Istation's Indicators of Progress (ISIP) Advanced Reading

Technical Report

Computer Adaptive Testing System for Continuous Progress Monitoring of Reading Growth for Students Grade 4 through Grade 8

Patricia Mathes, 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

Istation's Indicators of Progress (ISIP) Advanced Reading

Technical Manual (2016)

by Patricia Mathes, Ph.D.

Acknowledgements

Contributions by:

Reid Lyon, Ph.D.
Gale Roid, Ph.D.

With assistance from:

Kevin Kalinowski, Ph.D.
Dawn Levy, M.Ed.
Beverly Weiser, Ph.D.
Diane Gifford, M.Ed.
Jenny Lawton, M.Ed.
Olga Palchik, M.S.
Ratna Englehart, M.Ed.
Margaret Lamar
Tracey Roden, M.Ed.

Table of Contents

Chapter 1: Introduction 1-1

Background and Significance 2

The Need for Continuous Progress Monitoring 3

Computer Adaptive Testing 4

Continuous Monitoring of Advanced Reading Skills 5

ISIP Advanced Reading Assessment Domains and Subtests 8

ISIP Advanced Reading Item Design and Development 14

The ISIP Advanced Reading Link to Instructional Planning 17

Chapter 2: 2-1

Data Analysis and Results 4

CAT Algorithm 5

Chapter 3: 3-1

Reliability 1

Evidence of Validity 3

Concurrent Validity 21

Results 24

Discussion 27

Conclusion 29

Chapter 4: 4-1

Computing Norms 4

Instructional Tier Goals 5

References Ref-1

Chapter 1: Introduction

ISIP™, Istation's *Indicators of Progress, Advanced Reading* (ISIP Advanced 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, and across, academic years. ISIP Advanced Reading is the upward extension of a similar CAT reading assessment for Grades Pre-K to Grade 3, *Istation's Indicators of Progress, Early Reading* (ISIP Early Reading). ISIP Advanced Reading is the culmination of many years of work begun by Patricia G. Mathes, Ph.D. on extending computerized CPM applications to middle grades while assisting teachers with information about student reading ability across the academic year.

Designed for students in Grades 4–8, ISIP Advanced 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 Advanced Reading provides growth information in four critical domains of reading, including Word Analysis, Text Fluency, Vocabulary, and Comprehension. It is designed to (a) identify specific reading needs of the older struggling reader, (b) provide automatic continuous progress monitoring of skills, and (c) provide immediate and automatic linkage of assessment data to student-learning needs, which facilitates differentiated instruction.

ISIP Advanced Reading has been designed to automatically provide continuous measurement of Grade 4–8 students throughout the school year, as well as across school years, in critical areas of reading, including word analysis, text fluency, vocabulary, and comprehension. Importantly, there are no other valid reading assessment tools for the middle grades that can precisely identify reading abilities and deficits in all four critical domains for the purposes of continuous and differentiated instruction. This is accomplished through short tests, or "probes," administered at least monthly, that target critical areas to inform instruction. 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 thirty minutes or less. It is feasible to administer ISIP Advanced 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.

Background and Significance

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, more fully employed, and better informed.

Sadly, teaching every child to read is a goal we are far from achieving. Large portions of our children continue to struggle to become competent readers (National Reading Panel, 2000; Lyon, 2005). By the middle grades (Grades 4-8), students are expected to demonstrate the ability to read and comprehend grade-level, content-area texts. Yet, for most middle grade students, this is not their reality. The 2007 National Assessment of Educational Progress (Lee, Grigg, & Donahue, 2007) indicates that 74% of 8th graders nationwide struggle to read and gain information from their textbooks, making success in school very difficult. 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). Thus, the middle grades may be the last opportunity for older readers to "catch up" (Bryant et al., 2000).

Older struggling readers are often casualties of prior inadequate reading instruction that insufficiently taught the critical skills necessary for fluent reading and deep processing of text. Many of these students are able to "catch up" in critical reading areas with sufficient targeted instruction (Torgesen et al., 2007). However, many students in the middle grades have little access to effective reading instruction, simply because there

are no reliable and valid assessments that can help their teachers to provide targeted instruction tailored to their needs. Without effective assessments to assist teachers in providing data-informed instruction, many students make little progress year-to-year. This lack of progress is particularly damaging during the middle grade years (Grades 4-8), where learning content-area subject matter becomes a priority. Put succinctly, children who have not learned to read cannot read to learn.

These are not new findings. Overall reading achievement in the United States has remained flat since 1971, when national data were first reported. Because of this alarming and persistent trend, the National Institutes of Health (NIH), through the National Institute of Child Health and Human Development (NICHD), initiated a comprehensive, multidisciplinary effort in 1983 to (1) map the cognitive, linguistic, perceptual, genetic, and neurobiological foundations of reading development; (2) determine the causes of reading failure; and (3) identify and/or develop effective interventions for struggling readers (Lyon, 1985, 1999, 2002; Lyon & Gray, 1992; Lyon & Moats, 1997). Beginning in 1997, and every year until 2005, NICHD program scientists testified on the status of this research in response to requests from Congressional House and Senate Education and Health Committees (Lyon, 1997, 2002b-2005).

These requests were based, in part, on Congressional concerns that the consequences of reading failure went far beyond difficulties in school. NICHD scientists continue to report replicated data showing that reading failure not only constitutes an educational problem, but also a social and public health problem. Specifically, low reading performance is the strongest predictor of students dropping out of school. Consequently, dropouts are more than eight times as likely to be in jail or prison as high school graduates, and nearly 70% of prison inmates score at the lowest two levels of literacy (below fourth grade), with 19% being completely illiterate (Lyon, 1997, 1998). Equally alarming is that poor reading portends adverse health disparities and outcomes, including increased incidence of chronic illness, drug and alcohol abuse, risky sexual behavior, less than optimal use of preventive health services, difficulties accessing medical care, and difficulties understanding health risks (Lyon, 2002a).

The Need for Continuous Progress Monitoring

While the statistics for the long-term outcomes of reading failure are grim, the solution (i.e., reading success for all students) has thus far eluded our schools. While ultimately we want all children to leave the early grades reading, the fact that so many children leave the early grades without a firm foundation for reading suggests that teachers in the middle grades require help to better serve their students. Importantly, a number of efficacy studies have demonstrated that middle grade students are able to "catch up" in critical reading areas with sufficient differentiated instruction (Fletcher, Lyon, Fuchs, & Barnes, 2007; Torgesen et al., 2007). However, for students to receive such targeted instruction, their teachers must first have information about which areas and skills to target for which students.

Teaching that includes frequent monitoring of student progress has been shown to produce higher student outcomes in reading and mathematics than when monitoring is absent (Conte & Hintze, 2000; Mathes, Fuchs, Roberts, & Fuchs, 1998; Ysseldyke & Bolt, 2007). Also, teachers who use Continuous Progress

Monitoring (CPM) data to plan instruction 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, Fuchs, Hamlett, & Stecker, 1991; Mathes et al., 1998). Thus, in order to differentiate, teachers must have reliable and valid CPM assessment tools to (a) determine the specific reading needs of individual children at all levels of the achievement continuum, (b) determine which instructional methods and strategies would be most effective, and (c) monitor children's progress frequently (i.e., at least monthly) over time so that instructional changes can be made when necessary. Unfortunately, assessment tools for any grade level that meet all of these criteria are sorely lacking. Currently, the only CPM reading assessments available for students in the middle grades require one-to-one administration by a teacher to a student.

