading produces reliable and valid data for measuring key domains of emerging reading, such as letter knowledge, vocabulary, phonemic awareness, and comprehensive reading ability for students in Pre-Kindergarten.

Chapter 4: Reliability and Validity of ISIP ER for Kindergarten through 3rd Grade

The primary objective of this study is to establish the technical adequacy of the Computer Adaptive Testing (CAT)-based ISIP Early Reading assessment for students in Kindergarten through 3rd Grade. This consisted of conducting test-retest reliability and concurrent and predictive validity work. We compared ISIP Early Reading scores to scores from norm-referenced measures with good psychometric properties of similar constructs.

To establish reliability and validity evidence, data were collected during the 2008-2009 school year at five elementary schools (A-E) from a large north Texas independent school district, which was different from the district used in the Item Response Theory (IRT) calibration study. Demographics of the study participants are found in Table 4-1.

Table 4-1: Student Demographics

	Grade Level					
	K	1	2	3	K-3	
Students	122	103	95	96	416	
By School						
A	20	16	15	19	70	(16.8%)
B	21	15	18	18	72	(17.3%)
C	43	37	36	16	132	(31.7%)
D	17	15	11	12	55	(13.2%)
E	21	20	15	31	87	(20.9%)
By Gender						
Male	68	55	52	40	215	(51.7%)
Female	54	48	43	56	201	(48.3%)
By Ethnicity						
African American	21	28	17	10	76	(18.3%)
Caucasian	48	31	32	18	129	(31.0%)
Hispanic	40	38	40	65	183	(44.0%)
Asian	13	6	4	3	26	(6.3%)
Other	0	0	2	0	2	(0.5%)
Qualifying for Free/Reduced Lunch	63	52	44	73	232	(55.8%)
Qualifying for ESL Services	20	15	13	27	75	(18.0%)

	Grade Level					
	K	1	2	3	K-3	
Receiving ESL Services	17	15	10	25	67	(16.1%)
In a Bilingual Classroom	0	0	0	32	32	(7.7%)
Receiving Special Ed Services	1	5	6	7	19	(4.6%)

NOTE: Percentages may not add up to 100% for a given category due to rounding.

Research Design

A seven-group Latin square design was utilized to reduce ordering effect. Students were given assessments for reading skills appropriate for their age as indicated in Tables 4-2 and 4-3.

Table 4-2: CPM and Other Assessments Administered by Grade

	ISIP Early Reading							DIBELS					
Grade Level	PA	LK	AD	SPL	TF	CMP	VOC	PSF	NWF	ORF	TPRI[a]	ITBS[a]	TAKS[a]
K	X	X	X				X	X	X		X		
1	X		X	X	X	X	X	X	X	X		X	
2			X	X	X	X	X		X	X		X	
3				X	X	X	X			X			X

[a]**Tests administered by the district.**

Table 4-3: External Measures Administered by Grade

	External Measures							
Grade Level	CTOPP	LN/LS	WLPB-R	TOWRE	WIAT-II	WJ-III	GORT-4	PPVT-III
K	X	X	X	X				X
1	X		X	X	X	X	X	X
2			X	X	X	X	X	X
3			X		X	X	X	X

Seven thirty-minute testing sessions occurred every two weeks between October and February (Oct 20, Nov 3, Nov 17, Dec 8, Jan 12, Jan 26, and Feb 9). For each session, students were escorted to the school's computer lab in convenience groupings by trained data collectors from Southern Methodist University (SMU), for sessions on the CAT-based ISIP Early Reading program. On average, six items were needed per subtest to establish an ability estimate with a standard error below the threshold, resulting in 13-18 minute ISIP testing sessions, depending on the number of skills assessed. The remaining time in each session was spent administering external measures.

The key reading domains measured by ISIP Early Reading were Phonemic Awareness (PA), Letter Knowledge (LK), Alphabetic Decoding (AD), Spelling (SPL), Text Fluency (TF), Comprehension (CMP), and Vocabulary (VOC). All subtests, except Text Fluency, are CAT-based and are measured on a common scale. Text Fluency is a maze task and has a proprietary scoring mechanism.

The standard CPM measure against which our test was compared was the *Dynamic Indicators of Basic Early Literacy Skills* (DIBELS: Kaminski & Good, 1996; Good & Kaminski, 1996; 2002). *Phoneme Segmentation Fluency* (PSF) assesses a student's ability to fluently segment three and four phoneme words into their individual phonemes. The reliability coefficient is 0.88 for a single probe and 0.96 for the mean of 5 probes. Concurrent and predictive validity with a variety of reading tests ranges from 0.45 to 0.68. *Nonsense Word Fluency* (NWF) tests a child's alphabetic decoding ability. The reliability coefficient is 0.92 for a single probe and 0.98 for the mean of 5 probes. Concurrent and predictive validity with a variety of reading tests ranges from 0.59 to 0.82. *Oral Reading Fluency* (ORF) requires the student to orally read a passage geared to the student's grade level; predictive validity of ORF administered in January during Kindergarten with oral reading fluency administered in spring during First Grade is 0.45; predictive validity with the *Woodcock-Johnson Psycho-Educational Battery* Total Reading Cluster score is 0.36.