Computer Application

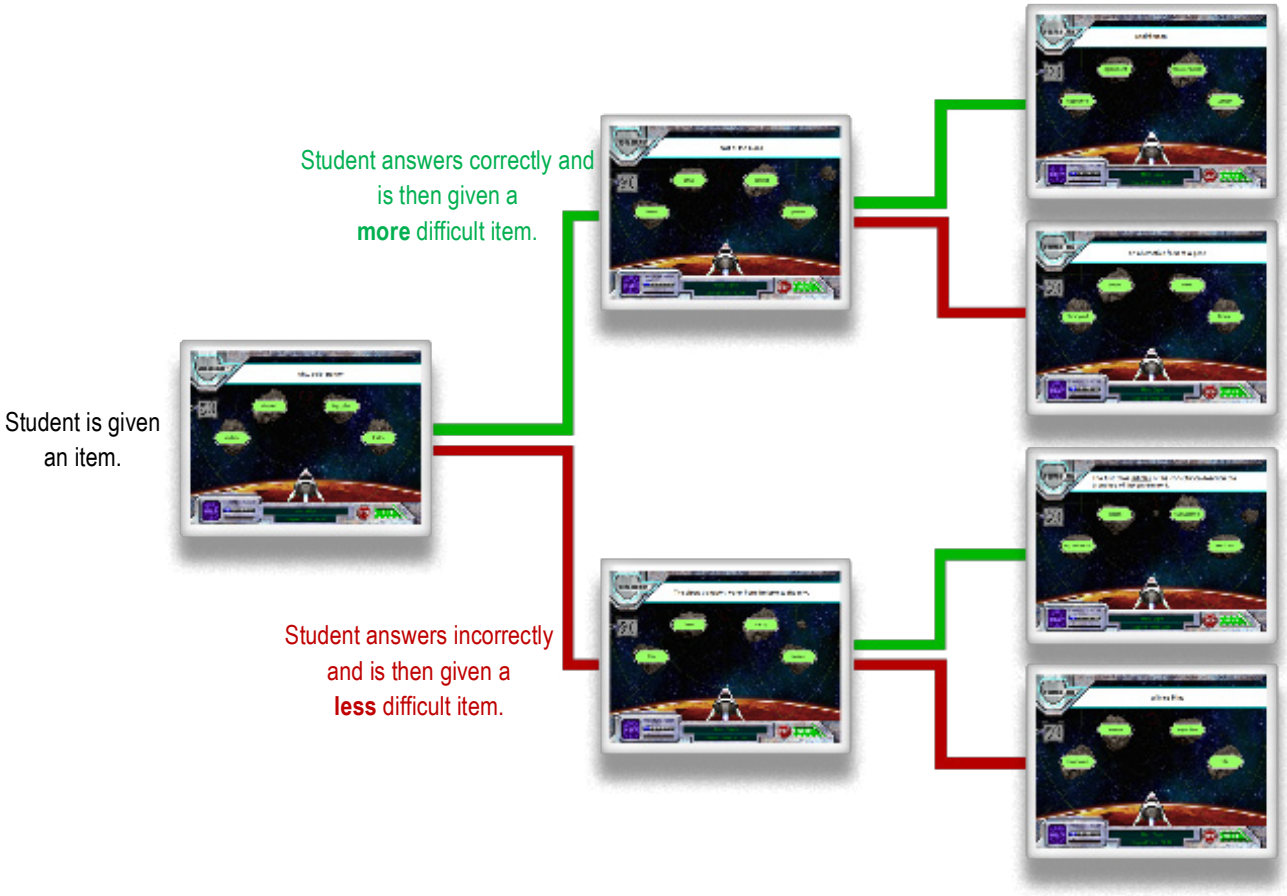
The problem with most CPM systems is that they have been cumbersome for teachers to utilize (Stecker & Whinnery, 1991). Teachers have to physically administer the tests to each child individually and then graph data by hand. The introduction of handheld technology has allowed for graphing of student 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 mathematics and spelling (Fuchs et al., 1995). Computerized applications save time and money. They eliminate burdensome test administrations and scoring errors by calculating, compiling, and reporting scores. They provide immediate access to student results that can be used to affect instruction. They provide information organized in formats that automatically group students according to risk and recommended instructional levels. Student results are instantly plotted on progress charts with trend lines projecting year-end outcomes based upon growth patterns, eliminating the need for the teacher to manually create monitoring booklets or analyze results.

Computer Adaptive Testing

With recent advances in Computer Adaptive Testing (CAT) and computer technology, it is now possible to create CPM assessments that adjust to the actual ability of each 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 student. 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 of using a CAT model rather than a more traditional parallel forms model, as is used in Dynamic Indicators of Basic Early Literacy Skills (DIBELS). First, it is virtually impossible to create alternate forms of any assessment that are truly parallel. Thus, 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.



Continuous Monitoring of Advanced Reading Skills

The typical infrastructure of the middle grades makes collecting frequent CPM data, one child at a time, onerous for teachers and schools. By the middle grades, reading classes are typically taught for 45 minutes to groups as large as 30 students by one teacher. Even if the actual assessment time per child is fairly short, teachers find the process of collecting data cumbersome and overwhelming (Foorman, Santi & Berger, 2007). Furthermore, just collecting the data does not help teachers determine how to respond to the

data. Even when provided with instructional data on their students, many teachers find it difficult to determine the specific needs shared by several students and to group students for differentiated instruction (Foorman, Santi, & Berger, 2007).

This situation is made more difficult because the teacher-administered CPM assessments currently on the market do not actually provide information on all critical areas of reading. Typically, reading fluency is the only area included for middle grade students (Silberglitt, Burns, Madyun, & Lail, 2006). Recent studies indicate that a more comprehensive assessment of reading ability is required for these students (Torgesen et al., 2007; Roberts, Torgesen, Boardman, & Scammacca, 2008; Scammacca et al., 2007). These syntheses suggest that four key areas of reading are significant in understanding comprehensive reading ability in middle grade students: (a) word analysis of multisyllabic words, (b) reading fluency that allows attention to be focused on understanding, (c) vocabulary development that helps students recall terms and provides interaction with students' prior knowledge by exploring semantic and syntactic relationships in text, and (d) reading comprehension skills.

Significance of Assessing Word Analysis

Accurate and automatic identification of multisyllabic words is critical to comprehension of middle grade content-area texts (Deshler et al., 2001; Gersten, Fuchs, Williams, & Baker, 2001) and distinguishes good readers from poor readers (Perfetti, 1986). Good readers use word components or parts, such as knowledge of syllable types, prefixes, suffixes, and roots, to identify long, multisyllabic words (Lenz & Hughes, 1990; Perfetti, 1986). Targeted instruction in advanced word analysis can improve reading outcomes by teaching students strategies to effortlessly recognize increasingly complex words that they encounter in text (Scammacca et al., 2007).

A valuable way to assess word analysis is though spelling. Correct spelling requires that a student possess a fully specified orthographic representation for each word, thus providing valuable information about the student's word analysis skills (Bourassa & Treiman, 2001; Ehri, 2000; Ehri & Wilce, 1987; Graham, 2000; Perfetti, 1997). Students are asked to spell multisyllabic words that are carefully selected to contain the various aspects of syllables, affixes, and roots. Scoring occurs at the syllable unit rather than the whole word, allowing for assessment not only growth in word analysis, but also to provide diagnostic information about which structural aspects of words particular students find challenging.

Significance of Assessing Fluency

The ability to read connected text with both speed and understanding is the true hallmark of a fluent reader. Successful older readers identify most of the words in text "automatically," allowing them to focus on higher order processes such as understanding, inferring, and interpreting (Archer, Gleason, & Vachon, 2003; Osborn, Lehr, & Hiebert, 2003). While fluency does not cause comprehension, it does play a facilitative role (Rasinski et al., 2005). Furthermore, measuring fluency has been shown to be a good gauge of overall reading health (Deno, 2003; Fuchs & Fuchs, 2008) in much the same way that a thermometer measures

general physical health. Current CPM fluency measures consist primarily of oral reading tasks. However, such a task does not measure if students are monitoring meaning. ISIP Advanced Reading uses a maze task to measure both text processing speed and understanding, as required for assessing comprehensive fluency. In a maze task, students read text in which every seventh word is blank. For each blank, students are given three choices with which to fill in the blank. Such tasks have been shown to highly correlate to oral reading tasks and to comprehension tasks (e.g., Deno, 2003; Fuchs & Fuchs, 2008).

Significance of Assessing Vocabulary

In the past decades, the importance of vocabulary knowledge in the development of reading skills has been extensively established in the literature (National Reading Panel, 2000). Moreover, for children historically at risk of reading difficulties due to poverty and language background, oral language, in general, and vocabulary, in particular, are critical to reading success (Hemphill & Tivnan, 2008; Pearson, Hiebert, & Kamil, 2007). Students need instruction that accelerates their acquisition of new vocabulary and provides deep knowledge about words. (Beck, McKeown, & Kucan, 2002) suggest breaking words into three tiers. Tier 1 words are words that students are likely to know (e.g., sad and funny). Tier 2 words appear frequently in many contexts (e.g., regardless and compromise). Tier 3 words appear rarely in text or are content-specific (e.g., irascible and biogenetics). Beck and colleagues suggest that teachers focus vocabulary instruction on Tier 2 words drawn from content-area materials that contain words students are likely both to need (because they are encountered across contexts) and to learn well (because students will have repeated opportunities for practice and use). However, Tier 3 words represent a specific challenge to students since these words are the jargon of the content areas (Bravo & Cervette, 2008). ISIP Advanced Reading focuses on both Tier 2 words (general vocabulary) and Tier 3 words (content-specific).

Significance of Assessing Reading Comprehension

Reading well is a demanding task requiring coordination of a diverse set of skills (Irwin, 1991). Struggling readers, even those with adequate word-level skills and acceptable fluency, often fail to use these types of strategies, either because they do not monitor their comprehension or because they lack the necessary tools to identify and repair misunderstandings when they occur. Effective reading comprehension interventions have focused on helping students to become strategic readers by teaching them how to think while they are reading. Effective interventions have included single strategies such as finding the main idea and self-monitoring (e.g., Chan, 1991; Malone & Mastropieri, 1992) and multi-component strategies that target reading sub-strategies (e.g., Jitendra, Hoppes, & Xin, 2000; Schumaker, Deshler, Alley, Warner, & Denton, 1982). Additionally, student-led discussions of predictions, text structure, and summary development within interactive small groups have produced improvements in understanding and recalling expository text (Englert & Mariage, 1991). It is important that assessments of comprehension provide information about specific comprehension abilities that are amenable to instruction.

ISIP Advanced Reading uses four broad areas of comprehension, which allow assessment of general growth in comprehension and provide diagnostic information to teachers to guide instruction. Specifically,

ISIP Advanced Reading assesses (a) main idea, (b) cause and effect, (c) inference, and (d) critical judgment. After silently reading passages, students answer questions representing these four areas of comprehension ability.

ISIP Advanced Reading Assessment Domains and Subtests

The ISIP Advanced Reading assessment will comprise four subtests, representing the four domains of reading previously identified. The domain of Word Analysis will be assessed through the spelling subtest. The domain of Fluency will be assessed through the connected text fluency subtest. The domain of Vocabulary will be assessed by the vocabulary subtest, which will include both general and content area vocabulary. The domain of Comprehension is assessed by the comprehension subtest, which includes several types of comprehension abilities, including determining main idea, making inferences, making critical judgments, and determining cause-and-effect relationships.