The *Texas Primary Reading Inventory* (TPRI; Texas Education Agency, 1998) was administered to all Kindergarten students by the district three times during the school year: beginning of the year (BOY), middle of the year (MOY), and end of the year (EOY). The *Iowa Tests of Basic Skills* (ITBS; Hoover, Dunbar, & Frisbie, 2007) was administered by the district in October to all students in Grades 1 and 2. The *Texas Assessment of Knowledge and Skills* (TAKS; Texas Education Agency, 2003) was administered by the district in October to all students in Grade 3. These data for students in the current study were provided by the district at the end of the school year.

Furthermore, one or more additional external measures were administered during each session. These additional assessments include well-known instruments in Phonemic Awareness: *Comprehensive Test of Phonological Processes* (CTOPP; Wagner, Torgesen, & Rashotte, 1999); Letter Knowledge: *Woodcock Language Proficiency Battery-Revised* (WLPB-R; Woodcock, 1991); Alphabetic Decoding: *Test of Word Reading Efficiency* (TOWRE; Torgesen, Wagner, & Rashotte, 1999), WLPB-R, and *Wechsler Individual Achievement Test* (WIAT-II; Wechsler, 2005); Spelling: *Woodcock-Johnson III Tests of Achievement* (WJ-III ACH; Woodcock, McGrew, & Mather, 2001) and WIAT-II; Vocabulary: *Peabody Picture Vocabulary Test* (PPVT-III; Dunn & Dunn, 1997) and WLPB-R; and Comprehension: *Gray Oral Reading Tests* (GORT-4; Wiedeholt & Bryant, 2001), WLPB-R, and WIAT-II.

The WLPB-R is a well-standardized instrument whose normative sample was concordant with 1980 US Census statistics, which consisted of 6,359 subjects (3,245 in K to 12), and was the same as that of the Woodcock-Johnson Psychoeducational Battery – Revised (Woodcock & Johnson, 1989). Median coefficient alphas range from 0.81 to 0.92 across all age ranges (and from 0.77 to 0.96 at ages 6 to 9) for the subtests utilized; test-retest measures for selected subtests in a sample of 504 ranged from 0.75 to 0.95. In addition, content, concurrent, and construct validity data is also available in the WLPB-R manual (Woodcock, 1991).

The CTOPP has nine subtests measuring phonological awareness (PA), rapid naming (RN), and phonological memory (PM). The normative base consisted of 1,656 individuals from ages 5 to 24, similar to the 1997 US Census statistics. Coefficient alphas for all three composites in the entire normative sample ranged from 0.83 to 0.95, and 0.83 to 0.92 in the age range of this sample; test-retest estimates in a small sample (n = 32) of children aged 5 to 7 ranged from 0.70 to 0.92 for the 3 composites. In addition, content, concurrent, predictive, and construct validity data is provided in the CTOPP manual (Wagner et al., 1999).

PPVT-III is a measure of expressive vocabulary. Reliability coefficients range from Alpha of 0.92 to 0.98. In addition, content, concurrent, predictive, and construct validity data is provided in the PPVT-4 manual (Dunn & Dunn, 2006).

The TOWRE is a measure of the accuracy and fluency of the word reading process (Torgesen, et al, 1999). The phonemic decoding efficiency subtest measures the number of nonwords students can pronounce in 45 seconds from a list that gradually increases in difficulty. The sight word (real-word) efficiency subtest has a similar structure, but the list is composed of high- frequency words. Reliability coefficients are 0.95 and 0.96 respectively. Content, concurrent, and construct validity data is also available in the TOWRE manual (Torgesen, et al., 1999).

The WIAT-II was standardized using a total sample of 5,586 individuals, with 2 standardization samples drawn for Pre-K through 12th grade (ages 4-19) and for the college-adult population. Both standardization samples were stratified based on the data from the 1998 U.S. Census Bureau, including grade, age, sex, race-ethnicity, geographic region, and parent education level. Age-based (4-19) average reliability coefficients on the spelling and reading comprehension subtests were .94 and .95, while grade-based (K-12) reliability coefficients were .93 and .93, respectively. In addition, content, concurrent, predictive, and construct validity data is provided in the WIAT-II manual (Wechsler, 2005).

The WJ-III ACH is a comprehensive instrument whose normative sample consisted of 8,818 subjects ranging in age from 24 months to 90 years (4,783 in K to 12) drawn from over 100 geographically diverse U.S. communities and selected to be representative of the U.S. population. Median reliability coefficient alphas for the standard battery for tests 1-12, all age groups, ranged from .81 to .94. Coefficient alphas for the spelling subtest of children aged 6-9, ranged from .89 to .92. The median coefficient alpha across all ages for the spelling subtest was .90. Test-retest reliabilities for the spelling subtest of children aged 4-7 (n=106) and 8-10 (n=145) were .91 and .88, respectively, with the median retest reliability of children aged 4 -17 (n=449) reported to be .95. In addition, content, concurrent, predictive, and construct validity data is provided in the WJ-III manual (Woodcock, et al, 2001).