Word Analysis Subtest

Students demonstrate if they have fully specified orthographic representations of words in the English language by spelling selections from among 1,090 carefully chosen words that incorporate the various aspects of English orthography. To choose these words, the development team first identified approximately five hundred words using grade-level word lists for Grades 2–14 and analyzed their spellings for number of syllables; syllable types; Anglo-Saxon, Greek, or Latin roots; affixes; derivatives; inflectional endings; consonant doubling; irregular elements; variant spellings; and unaccented syllable schwa. These grade-level lists of words were then coded by approximate difficulty with numbers 1 through 5, 1 being the most difficult (i.e., having the most elements). Thirty words from each of these difficulty levels were randomly selected from each grade-level list, resulting in a total of 150 words for Grades 4–8. Furthermore, we created an additional 150 items at Grade 3 and 75 additional Grade 2 items. An additional 300 items were developed to represent Grade 9–14 abilities. Although the difficulty levels for the items were determined based on theory, the IRT Calibration Study provided the definitive information regarding the difficulty of each item.

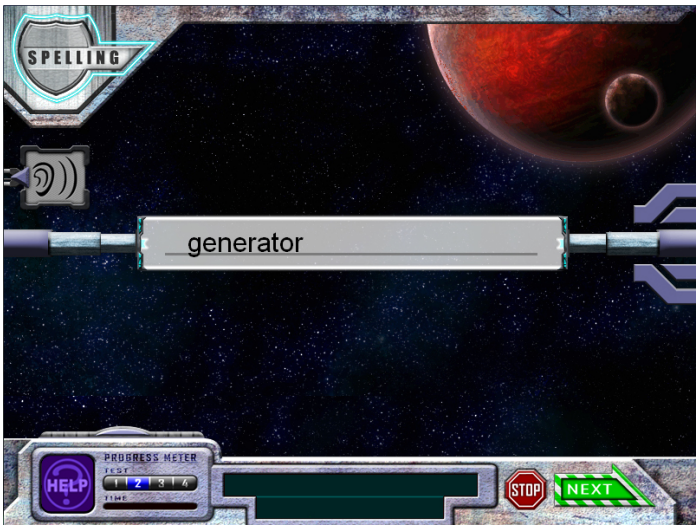
Theory and Research

It is known that proficient spellers almost always possess strong word recognition ability, and that good readers typically read at levels near their spelling ability (Foorman & Francis, 1994; Ehri, 2005). Furthermore, better spelling ability is associated with better word recognition, fluency, and comprehension ability (Harn & Seidenberg, 2004). Thus, there appears to be a synergy between spelling and reading (Joshi, Treiman, Carreker, & Moats, 2008; Moats, 2005; Weiser & Mathes, 2009). Learning to spell words and learning to read words are thought to be related like two sides of a coin, because they both rely on the same knowledge about the alphabetic system and memory for the spellings of specific words (Bourassa & Treiman, 2001; Ehri, 2000; Ehri & Wilce, 1987; Graham, 2000; Moats, 2000, 2005; Perfetti, 1997). Ehri's

connectionist theory (Ehri, 1997, 1998, 2000) suggests that spelling and reading, although independent skills, develop together reciprocally due to a logical symmetry relationship. Children who spell poorly demonstrate more problems with combining both phonological and orthographic processes than children who spell well, and children learn about language through print because print provides children with a schema for conceptualizing and analyzing the structure of speech (Ehri, 1998; Ehri 2005). Thus, if one wants to assess how well students are combining phonological and orthographical information with complex multisyllabic words, then assessing students' ability to spell such words is the logical choice.

Procedure

For this subtest, a line appears on the screen above the graphic of a keyboard. The computer asks students to spell a word. The computer then says the word in a sentence and repeats the word. Students use their computer keyboard to type the word. As they type, the letters light up on the keyboard that appears on the screen, and the letters appear on the line in the order they are typed. The purpose of the computer screen keyboard is to assist students in keeping their eyes on the screen rather than looking at their fingers as they type. If a student needs to hear a word again, the student has the option of clicking on an icon to hear the word repeated. Words are selected for each student based on the CAT procedure adapting to the student's estimated.



Connected Text Fluency Subtest

Students demonstrate their ability both to read words quickly and to monitor for meaning while reading grade level connected text. This subtest is constructed in a very different manner than the other subtests. Rather than increasing text difficulty across time, the test assesses students on passages of equivalent difficulty to measure growth over time against a constant level of difficulty. Thirty 500- to 700- word stories of near equivalent difficulty have been developed for each of the five target grades, for a total of 150 stories. Each of these stories is carefully written to conform to specific word-level features, follow linear narrative structure, and have readability according to Flesh-Kincaid and Lexile® units for end-of-grade level in the

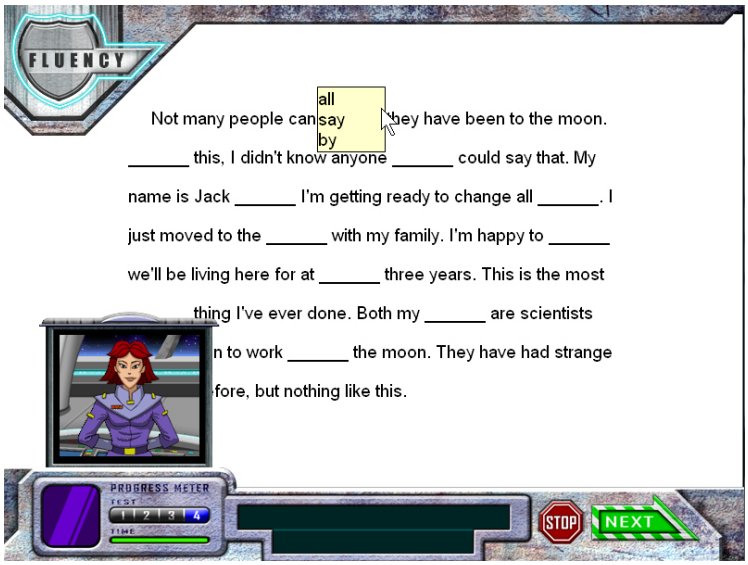
targeted grade. To assess text reading for understanding, a maze task is utilized in which every seventh word is left blank from the text. The student is given three choices for each blank from which to choose the word that works in the sentence. It is the student's job to read the text, selecting the correct maze responses for two and one-half minutes.

Theory and Research

Successful fluent readers read connected text with both speed and understanding (Archer, Gleason, & Vachon, 2003; Osborn, Lehr, & Hiebert, 2003). In order to assess the full scope of fluency, measures need to incorporate both speed and meaning aspects of fluency. The maze task has been shown to be highly correlated to measures of both fluency and comprehension and has high reliability and concurrent validity (Brown-Chidsey, Davis, & Maya, 2003; Fuchs & Fuchs, 1991; Jenkins, Pious, & Jewell, 1990; Shinn, Good, Knutson, Tilly, & Collins, 1992; Swain & Allinder, 1996). A similar task was part of the ISIP Early Reading assessment. The data confirms that the maze task, delivered via computer, correlates highly to measure oral reading fluency, comprehension measures, and high stakes assessments (Kalinowski, 2009).

Procedure

To complete connected text fluency, the computer tells students it is time to read a story and review the procedures. The first page then appears, and students perform the maze task for two and one-half minutes or until they complete the story. When students complete a page, they click on a button to turn the page and continue. The score obtained from this task incorporates the number and accuracy of maze items completed in the allocated time, and it accounts for the number of words read between mazes. This score, which Istation formulated for ISIP Early Reading, has been shown to better correlate to other measures of both DIBELS Oral Reading Fluency and Comprehension (Lyon & Kalinowski, 2008).



Vocabulary Subtest

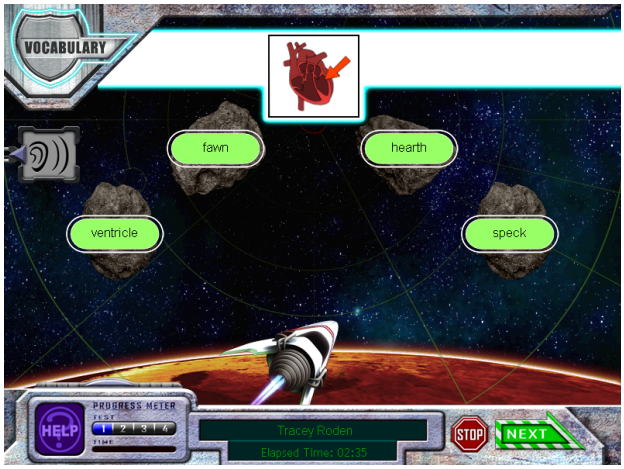
Students demonstrate their knowledge of word meanings through synonyms or definitions, as well as the ability to infer meaning through context. Four types of questions are used: (a) select the word that best matches the following definition, (b) select the word that is most similar in meaning to the following word, (c) select the word that best describes the following picture, and (d) select the word that is most similar in meaning to the underlined word. Distractor choices for each word include words with a similar spelling or pronunciation, antonyms, and words with an unrelated meaning.