The GORT-4 measures oral reading rate, accuracy, fluency, and comprehension. The normative sample consisted of 1,677 students ranging in age from 6 to 18 years old and was stratified to correspond with demographic characteristics reported by the U.S. Census Bureau in 1997. The coefficient alphas related to content sampling, test-retest, and scorer differences for the Form A comprehension subtest utilized are .97, .86., and .96, respectively. In addition, content, concurrent, predictive, and construct validity data is provided in the GORT-4 manual (Wiederholt & Bryant, 2001).

Reliability

Internal Consistency

Cronbach's (1951) coefficient alpha is typically used as an indicator of reliability across test items within a testing instance. However, Cronboch's Alpha is not appropriate for any IRT-based measure because alpha assumes that all students in the testing instance respond to a common set of items. Due to its very nature, students taking a CAT-based assessment, such as ISIP Early Reading, will receive a custom set of items based on their initial estimates of ability and response patterns. Thus, students do not respond to a common set of items.

The IRT analogue to classical internal consistency is marginal reliability (Bock & Mislevy, 1982) and thus applied to ISIP Early Reading. Marginal reliability is a method of combining the variability in estimating abilities at different points on the ability scale into a single index. Like Cronbach's alpha, marginal reliability is a unitless measure bounded by 0 and 1, and it can be used with Cronbach's alpha to directly compare the internal consistencies of classical test data to IRT-based test data. ISIP Early Reading has a stopping criteria based on minimizing the standard error of the ability estimate. As such, the lower limit of the marginal reliability of the data for any testing instance of ISIP Early Reading will always be approximately 0.90.

Test-Retest Consistency

To establish test-retest reliability evidence, Pearson product moment correlation coefficients between ISIP Early Reading sessions were computed. Results for overall reading ability range from 0.927 to 0.970 (N = 416) across all seven sessions spanning from October to February. Table 4-4 shows the individual test-retest reliability results for overall reading ability with all grades combined.

Table 4-4: ISIP Early Reading Overall Reading Test-Retest Reliability[a] between Testing Sessions for All Grades Combined

	Oct 20	Nov 3	Nov 17	Dec 8	Jan 12	Jan 26	Feb 9
Oct 20	---						
Nov 3	0.970	---					
Nov 17	0.962	0.975	---				
Dec 8	0.947	0.962	0.969	---			
Jan 12	0.946	0.963	0.964	0.960	---		
Jan 26	0.936	0.956	0.962	0.960	0.963	---	
Feb 9	0.927	0.945	0.951	0.949	0.958	0.961	---

[a]Pearson product moment correlations (r).

NOTE. Sessions were two weeks in length and started on the date indicated.

Validity Evidence

Construct Validity

Much prior work done has been done to establish construct validity of our item pool. The decision to include certain types of items builds on the vast amount of work alluded to in prior sections, describing what types of activities and skills predict a child's later reading performance. Thus, in designing ISIP Early Reading, we included only reading domains shown to meaningfully predict reading performance. In order to determine how to assess each domain, we utilized our collective expertise. In particular, we built upon Dr. Torgesen's prior work in developing items for the *Comprehensive Test of Phonological Processing* (CTOPP; Wagner, Torgesen, & Rashotte, 1999), and the *Test of Word Reading Efficiency* (TOWRE: Torgesen, Wagner, & Rashotte, 1999.) Of course, given that ISIP is computer-administered, we knew that many types of items could not be delivered in the same manner. Thus, we tested administration of each item, first in a graphic mock-up form, then as a computer delivered item. This procedure allowed us to "tinker" with item art and directions, until we were satisfied that there were no unintended confusions presented by the art, that the art was culture free, and that each item's correct response and distracters were operating as intended. The essence of this original art has been preserved in ISIP Early Reading. Items that were confusing to children were removed from the item pool. The result is a pool of items conforming to a current understanding of how reading develops and how to measure it.

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

Concurrent Validity

Concurrent validity evidence was established by computing Pearson product moment correlation coefficients between ISIP Early Reading subtests and appropriate external measures. Table 4-5 shows results by grade level. During each of the seven testing sessions, both ISIP Early Reading and DIBELS were administered to the students in the study. Pearson correlations between DIBELS and ISIP Early Reading are shown in Table 4-6. Prior to testing, the SMU testers were trained on administering DIBELS. Inter-rater reliability was ensured during training so that no more than a two point difference in scoring occurred between testers.

The *Texas Primary Reading Inventory* (TPRI; Texas Education Agency, 1998) was administered to all Kindergarten students by the district three times during the school year: beginning of the year (BOY), middle of the year (MOY), and end of the year (EOY). Data for students in the current study were provided by the district at the end of the school year. It is unknown when these testing administrations occurred, so

data from the most appropriate ISIP Early Reading testing sessions were used in the comparisons. The study concluded in February, so correlations for EOY (presumably administered in May) were not performed. Pearson correlations between TPRI subtests and ISIP Early Reading subtests for BOY and MOY are found in Table 4-7. The training and inter-rater reliability of the district testers is unknown.

Table 4-5: Correlations between External Measures and ISIP Early Reading Subtest Scores

ISIP Early Reading Subtest	External Measure		Grade Level				
			K	1	2	3	K-3
Phonemic Awareness (PA)	CTOPP Blending Words	*r*	**.688**	**.431**			**.702**
		N	120	100			220
	CTOPP Blending Non Words	*r*	**.676**	**.336**			**.650**
		N	120	100			220
	CTOPP Segmenting Words	*r*	**.644**	**.344**			**.620**
		N	122	101			223
	CTOPP Sound Matching	*r*	**.624**	**.474**			**.662**
		N	122	101			223
Letter Knowledge (LK)	Letter Names	*r*	**.593**				**.593**
		N	121				121
	Letter Sounds	*r*	**.693**				**.693**
		N	121				121
	WLPB-R Letter Word ID	*r*	**.711**				**.711**
		N	120				120
Alphabetic Decoding (AD)	TOWRE Phonemic Decoding	*r*	**.582**	**.679**	**.539**		**.838**
		N	122	103	93		313
	TOWRE Sight Word Efficiency	*r*	**.583**	**.626**	**.586**		**.811**
		N	120	100	93		313
	WLPB-R Word Attack	*r*	**.535**	**.701**	**.702**		**.830**
		N	122	102	94		316
	WIAT-II Target Words	*r*		**.624**	**.507**		**.589**
		N		101	92		193
Spelling (SPL)	WJ-III Spelling	*r*		**.800**	**.823**	**.798**	**.890**
		N		103	94	96	293
	WIAT-II Spelling	*r*		**.726**	**.774**	**.788**	**.875**
		N		101	91	96	288
Connected Text Fluency (TF)	DIBELS ORF[a]	*r*		**.741**	**.667**	**.627**	**.766**
		N		103	92	94	289
Comprehension	GORT-4 Comprehension	*r*		**.456**	**.354**	**.473**	**.621**

ISIP Early Reading Subtest	External Measure		Grade Level				
			K	1	2	3	K-3
(CMP)		*N*		102	95	94	291
	WLPB-R Comprehension	*r*		**.707**	**.597**	**.569**	**.794**
		N		102	92	93	287
	WIAT-II Comprehension	*r*		**.630**	**.554**	**.596**	**.682**
Vocabulary (VOC)	PPVT-III	*r*	**.687**	**.696**	**.582**	**.785**	**.814**
		N	121	101	94	95	411
	WLPB-R Vocabulary	*r*	**.368**	**.656**	**.702**	**.716**	**.836**
		N	121	103	94	96	414

[a]Feb 9 session data used for correlations.

NOTE: Empty cells indicate no students were administered the instrument for the grade level.

Table 4-6: Correlations between DIBELS and ISIP Early Reading Subtest Scores for Grades K-3

		PSF & PA					NWF & AD					ORF & TF				
		Grade Level					Grade Level					Grade Level				
		K	1	2	3	K-3	K	1	2	3	K-3	K	1	2	3	K-3
Oct	r	**.65**	**.48**			**.71**	**.45**	**.43**	**.38**		**.72**		**.66**	**.70**	**.81**	**.83**
	N	98	92			190	96	94	84		274		87	81	73	241
Nov[1]	r	**.61**	**.39**			**.68**	**.43**	**.52**	**.50**		**.79**		**.59**	**.71**	**.71**	**.79**
	N	121	103			224	121	103	93		317		100	93	91	284
Nov[2]	r	**.71**	**.37**			**.71**	**.58**	**.57**	**.52**		**.81**		**.66**	**.74**	**.73**	**.83**
	N	121	102			223	121	102	93		316		102	93	96	291
Dec	r	**.65**	**.41**			**.65**	**.57**	**.64**	**.61**		**.82**		**.64**	**.68**	**.62**	**.75**
	N	121	102			223	121	102	92		315		101	93	94	288
Jan[1]	r	**.62**	**.24**			**.56**	**.61**	**.49**	**.65**		**.80**		**.59**	**.71**	**.60**	**.75**
	N	120	102			222	120	102	86		308		102	91	95	288
Jan[2]	r	**.53**	**.17**			**.48**	**.55**	**.59**	**.51**		**.78**		**.66**	**.71**	**.65**	**.78**
	N	121	102			223	121	102	91		314		102	91	94	287
Feb	r	**.50**	**.25**			**.52**	**.60**	**.54**	**.44**		**.76**		**.74**	**.67**	**.63**	**.77**
	N	122	102			224	122	103	92		317		103	92	94	289

NOTE: Empty cells indicate no students were administered the instrument for the grade level.

Table 4-7: Correlationsa between TPRI Subtest Scores and ISIP Early Reading Subtest Scores for Kindergarten

		ISIP Early Reading Phonemic Awareness					ISIP EARLY READING Letter Knowledge	
		Rhy[b]	*BWP*[c]	*BP*[d]	*DIS*[e]	*DFS*[f]	*LN*[g]	*LtSL*[h]
BOY[i]	*r*	**.48**	**.56**	**.56**	**.48**	**.40**	**.73**	**.56**
	N	109	97	91	88	88	109	97
MOY[j]	*r*	**.33**	**.60**	**.60**	**.56**	**.56**	**.63**	**.55**
	N	109	101	98	97	88	109	106

[a]Pearson product moment correlations (r). TPRI subtest = [b]Rhyming. [c]Blending Word Parts. [d]Blending Phonemes. [e]Deleting Initial Sounds. [f]Deleting Final Sounds. [g]Letter Name Identification. [h]Letter to Sound Linking. [i]BOY = ISIP Early Reading Nov 17 session data used for correlations. [j]MOY =ISIP Early Reading Jan 12 session data used for correlations.