Theory and Research

In order to assess students' knowledge of word meaning, decontextualized types of items (synonyms, pictures, and definitions) are used. However, it is known that students acquire vocabulary best when it is used in a meaningful context. Thus, contextual types of questions are also included, in which students must infer the correct meaning of a word based on its use in a sentence. Passive recognition tasks have been chosen for assessment, based on reports that the ability to establish the link between word form and word meaning is the most important component of word knowledge (Laufer et al., 2004; Read, 2007).

Procedures

Throughout the vocabulary assessment, there is a mix of general vocabulary words and content vocabulary words. The narrator reads the stem for each item. Students can choose to hear the word choices by scrolling over each word on the screen. Students choose from among four possible answers by clicking their mouse on their selected answer. The CAT program matches the difficulty of the items to the students' abilities, regardless of their age or grade level. Teachers are able to access reports of their students' progress and necessary areas of vocabulary instruction.



Reading Comprehension Subtest

The objective of the Reading Comprehension subtest is to determine how well students are processing text of increasing difficulty for meaning. Two hundred twenty graduated passages were constructed (ranging in readability from 2.0 through 12.9 on the Flesh-Kinkaid scale) for students to initially read silently. After reading, students answer a series of four multiple-choice questions. Passages are a mix of narrative and expository text and target main idea, cause/effect or problem/outcome, inference, and critical judgment of the text. The underlying theory driving this assessment is that comprehension requires both low-level and high-level processing of text information. It is in the higher level processing that the deeper message of the text comes through. Thus, the Reading Comprehension subtest is being crafted to assess higher cognitive levels of comprehension, with the goal of constructing questions that are both conceptually and instructionally valid.

Theory and Research

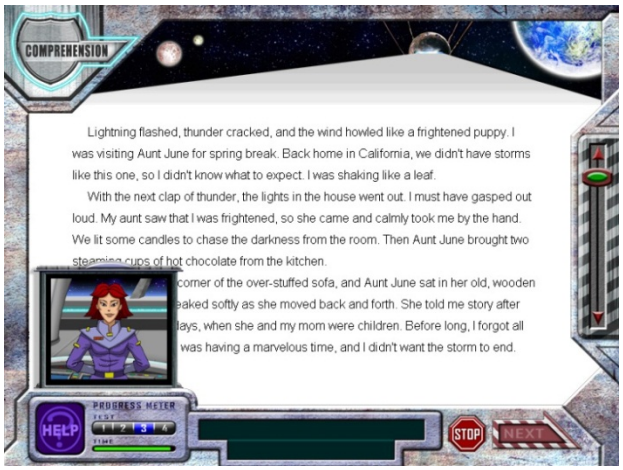
The proposed view of comprehension aligns with the most current understanding of reading comprehension. Higher-level processing of text is defined as the reader's ability to determine the overall idea of the passage, differentiate and switch between broader and narrower concepts (essence vs. details), inhibit irrelevant information from intruding upon meaning, monitor comprehension, reason, make inferences, and integrate information into long-term memory (Gamino & Chapman, in press; Kintsch, 1998; Oakhill, Hartt, & Samols, 2005; Sesma et al., 2009; Williams, 2003; Yuill & Oakhill, 1991).

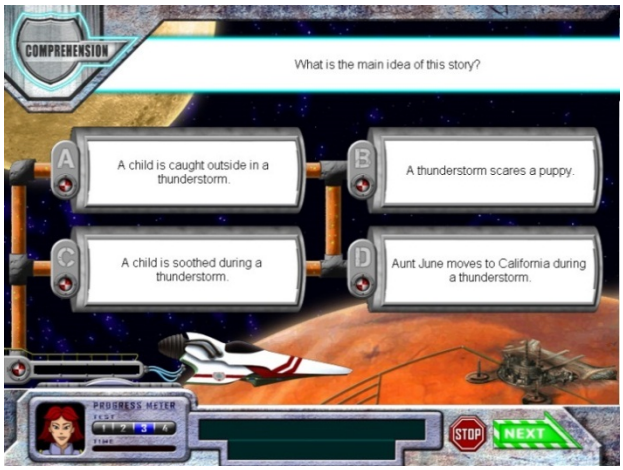
In item construction, care was taken to assess students' coherence of knowledge generation (Kintch, 1998), or the ability to make higher-level links between individual sentences to establish local coherence (i.e., cause/ effect and inference question types) and to integrate new information into existing representations to establish global coherence of text (i.e., main idea, problem/outcome, and critical judgment question types) (Cain & Oakhill, 1999; Cain, Oakhill, Barnes, & Bryant, 2001; Oakhill, 1982; Wixson & Peters, 1987). Furthermore, all questions are being designed to be dependent upon information in the passage in order to avoid the testing of background knowledge or having questions that can be answered without reading the text. This situation has been a pitfall of other well-known tests (Keenan & Betjemann, 2006). All answer

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

Procedures

To complete the comprehension subtest, students first read a passage that appears on the screen. The computer tells them to read the passage for meaning. When they are ready, they turn the page, and the first of four questions appear. When they complete a question, the next question automatically appears. During the test, students are not allowed to go back and review the passage. All assessment items are multiple choice, allowing students to select from four possible answers. Students select answers by clicking the mouse on their selected responses. Teachers are able to access information about their students' text levels, such as overall performance in comprehension based on the student ability index score. Teacher reports include diagnostic information about skill-specific deficits and recommendations for interventions to meet deficiencies.





ISIP Advanced Reading Item Design and Development

Along with the authorship team, graduate students from the Institute for Evidence-Based Education at Southern Methodist University (SMU) began item development by asking the following question: What are the best ways to assess these domains with middle-grade students via computer administration?

Knowing that middle-grade students need to be assessed in Word Analysis, Fluency, Vocabulary, and Comprehension, a search of the literature was then completed to locate studies that focused on how to best assess each of these four dimensions of reading, as well as possible confounds to these design assessments. An extensive of the extant research literature base on how to best assess each of the four areas was conducted to provide our team clarity about the most current understanding about assessment techniques for assessing advanced Word Analysis, Vocabulary, Fluency and Comprehension of students in the middle grades. The results of this search provided great insight into the issues involved in assessing each of the four domains, as well as current thinking about how best to assess each domain. The authorship team was greatly influenced by Cutting and Scarborough's (2006) call to develop new instruments that correspond more closely to theoretical models of the constructs being measured. Thus, much time was spent defining our models for each of the four constructs and designing items to assess the models. It was further examined how each of the four domains of reading has been assessed in other widely accepted assessments. 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.

This work was particularly helpful in guiding decisions on how to assess comprehension. Reading comprehension difficulties are found in as many as 15% of students, even though they may not display lower level or surface processing deficits (i.e., decoding, word recognition, fluency, and/or language comprehension) (Cain & Oakhill, 2007; Fletcher et al., 2007; Nation, 1999; Nation et al., 1999; Yuill & Oakhill, 1991). Understanding how students comprehend text at higher cognitive levels is necessary for advancement and intervention. There is consensus among the reviewed literature that reading comprehension assessments have been one-dimensional and have had little variation in reading material or response formats, and that current assessments provide little diagnostic information because they lack precision in measuring the underlying latent variables that comprise comprehension (Cutting & Scarborough, 2006; Deane, Sheehan, Sabatini, Futagi, & Kostin, 2006; Fletcher, 2006; Francis, Snow, August, Carlson, Miller, & Iglesius, 2006; Millis, Magliano, & Todaro, 2006; Rayner, Chace, Ashby, & Slattery, 2006).

Taking this into consideration, comprehension measures were created that provided the precision needed to guide instruction by carefully constructing passages and designing questions to target specific skills within the construct of comprehension. The review of the literature also resulted in the awareness of the need to ensure that passages and their associated comprehension questions are dependent, to ensure that students must actually process text in order to answer questions correctly. Passage independence has been found to be problematic in a number of comprehension assessments (Keenan & Betjemann, 2006). As

comprehension items were developed, work was checked by asking high-performing middle-grade students the questions without asking them to read the associated passages. If questions could be answered correctly, they were removed from the item bank. Likewise, much attention was placed on the fact that many common measures of comprehension appear to be more highly linked to decoding ability than to comprehension (Keenan, Betjemann, & Olson, 2008; Cutting & Scarborough, 2006). This problem was solved by matching text difficulty level to a student's text reading ability (i.e., matching passage difficulty to student ability), allowing the assessment of ability to be in processing text for meaning.

In building the blueprint for the items within each domain, in terms of item types and number of items representing the span of skills development, the state standards for California, Florida, New York, and Texas, were reviewed for Grades 4-8. 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. Next, the "big ideas" that were consistent across all states were identified, which can be summed up in three statements: (a) students should easily recognize increasingly complex words, (b) students should fluently process grade-level materials in a variety of genres, and (c) students should be able to derive meaning from grade-level texts representing a variety of genres. The common skills associated with deriving meaning identified by all states examined included: (a) determining a text's main ideas and how they are supported in the text, (b) analyzing text to determine the author's purpose, (c) analyzing plot structure and literary devices in a story, (d) determining and explaining cause-and-effect relationships, (e) drawing conclusions and making predictions based on the text, (f) comparing and contrasting information in the text, (g) determining the sequence of events, and (h) distinguishing between fact and opinion. Embedded in these skills is knowledge of increasingly sophisticated vocabulary. Beyond these skills categories, the states that were analyzed also specified expectations for the level of refinement expected of students within each skill area for each grade. Using this information, a flow chart by grade was created, illustrating each domain, skills within each domain, and plotted skill-development expectations. This served as the foundation of the assessment blueprint.