NOTE: TPRI administered by the district. It is unknown when in the school year TPRI was administered, by whom, or under what conditions.

The *Iowa Tests of Basic Skills* (ITBS; Hoover, Dunbar, & Frisbie, 2007) was administered by the district in October to all students in Grades 1 and 2. Data for students in the current study were provided by the district at the end of the school year. Pearson correlations between ITBS Reading and ISIP Early Reading overall reading ability scores are shown in Table 4-8.

Table 4-8: Correlations[a] between ITBS Reading Scale Scores and ISIP Early Reading Overall Reading Scores for Grades 1 and 2

Testing		Grade Level		
Session		1	2	1-2
Oct 20	*r*	**.807**	**.845**	**.895**
	N	62	75	137
Nov 3	*r*	**.808**	**.821**	**.884**
	N	65	78	143
Nov 17	*r*	**.793**	**.839**	**.888**
	N	65	78	143
Dec 8	*r*	**.806**	**.741**	**.845**
	N	65	78	143
Jan 12	*r*	**.748**	**.837**	**.874**
	N	64	78	142
Jan 26	*r*	**.725**	**.806**	**.854**
	N	65	78	143
Feb 9	*r*	**.699**	**.768**	**.829**
	N	65	77	142

[a]Pearson product moment correlations (r).

NOTE: ITBS administered by the district in October.

To establish predictive validity evidence, Pearson correlations between ISIP Early Reading overall reading ability and the state-mandated *Texas Assessment of Knowledge and Skills* (TAKS; Texas Education Agency, 2003) were computed for Grade 3. Results are found in Table 4-9. TAKS was administered by the district in March. Furthermore, ROC analysis was conducted to determine the power to which ISIP Early

Reading Overall Reading scores from January predicted a passing status on TAKS Reading in March (Macmillan & Creelman, 2005). Table 4-10 shows the contingency table for the data, resulting in an instrument sensitivity of 85.7%, specificity of 95.7%, positive prediction power (precision) of 66.7%, and a false positive rate of 4.3%. The subsequent ROC graph, with an area under the curve (Az) of 89.8%, is displayed in Figure 4-A.

Table 4-9. Correlations[a] between TAKS Reading Scale Scores and ISIP Scores plus DIBELS ORF Scores for Grade 3

Testing		ISIP				DIBELS
Session		Fluency with Text	Vocabulary	Comprehension	Overall Reading	ORF
Oct 20	*r*	**.641**	**.697**	**.678**	**.740**	**.630**
	N	63	64	64	64	60
Nov 3	*r*	**.665**	**.660**	**.598**	**.741**	**.551**
	N	75	74	74	74	75
Nov 17	*r*	**.677**	**.652**	**.625**	**.698**	**.598**
	N	77	77	77	77	77
Dec 8	*r*	**.617**	**.652**	**.586**	**.695**	**.450**
	N	77	77	77	77	76
Jan 12	*r*	**.649**	**.645**	**.580**	**.698**	**.582**
	N	76	76	76	76	77
Jan 26	*r*	**.492**	**.687**	**.648**	**.741**	**.555**
	N	75	74	74	74	75
Feb 9	*r*	**.667**	**.637**	**.607**	**.710**	**.533**
	N	76	77	77	77	76

[a]Pearson product moment correlations (r).

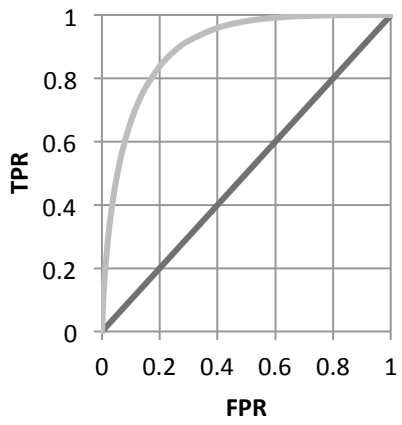
NOTE: TAKS administered by the district in March.

Table 4-10: Contingency Table for ISIP Early Reading Overall Reading Score in January Predicting TAKS Reading Passing Condition in March for Grade 3

		TAKS Reading		
		Not Passing	Passing	Total
ISIP Early Reading Overall Reading Score	< 227[a]	6	3	9
	>= 227	1	67	68
	Total	7	70	77

[a]The Overall Reading score of 227 is associated with the 20th percentile for students in Grade 3 taking ISIP Early Reading in January.

Figure 4-A. ROC Graph for ISIP Early Reading Overall Reading as a TAKS Reading Predictor for Grade 3



Discussion

Reliability and validity are two important qualities of measurement data. Reliability can be thought of as consistency, either consistency over items within a testing instance or over scores from multiple testing instances, whereas validity can be thought of as accuracy, either accuracy of the content of the items or of the constructs being measured. In this study, both qualities were examined using ISIP Early Reading data collected from Kindergarten through Grade 3 students in north Texas elementary schools during the 2008-2009 school year.