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

With a blueprint in hand, the team developed items. ISIP Advanced Reading is composed of 3,100 items (Spelling = 1,090, Vocabulary = 760, Connected Fluency Stories = 150, Comprehension passages = 220, and Comprehension questions = 880 [4 per passage]). Within the 4 domains, the complete item pool is distributed across the full continuum of middle school ability (i.e., Grades 2-14).

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.

Teacher Friendly

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

Student Friendly

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

The ISIP Advanced Reading Link to Instructional Planning

ISIP Advanced Reading provides continuous assessment results that can be used in recursive assessment instructional decision loops. First, ISIP Advanced Reading identifies students in need of support. Second, validation of student results and recommended instructional level can be easily verified by re-administering assessments, increasing 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 Advanced 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 level of support needed as well as skill needs. Data are provided in both graphical and detailed numerical formats 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 Advanced 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 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 those from the current year and all prior years, is maintained for each student. On the report, teachers may acknowledge that suggested interventions have been provided. A record of these interventions is maintained with the student history as an Intervention Audit Trail. This history can be used for special education Individual Education Programs (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.

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 Advanced Reading results is available. Administrators, principals, and teachers may use their reports to evaluate and modify curriculum, interventions, Adequate Yearly Progress (AYP), the effectiveness of professional development, and personnel performance.

Chapter 2: IRT Calibration and the CAT Algorithm of ISIP AR

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 2009-2010 school year, data were collected from two large north Texas independent school districts (ISD), and one large independent private school organization, labeled AISD, BISD and CISD henceforth. Seven elementary schools from AISD, two middle schools from BISD, and 2 K-8 campuses in CISD were recruited for the study. At each AISD school, all 3rd-, 4th-, and 5th-grade students in general education classrooms were asked to bring home introductory letters and study consent forms, which had prior approval by both the school district and Southern Methodist University's (SMU's) Institutional Review Board. All 6th-grade students at both BISD schools, all 7th graders and 2 classes of 8th graders at BISD school 1, all 8th graders at BISD school 2, as well as all 6th, 7th, and 8th graders at CISD school 1 and all 5th, 7th, and 8th graders at CISD school 2 were given the aforementioned introductory letters and study consent forms. Table 2-1 shows the number of students recruited at each school and the number of students with signed consent (participating students). The three districts represented socially and ethnically diverse populations. Table 2-2 shows the demographics of participating students from each district.

Table 2-1: Number of Students in Study

School District	Signed Consent Forms	Total Students	Percent of Students with Signed Consent Forms
AISD	1,711	2,017	84.43
A.1	303	356	85.11
A.2	149	214	65.89
A.3	253	277	91.34
A.4	353	389	90.75
A.5	273	298	91.61
A.6	212	272	77.94
A.7	168	211	79.62
BISD	790	1,024	76.95
B.1	399	540	73.89
B.2	391	484	80.37
CISD	262	284	92.25
C.1	180	194	92.78
C.2	82	90	91.11
TOTAL	2,763	3,325	82.80

Table 2-2: Demographics of Participating Students

Demographic	n	%
Total Number of Students	2,763	
Districts		
Dallas Catholic Diocese	262	9.48%
Garland ISD	1,711	61.93%
Granbury ISD	790	28.59%
Ethnicity		
African American	258	9.34%
Alaska Native	1	0.04%
American Indian	13	0.47%
Asian	140	5.07%
Pacific Islander	1	0.04%
Latino/a	744	26.94%
Filipino	6	0.22%
White	1564	56.63%
Other	35	1.27%
Gender		
Male	1,388	50.24%
Female	1,375	49.76%
Enrolled in Special Ed.		
Yes	215	7.78%
No	2,548	92.22%
Classroom Instruction		
General Education	2,424	87.76%
ESL	122	4.42%
Bilingual	216	7.82%
Economic Disadvantage		
Yes	1,163	42.11%
No	1,599	57.89%
English Proficiency		
Native	2,479	89.72%
ELL	238	8.61%
Former ELL	46	1.66%
Disability Condition†		
Auditory Impairment	2	0.50%
Autistic	6	1.50%
Emotional Disturbance	18	4.51%
Learning Disability	127	31.83%

Demographic	n	%
Mental Retardation	8	2.01%
Other Health Impairment	50	12.53%
Orthopedic Impairment	4	1.00%
Speech Impediment	182	45.61%
Traumatic Brain Injury	1	0.25%
Visual Impairment	1	0.25%
Grade Level		
3rd	586	21.21%
4th	605	21.90%
5th	541	19.58%
6th	415	15.02%
7th	293	10.60%
8th	323	11.69%

†Percentage is relative to total number of students with a Disability Condition, 399.

*Totals 2,762 students. These categories were not provided for 1 student.

Students were escorted by trained SMU data collectors, typically graduate students, project coordinators and/or research associates, in convenience groupings to the school's computer lab for 30-minute sessions on the ISIP- AR program.

It was unrealistic to administer all of the items to each student participating in the study. Therefore, items were divided into grade-specific subpools. Except for 3rd grade, each grade-specific subpool also included 10% of the items from the grade below (eg. the 5th-grade pool included 10% of the 4th-grade items), to be used for comparison and vertical scaling. Each participant was administered all of the items in the subpool for their grade level. Table 2-3 shows the numbers of items in each grade subpool, not including the 10% overlap items.

Table 2-3: Items Used in Study

Skill	Grade					
	3rd	4th	5th	6th	7th	8th
Spelling	150	150	154	149	150	151
Vocabulary	44	122	120	119	116	117
Comprehension	108	136	136	120	120	116
TOTAL	302	408	410	388	386	384

To control for order main effects, participating students were assigned items from their grade subpool in random order until they had answered all of the items in the subpool. The total number of sessions required to answer all items varied by participant.

Testing for all three districts took place between March 2010 and May 2010. 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 700 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

$$P(x=1) = \frac{1}{1+e^{-a_j}(\theta_i - b_j)}$$

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

To estimate the item parameters, BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003) was used. BILOG-MG uses marginal maximum likelihood estimation (MMLE) to maximize the person response vector across both the item difficulty and discriminability dimensions. For example, Equation 2 represents the probability of a response vector of dichotomous items, *X*, in an instrument of length *L*,

$$P(X \mid \theta, J) = \prod_{j=1}^{L} p_j^{x_j} (1-p_j)^{1-x_j}$$

where the probability of a set of responses is conditioned on the person's ability (θ) and the matrix of item parameters, *J* (i.e., the collection of a s and b s for each item, *j*). In MMLE, an unconditional, or marginalized, probability of a randomly selected person from the population with a continuous latent distribution is specified as an integral function over the population distribution (Bock & Aitken, 1981). Subsequently, the resulting marginal likelihood function underwent maximum likelihood estimation (MLE) by BILOG-MG to generate item parameters.

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, 2,240 items were used for the ISIP Advanced 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 0.868, indicating that ISIP Advanced Reading is very reliable. The standard error of measurement (SEM) is also computed from IRT point of view. Since the ISIP Advanced Reading scale score is $(200 * \theta) + 2,000$, $SEM(\theta) = 200 * \mathrm{SE}(\theta)$. It is 72.779.

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 Maximum Likelihood Estimation (MLE) 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 end ISIP Advanced Reading when the ability score's standard error drops below a preset threshold or when four consecutive items have each reduced the standard error by less than a preset amount.

Production Assessment

Item types were grouped according to key reading domains for the production assessment. Each grade-level (4th, 5th, 6th, etc.) was given the same set of subtests: 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 uses 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 2,240 Grades 3-8 items with reliable parameter estimates aligned on a common scale, with the majority of items ranging from 650 to 3,000 in difficulty.

After completing this study, which included determining an appropriate IRT model, calibrating the items, and constructing the CAT algorithm, the ISIP Advanced Reading assessment went into full production starting with the 2010-2011 school year.

Chapter 3: Reliability and Concurrent Validity of ISIP AR for Grades 4–8

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™ Advanced Reading data collected from 4th- to 8th-grade students in north Texas elementary and middle schools during the 2010-11 school year.

Reliability

Marginal Reliability

Lee Cronbach's (1951) coefficient alpha is typically used as an indicator of reliability across test items within a testing instance. However, Cronbach's alpha is not appropriate for any item response theory (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 computer adaptive testing (CAT)-based assessment, such as ISIP Advanced Reading, received a custom set of items based on their initial estimates of ability and response patterns. Thus, students did 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 Advanced 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. Marginal reliability coefficient, operates on the variance of the ability scores (σ^2_θ) and the average of the expected error variance ($\bar{\sigma}^2_E$; Sireci, Thissen, & Wainer, 1991):

$$\rho^2 = \frac{\sigma^2_\theta - \bar{\sigma}^2_E}{\sigma^2_E}$$

ISIP Advanced 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 Advanced Reading will always be approximately 0.90.

Test-Retest Reliability

Test-Retest reliability was examined for ISIPAR with a large sample of all 4th- to 8th- grade students in a north Texas Independent School District who were administered ISIPAR during the fall of 2010 as part of their typical instructional program. Assessments were exported from administration of ISIP in October, then again in November. The time between administrations ranged from three to seven weeks. Table 3-1 presents the correlations between the first and second administration.

Table 3-1: ISIP Advanced Reading Test-Retest Reliability between Testing Sessions

	ISIP AR OVR1	ISIP AR CMP1	ISIP AR SPL1	ISIP AR VOC1	ISIP AR TF1	ISIP AR OVR2	ISIP AR CMP2	ISIP AR SPL2	ISIP AR VOC2	ISIP AR TF2
ISIP AR OVR1	1.000[d]	0.910[n]	0.879[f]	0.865[e]	0.785[q]	0.910[u]	0.830[dd]	0.841[y]	0.808[v]	0.775[gg]
ISIP AR CMP1		1.000[n]	0.678[o]	0.730[p]	0.730[r]	0.836[ii]	0.877[mm]	0.681[kk]	0.721[jj]	0.733[qq]
ISIP AR SPL1			1.000[f]	0.663[g]	0.695[s]	0.828[aa]	0.652[ff]	0.921[cc]	0.646[bb]	0.692[ll]
ISIP AR VOC1				1.000[e]	0.682[t]	0.800[w]	0.700[ee]	0.664[z]	0.885[x]	0.656[hh]
ISIP AR TF1					1.000[q]	0.761[nn]	0.695[rr]	0.704[pp]	0.649[oo]	0.823[ss]
ISIP AR OVR2						1.000[a]	0.9.00[h]	0.881[c]	0.852[b]	0.782[k]
ISIP AR CMP2							1.000[h]	0.666[j]	0.708[i]	0.735[k]
ISIP AR SPL2								1.000[c]	0.656[c]	0.705[m]
ISIP AR VOC2									1.000[b]	0.648[l]
ISIP AR TF2										1.000[k]

[a]N = 10292 [b]N= 10291 [c]N= 10227 [d]N = 10119 [e]N= 10114 [f]N= 10049 [g]N = 10044[h]N= 9737 [i]N= 9736 [j]N = 9734
[k]N = 9472 [l]N= 9471 [m]N= 9469 [n]N = 9447 [o]N= 9443 [p]N= 9442 [q]N = 9119 [r]N = 9118 [s]N= 9115 [t]N= 9114
[u]N = 8995 [v]N= 8994 [w]N= 8990 [x]N = 8989 [y]N= 8954 [z]N= 8949 [aa]N = 8936 [bb]N= 8935 [cc]N= 8895 [dd]N = 8643
[ee]N = 8638 [ff]N= 8593 [gg]N= 8448 [hh]N = 8443 [ii]N= 8440 [jj]N= 8439 [kk]N = 8404 [ll]N= 8400 [mm]N= 8195 [nn]N = 8175
[oo]N = 8174 [pp]N= 8143 [qq]N= 8037 [rr]N = 7956 [ss]N= 7814

Evidence of Validity

Construct Validity

Much work had been previously done to establish construct validly of our item pool. The theoretical underpinnings for each subtest had previously been presented (see Chapter 2). Thus we will not repeat that information here. In addition to reviewing the research literature and determining our theoretical approach to measuring the constructs included in ISIP AR (as previously discussed), our team spent considerable time building a precise blueprint guiding construction of the item pool. Creation of this item pool began with a review of state standards for California, Florida, Texas, and New York. Our team then designed a multi-stage process that was followed for each subtest as described below.

Once the item pool was created, 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 to 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 in which they intend to measure.

Spelling (Word Analysis)

For the Spelling subtest, the state standards were listed and then cross-referenced, identifying standards that appeared in more than one state. The wording of the standards was condensed to reduce wordiness and repetition. The most important standards, as evidenced by the state adoptions, were listed in concise terms and are presented in Table 3-2 below.

Table 3-2: Spelling Standards

Spelling Standards					
3rd Grade	4th Grade	5th Grade	6th Grade	7th Grade	8th Grade
One-Syllable Words					
CVC Words	CVC Words	CVC Words			
VC e Silent e	VC e Silent e	VCe Silent e			
Contractions					
Compound Words					
Orthographies	Orthographies	Orthographies	Orthographies	Orthographies	Orthographies
Spelling Rules	Spelling Rules	Spelling Rules	Spelling Rules	Spelling Rules	Spelling Rules
Homophones					
Consonant Doubling					
Changing the -y					
Adding Affixes	Adding Affixes	Adding Affixes	Adding Affixes	Adding Affixes	Adding Affixes
	Roots	Roots	Roots		
	Inflectional Endings				
	Syllables	Syllables			
			Frequently Misspelled	Frequently Misspelled	Frequently Misspelled
	Derivatives	Derivatives	Derivatives	Derivatives	Derivatives
		Spelling Conventions	Spelling Conventions	Spelling Conventions	Spelling Conventions
	Using Resources	Using Resources	Using Resources	Using Resources	Using Resources

We then consulted the following internet sites to select spelling words to be used as possible test items:

- http://www.all-about-spelling.com/ (Grades 3–7)
- http://sk043.k12.sd.us/8th.grade.spelling.lists.htm (Grade 8)
- http://www.eduplace.com/rdg/hmsv/3/wordpuzzles/wp_unit04_wordlist.pdf (Grade 3)
- http://www.essortment.com/family/spellingbeewor_ttyq.htm (Grades 10–12)

Likewise, the following books were utilized to select spelling words as possible test items:

- *The Reading Teacher's Book of Lists*; Edward B. Fry & Jacqueline E. Kress; 5th ed., 2006; Jossey-Bass, San Francisco, CA.
- *Building Words: A Resource Manual for Teaching Word Analysis and Spelling Strategies*; Thomas G. Gunning; 2001; Pearson Education Company, Needham Heights, MA.

Our team then selected words from these sources to align to two or more standards at each grade level. Selected words were organized on Excel lists by grade level and their predicted difficulty was determined by the following: (a) number of standards met, and (b) number of syllables in each word. Each word was then coded as follows, and appropriate sentences using each spelling word were created to be given to the students verbally during assessment.

3rd Grade Spelling Codes: (578 items)

Words were coded # 1-4. Appropriate third-grade sentences using each spelling word were created to be given to the students verbally during assessment.

1 – Words occurring in more than one (grade) list and including at least two standards

2 - Words including two or more standards

3 – Words including a standard other than orthographic knowledge

4 – Words which have only orthographic knowledge as the single standard

4th Grade Spelling Codes: (604 items)

Words were coded # 1-5. Appropriate fourth-grade sentences using each spelling word were created to be given to the students verbally during assessment.

1 – Words containing four or more standards

2 – Words containing three standards or appearing on multiple lists

3 – Words containing two standards

4 – Words containing one standard

5 – Words containing no standards

5th, 6th, and 7th Grade Codes (625 5th Grade items, 733 6th Grade items, 696 7th Grade items)

1 – Words on multiple lists, three or more syllables and three or more standards

2 – Words containing 3 or more syllable and one or two standards

3 – Words containing 2 syllables and three standards

4 – Words containing 2 syllables and one or two standards

5 – All other words

8th Grade and Up Spelling codes (705 8th Grade items, 541 9th Grade and higher items)

1 – Words on multiple lists, 4 or more syllables and/or 4 standards

2 – Words containing 3 syllables and 3 or more standards

3 – Words containing 3 syllables and one or two standards

4 – Words containing 2 syllables and 2 or more standards

5 – All other words

Text Fluency

As with spelling, the state standards for fluency were listed and then cross-referenced, identifying standards that appeared in more than one state. The wording of the standards was condensed to reduce wordiness and repetition. The most important standards as evidenced by the state adoptions were listed in concise terms. Standards amenable to assessment using computer administration we culled to guide item creation.

Table 3-3: Text Fluency Standards

3rd Grade	4th Grade	5th & 6th Grade	7th & 8th Grade
Reading Expository Text with Correct Pacing	Reading Expository Text with Correct Pacing	Reading Expository Text with Correct Pacing	
Reading Narrative Text with Correct Pacing	Reading Narrative Text with Correct Pacing	Reading Narrative Text with Correct Pacing	
	Adjust Reading Rate Based on Purpose	Adjust Reading Rate Based on Purpose	Adjust Reading Rate Based on Purpose
	Read with Confidence from a Variety of Grade-Level Texts with Appropriate Speed and Accuracy	Read with Confidence from a Variety of Grade-Level Texts with Appropriate Expression	Read with Confidence from a Variety of Grade-Level Texts with Appropriate Expression
Read Silently for Increasing Periods of Time	Read Silently for Increasing Periods of Time		
			Ability to Read Grade-Level Texts Fluently

Our team then constructed passages to ensure that these standards were represented in the passages.

Vocabulary

As with the other subtests, standards that were common to at least two states were selected and put into an excel spreadsheet of standards per grade (3-8). The state standards were listed and then cross-referenced, identifying standards that appeared in more than one state. The wording of the standards was condensed to reduce wordiness and repetition. The most important standards as evidenced by the state adoptions were listed in concise terms.

1. We then examined the standards to determine which standards were measureable as follows:
    - Deleted "Derivations" standard because it falls under "Affixes" category
    - Combined "Root Words for Meaning" and "Common Roots" and labeled the standard "Roots."
    - Deleted the following standards because they were not measureable: "Use Thesaurus," "Use Dictionary/ Glossary," "Use Online Sources," "Homophones," "Use Experiences to Bring Meaning…," "Developing Vocabulary by Listening to Read Alouds," "Applying Words in Different Content Areas," "Using Context Clues," "Acquire New Vocabulary through Reading," "Direct Instruction of New Vocabulary," "Denotative and Connotative Meanings," and "Relating New Vocabulary to Familiar Words."
2. The standards spreadsheet was used as a resource when collecting vocabulary words for the Word Bank.