Regarding measures of reliability, the data from the current study suggest consistently high levels of internal consistency, both in the subtest ability scores as well in the overall reading ability scores. In addition, ISIP Early Reading produced extremely stable scores over time, even between testing instances five months apart. These outstanding results could stem from a number of converging reasons. First, the

authors, reading experts Drs. Patricia Mathes and Joe Torgesen, took great care in constructing the ISIP Early Reading item pool. They utilized the most up-to-date findings in early-reading research as a basis for the item types and content they produced for Istation. Furthermore, the ISIP Early Reading items have been operational for several years in previous versions of the program. Inconsistent items have been culled over time, resulting in a very stable item pool. Finally, ISIP Early Reading is an engaging and adaptive computer-based assessment program. Items are presented to students at their ability and using high quality computer animation. Students feel they are "playing a game" rather than "taking another test," which probably results in less off-task behavior during assessment, producing more consistent results.

Evidence of concurrent validity, can be found in the numerous strong, positive relationships to external measures of reading constructs. Cohen (1988) suggested correlations around 0.3 could be considered moderate and those around 0.5 could be considered large. Hopkins (2009) expanded the upper end of Cohen's scale to include correlations around 0.7 as very large, and those around 0.9 as nearly perfect. Given those criteria, the data from the current study show mostly large to very large criterion validity with scores from well-known external measures, such as CTOPP, GORT-4, PPVT-III, TOWRE, WJ-III ACH, WLPB-R, and WIAT-II, as well as with TPRI and ITBS. In addition, validity results show that ISIP Overall Reading is a stronger predictor than DIBELS ORF for TAKS Reading, using scores from 1 to 5 months prior to TAKS administration.

Taken together, the evidence supports the claim that ISIP Early Reading produces reliable and valid data for measuring key areas of reading development, such as phonemic awareness, alphabetic knowledge, vocabulary, and reading comprehension, as well as overall reading ability.

Chapter 5: Determining Norms

Norm-referenced tests are designed so that test administrators have a way of comparing the results of a given test-taker to the hypothetical "average" test taker to determine whether they meet expectations. In the case of the Computerized Adaptive Testing (CAT)-based ISIP Early Reading test, we are interested in comparing students to a national sample of students who have taken the ISIP Early Reading test. We are also interested in knowing what the expected growth of a given student is over time, and in administering our test regularly to students to determine how they are performing relative to this expected growth. By determining and publishing these norms, called Instructional Tier Goals, we enable teachers, parents, and students to know how their scores compare with a representative sample of children in their particular grade for the particular period (month) in which the test is administered. The norming samples were obtained as part of Istation's ongoing research in assessing reading ability. The samples were drawn from enrolled ISIP Early Reading users during the 2014-2015 school year in grades PreK - 3. The state distributions for the sample are found in Table 5-1.

Table 5-1: State Distributions & Demographics For ISIP Early Reading Norming Sample

	Grade				
	Pre-K	K	1st	2nd	3rd
	Frequency (%)	Frequency (%)	Frequency (%)	Frequency (%)	Frequency (%)
Gender					
Female	3,118 (26.1)	39,963 (29.2)	57,305 (31.8)	57,629 (32.9)	7,630 (4.8)
Male	3,233 (27.1)	42,231 (30.8)	61,367 (34.1)	61,388 (35.0)	8,442 (4.7)
Special Education					
No	5,294 (44.3)	63,805 (46.6)	77,061 (42.8)	77,431 (44.2)	8,870 (5.0)
Yes	224 (1.9)	3,719 (2.7)	5,507 (3.1)	6,097 (3.5)	1,485 (0.8)
State					
Alabama	2 (0.1)	1,628 (1.2)	1,898 (1.1)	1,382 (0.8)	1,183 (0.7)
Arizona	15 (0.1)	211 (0.2)	176 (0.1)	151 (0.1)	147 (0.1)
California	52 (0.4)	1,237 (0.9)	1,739 (1.0)	1,669 (1.0)	1,583 (0.9)
Colorado	-	204 (0.1)	360 (0.2)	252 (0.1)	265 (0.1)
District of Columbia	-	41 (0.1)	-		
Florida	111 (0.9)	10,238 (7.5)	14,971 (8.3)	8,310 (4.7)	6,559 (3.7)
Georgia	202 (1.7)	1,970 (1.4)	2,291 (1.3)	2,124 (1.2)	2,117 (1.2)
Illinois	107 (0.9)	365 (0.3)	378 (0.2)	410 (0.2)	482 (0.3)
Indiana	201 (1.7)	437 (0.3)	454 (0.3)	345 (0.2)	289 (0.2)
Iowa	-	95 (0.1)	134 (0.1)	108 (0.1)	23 (0.1)
Maine	-	10 (0.1)	41 (0.1)	34 (0.1)	7 (0.1)
Maryland	28 (0.2)	207 (0.2)	331 (0.2)	395 (0.2)	244 (0.1)