Creation of a Vocabulary Word Bank

In the process of creating the vocabulary word bank, two types of words were identified: (a) content vocabulary words—social studies, science, and math, and (b) general vocabulary words.

a. General vocabulary words were selected from nationally recognized sources based on the standards previously defined.

b. Since most nationally recognized sources of vocabulary words do not classify words based on content, other sources were used for generating a content-driven vocabulary word list. Microsoft Access was used to cross-reference the national list to determine content classification.

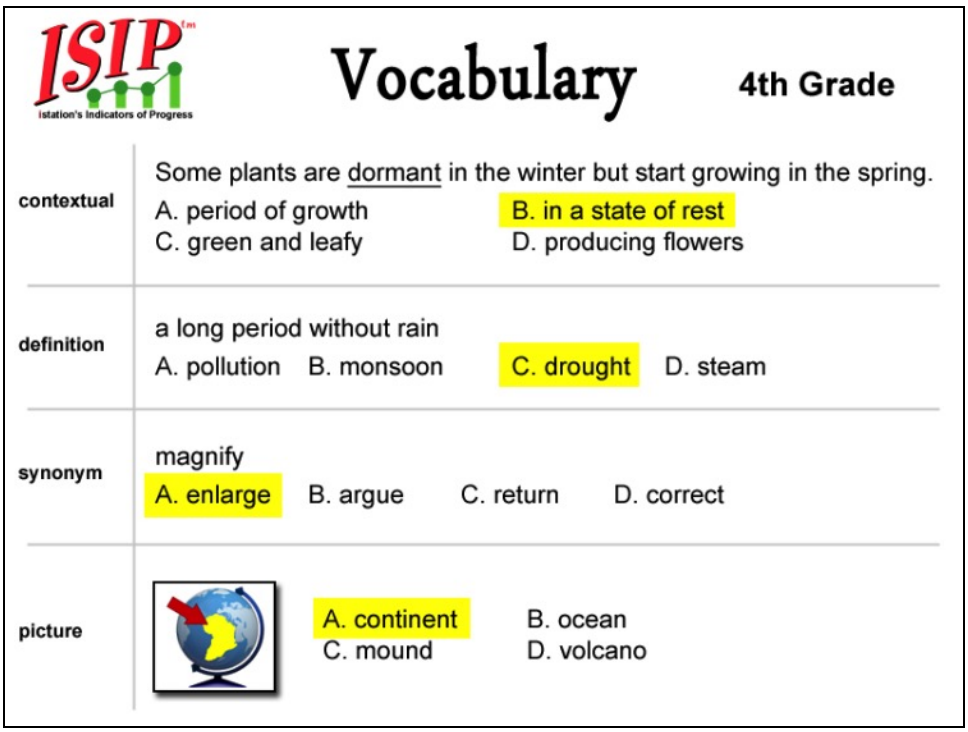
3. General and content vocabulary lists were combined into one Microsoft Access database.
4. Each word was assigned several criteria: grade level, root origin, affix, synonyms, antonyms, and definition.
    a. Root origin was used to assure our words meet the grade-level standards. According to the standards we previously defined, Latin and Greek roots are taught 5th-7th grades; and words from other languages in the 6th grade; Anglo Saxon origin words in 7th grade.
    b. Synonym and antonym pairs and definitions were used to generate test items.
    c. Affixes were used to generate distracters.

5. The word lists were then cross-referenced on each criterion to ensure that all words in the database (on all source-specific lists) had the appropriate information attached to them.
6. At the end of this process, the database included 14,000 words.
7. The process of narrowing down the list began with the exclusion of sources that were not nationally recognized.
8. The remaining 10,000-word list was further reduced based on consistency of grade-level assignment across sources.
9. Words deemed inappropriate by graduate students in the Literacy Acquisition Department were removed from the list.
10. Items were then selected for the final item bank so that all standards were represented.

Table 3-4: Vocabulary Standards

Vocabulary Standards					
3rd Grade	4th Grade	5th Grade	6th Grade	7th Grade	8th Grade
Synonyms	Synonyms	Synonyms	Synonyms		
Antonyms	Antonyms	Antonyms			
Roots	Roots	Roots	Roots	Roots	Roots
Affixes	Affixes	Affixes	Affixes	Affixes	Affixes
		Latin and Greek Roots	Latin and Greek Roots	Latin and Greek Roots	
				Anglo-Saxon Origins	
			Words from Other Languages		
Homographs	Homographs	Homographs	Homographs	Homographs	Homographs

Further, a search of literature was conducted to locate studies that focused on vocabulary assessment and possible confounds to the design of comprehension assessments. From this review we determined that a weakness of vocabulary assessment was in neglecting to assess students' abilities to infer vocabulary meaning from context. Our final blueprint included four item types: pictures, synonyms, definitions, and contextual as exemplified below.



Comprehension

Creation of the comprehension subtest blueprint possessed the greatest challenge for our team since there is little agreement in the field as to what comprises the construct of reading comprehension. As with the other subtest, our team consulted with the state standards. We then conducted the following steps:

1. A search of literature was conducted to locate studies that focused on comprehension assessment and possible confounds to the design of comprehension assessments. See abstracts and references.
    - Databases used include Google Scholar, PsycARTICLES®, PsycINFO®, and Academic Search Complete.
    - Search terms included reading comprehension, assessment, diagnostic, cloze procedure, MAZE task, skills, inference, analysis, and academic progress.
2. Internet sites were viewed to become familiar with other sources of information about comprehension and assessment. They included:
    - Northwest Evaluation Association, Measures of Academic Progress, at http://www.nwea.org/assessments/map.asp
    - Reading in America, Comprehension in Beginning Reading, at http://reading.uoregon.edu/scope/trial_scope_index.php

- The National Reading Panel, Comprehension, Reports of the Sub-groups, Comprehension, at http://www.nationalreadingpanel.org/
- The Center for Applied Linguistics, Project DARC, at http://www.cal.org/projects/darc.html

3. Meetings included discussions of the levels of higher-order thinking and the questioning strategies to be included in the assessment.
4. The feasibility of using innovative techniques, such as the use of graphic organizers, was explored. Options discussed were: story maps, concept maps, Venn Diagrams, timelines, and cause-effect charts. A literature review, Grades 2 – 8, was conducted to discover more information about the evidence base for using graphic organizers in instruction and assessment.
5. Science standards were reviewed and analyzed for CA, FL, NY, and TX, to determine grade-level content knowledge expectations for the creation of reading passages. Major science themes were isolated and cross-referenced by grade level. This indicated when students should have knowledge of these themes for assessment purposes.
6. Social studies standards were reviewed and analyzed for CA, FL, NY, and TX, to determine grade-level content knowledge expectations for the creation of reading passages. Major social studies themes were isolated and cross-referenced by grade level. This indicated when students should have knowledge of these themes for assessment purposes.

Standards that were identified in two or more states included:

Grade 4

- Summarize text content
- Analyze text to determine author's purpose
- Recognize plot, setting, characters and theme of a story
- Determine and infer main idea and supporting details
- Determine and explain cause-and-effect relationships
- Draw conclusions and make predictions based on the text
- Interpret graphic features
- Compare and contrast information in text
- Determine sequence of events
- Distinguish between fact and opinion.

Grade 5

- Distinguish and understand the elements of plot, setting, characterization, and problem resolution
- Determine main idea through summarizing and identifying relevant details
- Identify the purpose of different types of text such as to inform, influence, express, or entertain
- Determine how meaning in prose and poetry is affected by imagery, rhythm, flow, or figurative language, such as:
  - personification
  - metaphor
  - simile
  - hyperbole
- Interpret the author's use of dialogue and description
- Understand that theme refers to the implied or stated message about life and the world
- Make judgments and inferences about plot, setting, characters, and theme (implied or stated)
- Distinguish between fact and opinion in various texts

Grade 6

- Use information from text to answer questions related to explicitly stated main ideas or relevant details
- Interpret the author's use of dialogue, description, tone, purpose and perspective
- Draw conclusions from information gathered from multiple sources
- Decipher and analyze features of themes conveyed through characters, actions, and images
- Determine the main idea through inference
- Present a point of view or interpretation of a text and support it with relevant details from the text
- Identify cause-and-effect relationships in text
- Compare and contrast a variety of text structures
- Determine and describe elements of story structure, such as: setting, characterization, plot, and conflict.
- Establish and adjust purposes for reading

Grade 7

- Determine the theme in a selection and distinguish theme from topic
- Describe and connect the essential ideas, arguments, and perspectives of text
- Analyze characterization as evidenced through:
  - a character's thoughts, words, speech patterns, and actions
  - the narrator's description
  - the thoughts, words, and actions of other characters
- Relate a literary work to information about its setting or historical moment
- Use predicting, questioning, and summarizing as comprehension strategies
- Determine a text's main (or major) ideas and how those are supported in the text
- Draw inferences such as conclusion or generalizations, and support them with text evidence and experience
- Determine which events advance the plot and determine how each event explains past or present action(s) or foreshadows future action(s)
- Analyze how an author's use of words creates tone and mood, giving supporting evidence from text
- Compare and contrast traditional literature from different cultures

Grade 8

- Determine the difference between concepts of theme in a literary work and author's purpose in an expository text
- Describe and connect the essential ideas, arguments, and perspectives of text
- Identify and analyze recurring themes (e.g., good versus evil) across traditional and contemporary works
- Analyze a character's traits, emotions, or motivations and give supporting evidence from the text
- Identify literary devices that define a writer's style and use those to interpret the work
  - irony
  - symbolism
  - imagery
- Explain how an author's use of words creates and establishes tone and mood
- Determine a text's main (or major) ideas and how those are supported in the text

- Identify the purposes of different types of texts such as to inform, influence, express, or entertain
- Locate and analyze elements of plot including setting, conflict, and problem resolution

ISIP AR Theory of Comprehension

Unlike the other subtests, our team found that identification of the standards was less helpful in determining the exact nature of what comprises comprehension as measured on a computer administered test. Thus, our team spent considerable time creating an underlying theory of comprehension to drive item creation.