Montana	96 (0.8)	324 (0.2)	1,038 (0.6)	275 (0.2)	299 (0.2)
Massachusetts	-	19 (0.1)	31 (0.1)	7 (0.1)	2 (0.0)
North Carolina	12 (0.1)	630 (0.5)	938 (0.5)	1,157 (0.7)	1,222 (0.7)
North Dakota	29 (0.2)	165 (0.1)	200 (0.1)	178 (0.1)	132 (0.1)
New Jersey	60 (0.5)	346 (0.3)	734 (0.4)	806 (0.5)	672 (0.4)
New Mexico	-	16 (0.1)	37 (0.1)	51 (0.1)	73 (0.1)
New York	61 (0.5)	168 (0.1)	266 (0.1)	160 (0.1)	170 (0.1)
Ohio	-	23 (0.1)	17 (0.1)	51 (0.1)	42 (0.1)
Oregon	-	10 (0.1)	9 (0.1)	-	-
Pennsylvania	56 (0.5)	1,042 (0.8)	703 (0.4)	678 (0.4)	212 (0.1)
Rhode Island	-	37 (0.1)	56 (0.1)	1 (0.0)	2 (0.0)
South Carolina	340 (2.8)	1,455 (1.1)	1,692 (0.9)	1,395 (0.8)	1,114 (0.6)
South Dakota	-	33 (0.1)	59 (0.1)	48 (0.1)	34 (0.1)
Tennessee	43 (0.4)	7,566 (5.5)	7,536 (4.2)	7,252 (4.1)	7,016 (3.9)
Texas	10,285 (86.1)	104,169 (76.1)	138,403 (76.8)	141,829 (80.9)	149,911 (83.8)
Utah	-	388 (0.3)	851 (0.5)	878 (0.5)	494 (0.3)
Virginia	192 (1.6)	2,181 (1.6)	2,979 (1.7)	3,010 (1.7)	2,731 (1.5)
West Virginia	-	62 (0.1)	120 (0.1)	123 (0.1)	113 (0.1)

Sample

We last updated the ISIP Early Reading Instructional Tier Goals in August 2011. Since that time, there has been substantial growth in the number of students using the ISIP Early Reading assessment. Due to this growth in population, it was necessary to establish a new norming sample in order to derive updated expected growth and goals that represent the current population of students using ISIP Early Reading. Students completing three assessments in September (BOY), January (MOY), and May (EOY) during the 2014-2015 school year were sampled from the total population to establish the norming sample. The total population by grade (N) and the sample size (n) by grade are found in Table 5-2. In total, the ISIP Early Reading scores from 683,379 students were considered to establish norms. This sample used in establishing the Instructional Tier Goals for the ISIP Early Reading Overall ability score, as well as all subtests within ISIP Early Reading.

Table 5-2: ISIP Early Reading Population and Norm Sample

ISIP	Size	Pre-K	K	1st	2nd	3rd
ISIP_BOY	N	16,540	174,466	226,624	226,847	264,061
	n	11,951	136,930	180,168	175,352	178,978)
ISIP_MOY	N	42,448	247,507	279,285	274,543	327,372
	n	11,951	136,930	180,168	175,352	178,978
ISIP_EOY	N	46,011	244,386	273,065	266,384	280,635
	n	11,951	136,930	180,168	175,352	178,978

Computing Norms

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

All ISIP Early Reading scores of Overall Reading Ability, Alphabetic Decoding, Comprehension, Letter Knowledge, Phonemic Awareness, Spelling, Text Fluency, and Vocabulary were used to develop the updated Instructional Tier Goals. Table 5-3 shows which ISIP Early Reading subtests by grade level that have associated Instructional Tier Goals. Alphabetic Decoding goals are available for only Grade 1. Comprehension, Spelling, and Text Fluency goals are available for Grades 1–3. Letter Knowledge goals are available for Grades Pre-K – 1. Phonemic Awareness goals are available for Kindergarten and Grade 1, whereas Overall Reading ability and Vocabulary are available for Grades PreK-3.

Table 5-3: Availability of Instructional Tier Goals by ISIP Early Reading Subtests by Grade

Subtest	Pre-K	K	1st	2nd	3rd
Overall Reading Ability	√	√	√	√	√
Alphabetic Decoding (AD)	-	-	√	-	-
Comprehension (CMP)	-	-	√	√	√
Letter Knowledge (LK)	√	√	√	-	-
Phonemic Awareness (PA)	-	√	√	-	-
Spelling (SPL)	-	-	√	√	√
Text Fluency (TF)	-	-	√	√	√
Vocabulary (VOC)	√	√	√	√	√

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

Instructional Tier Goals

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

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

- Tier 1 are performing at grade level.
- Tier 2 are performing moderately below grade level and in need of intervention.
- Tier 3 are performing seriously below grade level and in need of intensive intervention.

References

Beck, I.L., McKeown, M.G., & Kucan, L. (2002). *Bringing words to life: Robust vocabulary instruction.* New York: The Guilford Press.

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

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

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

Degraff, A. & Torgesen, J. (2005) Monitoring growth in early reading skills: Validation of a computer adaptive test. Florida State University College of Arts and Sciences. Dissertation, spring 2005.

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

Dickinson, D., & Tabors, P. (2001). *Beginning literary with language*. Baltimore: Paul E. Brookes.

Ehri, L. (2000). Learning to read and learning to spell: Two sides of a coin. *Topics in Language Disorders*, 20(3), 19-49.

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.M., Foorman, B.R., Francis, D.J., & Schatschneider, C. (1997). Prevention of reading failure. *Insight*, 22-23.

Fletcher, J.M., Foorman, B.R., Boudousquie, A., Barnes, M., Schatschneider, C., & Francis, D.J. (2002). Assessment of reading and learning disabilities: A research-based, treatment-oriented approach. *Journal of School Psychology*, 40, 27-63.

Foorman, B. R., Anthony, J., Seals, L., & Mouzaki, A, (2002). Language development and emergent literacy in preschool. *Seminars in Pediatric Neurology*, 9, 172-183.

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

Foorman, B. R., & Torgesen, J. (2001). Critical elements of classroom and small-group instruction promote reading success in all children. *Learning Disabilities Research & Practice*, 16, 203-212.

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

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

Fuchs, L. S., Deno, S. L., & Mirkin, P. K. (1984). The effects of frequent curriculum-based measurement and evaluation on pedagogy, student achievement, and student awareness of learning. *American Educational Research Journal*, 21, 449-46.

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

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

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

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Ferguson, C. (1992). Effects of expert system consultation within curriculum-based measurement using a reading maze task. *Exceptional Children*, 58, 436-4.

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

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

Good, R. H., Shinn, M. R., & Bricker, D. (1991). *Early Intervention to prevent special education: Direct assessment of student progress on pre-reading skills*. Paper submitted to the Field-Initiated Research Projects (84.023C) Competition United States Department of Education Office of Special Education and Rehabilitative Services. Oregon.

Good, R. H., & Kaminski, R.A. (2002). DIBELS. *Oral reading fluency passages for first through third grade* (Technical Report No. 10). Eugene, OR: University of Oregon.

Good, R. H., & Kaminski, R.A. (1996). Assessment for instructional decisions: Toward a proactive/prevention model of decision making for early literacy skills. *Scholastic Psychology*, 11, 325-336.

Good, R. H., & Kaminski, R. A. (Eds.). (2002b). Dynamic Indicators of Basic Early Literacy Skills (6th ed.). Eugene, OR: Institute for the Development of Education Achievement.

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.

Jenkins, J. R., & Jewell, M. (1993). Examining the validity two measures for formative teaching: Reading aloud and maze. *Exceptional Children*, 59, 421- 432.

Jenkins, J. R., Pious, C.G., & Jewell, M. (1990). Special education and the regular education initiative: Basic assumptions. *Exceptional Children*, 56, 479-91.

Kaminski, R. A., & Good, R.H. (1996). Toward a technology for assessing basic early literacy skills. *School Psychology Review*, 25, 215-227.

Lonigan, C. J., Burgess, S.R., & Anthony, J. L. (2000). Development of emergent literacy and early reading skills in preschool children: Evidence from a latent variable longitudinal study. *Developmental Psychology*, 36, 596 - 613.

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

Mathes, P. G., Fuchs, D., & Roberts, P. H. (1998). The impact of curriculum-based measurement on transenvironmental programming. *Journal of Learning Disabilities*, 31(6), 615-624.

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.

O'Connor, R. E., & Jenkins, J. R. (1999). The prediction of reading disabilities in kindergarten and first grade. *Scientific Studies of Reading*, 3, 159-197.

Rayner, K., Foorman, B. R., Perfetti , C.A. , Pesetsky, D., & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest*, 2 (2), 31-74.

Scarborough, H. S. (1998). Early identification of children at risk for reading disabilities: Phonological awareness and some other promising predictors. In B. K. Shapiro, P. J. Accardo, & A. J. Capute (Eds.), *Specific reading disability: A view of the spectrum* (pp. 75-107). Timonium, MD: York Press.

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.

Shin, J., Deno, S. L., & Espin, C. (2000). Technical adequacy of probes for curriculum-based measurement of reading growth. *The Journal of Special Education*, 34(3), 140-153.

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.

Snow, C. E., Burns, S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children*. Washington, D.C.: National Academy Press.

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.

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

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

Torgesen, J. K. (2002). The prevention of reading difficulties. *Journal of School Psychology,* 40, 7-26.

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.

Vellutino, F. R., Scanlon, D. M., & Lyon, G. R. (2000). Differentiating between difficult-to-remediate and readily-remediated poor readers: More evidence against the IQ-achievement discrepancy definition of reading disability. *Journal of Learning Disabilities*, 33, 223-238.

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.

Wagner, R. K., Torgesen, J. K., Laughon, P., Simmons, K., & Rashotte, C. A. (1993). Development of young readers' phonological processing abilities. Journal of Educational Psychology, 85, 83-103.

Wood, F. B., Hill, D. F., Meyer, M. S., & Flowers, D. L. (2001). *Predictive Assessment of Reading*. Winston-Salem, NC: Wake Forest University School of Medicine.