We determined that ISIP-AR Comprehension subtest should assess higher-level text comprehension, which is the ultimate goal of reading. Toward this end, our team developed 220 graduated testlets, consisting of either narrative or expository text and four types of multiple-choice questions. The question types chosen to assess higher-level text comprehension include main idea, cause/effect or problem/outcome, inference, and critical judgment of the text. Students choose from four possible answers for each question.

The comprehension subtest has been conceived and developed based on several key pieces of research. They include the following:

- Text comprehension difficulties are found in 5-15% of children, even though they may not display lower level or surface processing deficits, i.e., decoding, word recognition, fluency, and/or language comprehension (Cain & Oakhill, 2007; Fletcher et al., 2007; Nation, 1999; Nation et al., 1999; Yuill & Oakhill, 1991). Understanding how children comprehend text at higher cognitive levels is necessary for advancement and intervention.

- Higher level processing of text is defined as the reader's ability to determine the overall gist of the passage, differentiate and switch between broader and narrower concepts (gist versus details), inhibit irrelevant information from intruding upon meaning, monitor comprehension, reason, make inferences, and integrate information into long-term memory (Gamino & Chapman, in press; Kintsch, 1998; Oakhill, Hartt, & Samols, 2005; Sesma et al., 2009; Williams, 2003; Yuill & Oakhill, 1991). Cain & Oakhill (2007) refer to this ability to comprehend as being able "to derive an overall interpretation of the state of affairs described, rather than to simply retrieve the meaning of individual words and sentences." Johnson-Laird (1983) calls it "a mental model," while Kintsch (1998) refers to it as "the situation model." All questions have been developed to meet the criteria of higher-level processing.

- Questions have been constructed to assess the students' ability to make higher-level links between individual sentences to establish local coherence (i.e., cause/effect and inference question types) and to integrate new information into existing representations to establish global coherence of text (i.e., main idea, problem/outcome, and critical judgment question types) (Cain & Oakhill, 1999; Cain, Oakhill, Barnes, & Bryant, 2001; Oakhill, 1982; Wixson & Peters, 1987). Additionally, Kintsch

(1998) refers to this coherence as knowledge generation, or the ability to derive new information from propositions in the text by some inference or critical judgment procedure.

- All questions are dependent upon information in the passage in order to avoid the testing of background knowledge and having questions that can be answered without reading the text. This situation has been a pitfall of other well-known tests (Keenan & Betjemann, 2006). All answer choices (i.e., correct answer, two distractors, and wrong answer) relate to the passage in some form.

- Because proficient memory has been associated with reading ability and skilled text comprehension (Cain, 2006; Daneman & Merikle, 1996; Sesma et al., 2009; Swanson, Howard, & Saez, 2007), the text will not be available to students when they are answering questions. However, specific details that do not add to an understanding of the general or global coherence of the passage will not be questioned.

- In the subtest, all types of story structures have been included since comprehension failure has been linked to inadequate knowledge about how texts are structured (Perfetti, 1994). Understanding children's deficiencies in different types of story structures will help when intervening.

In theory, then, comprehension requires both low level and high level processing of text information. It is in the higher level processing that the deeper message of the text comes forth. The subtest has been developed to specifically address higher cognitive level comprehension with the goal of constructing questions that are both conceptually and instructionally valid.

With our theory of how to assess comprehension in hand, our team devised parameters for testlet construction representative of each grade.

Table 3-5: Parameters for ISIP AR Comprehension Testlets

Fourth Grade	
Minimum WPM (per DIBELS)	< 110 WPM for Low Risk
Word Count	150 < 250
Text Genre	• 50% Narrative • 50% Expository

Suggested Text Structures	• Narrative story • Personal narrative • Fable • Tall tale • Historical fiction • Explanatory / Factual story or report • Descriptive • Compare / Contrast • Sequence of events / Time order • Define / Explain
Sentence Structures	• Simple sentence (including declarative, interrogative, imperative, and exclamatory sentences) • Compound sentence (using simple conjunctions such as: and, but, or, because) • Punctuation (period, question mark, exclamation point, comma, apostrophe) • Simple appositives (e.g., Mr. Miller, our coach, will not be at practice today.)
Word Structures	• Prefixes (i.e., re-, un-, under-, over-, -dis-, non-, pre-, in-, bi-, tri-, quad-, oct-, etc.) • Suffixes (i.e., -ly, -er, -est, -y, -ion, -ation, -sion, -ible, -ness, -less, -or, -ful, etc.) • Compound words • Contractions (i.e., I'm, I've, won't, we'll, don't, you're, I'll, they're, etc.) • Possessives (e.g., Mr. Miller's tie, the child's toy, the cats' food dishes, etc.)
Word Analysis	• Figurative language – similes • Homophones (e.g., there, their, they're; to, two, too)
Question Types	• Main idea – 75% explicit; 25% implied • Cause/effect or problem/outcome • Inference • Critical judgment (i.e., determining author's point of view, fact and opinion, drawing conclusions)

Fifth Grade	
Minimum WPM (per DIBELS)	< 118 WPM for Low Risk
Word Count	175 < 275
Text Genre	• 50% Narrative • 50% Expository
Suggested Text Structures	• Narrative story • Personal narrative • Fable • Tall tale • Folktale • Historical fiction • Explanatory / Factual story or report • Descriptive • Compare / Contrast • Sequence of events / Time order • Define / Explain • Biographical sketch
Sentence Structures	• Simple sentence (including declarative, interrogative, imperative, and exclamatory sentences) • Compound sentence (using two independent clauses) • Simple complex sentences • Punctuation (period, question mark, exclamation point, comma, apostrophe) • Appositives in subjective case

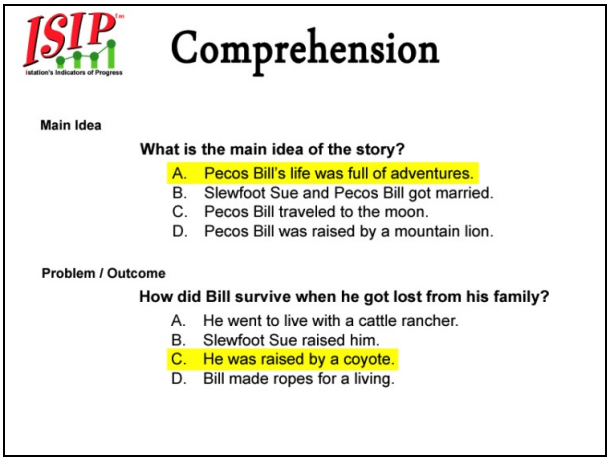
Word Structures	• Prefixes (i.e., in addition to 4th grade, semi-, super-, multi-, poly-, tele-, in-, il-, im-, ir-, mis-, inter-, mid-, sub-, deci-, deca-, di-, dia-, kilo-, milli-, centi, etc.) • Suffixes (i.e., in addition to 4th grade, -ian, -an, -ment, -en, -dom, -ship, -ness, -ist, -ess, etc.) • Latin and Greek roots (i.e., max(i), meter/metr, photo, scope, port, tract, form, etc.) • Compound words • Contractions • Possessives
Word Analysis	• Figurative language – similes, metaphors, personification • Homophones (e.g. council, counsel, etc.)
Question Types	• Main idea – 50% explicit; 50% implied • Cause/effect or problem/outcome • Inference • Critical judgment (i.e., determining author's point of view, fact and opinion, drawing conclusions, theme, motivation)
Sixth Grade	
Minimum WPM (per DIBELS)	< 124 WPM for Low Risk
Word Count	200 < 325
Text Genre	• 50% Narrative • 50% Expository
Suggested Text Structures	• Narrative story • Myths and legends • Historical fiction • Explanatory / Factual story or report • Descriptive • Compare / Contrast • Persuasive • Biographical sketch

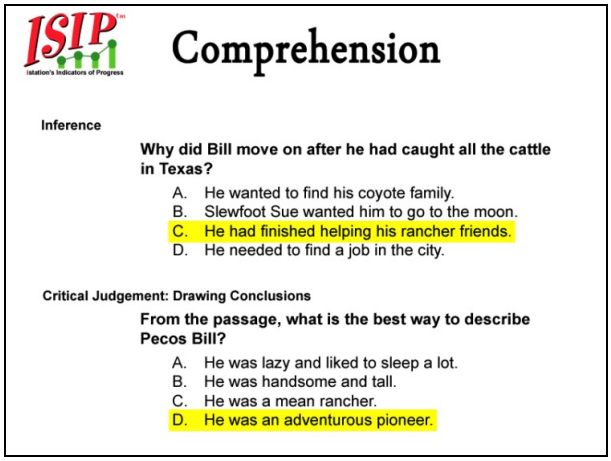
Sentence Structure	• Variety of simple sentences • Compound sentences (using two independent clauses) • Complex sentences (using dependent clauses and phrases with independent clauses) • Punctuation (period, question mark, exclamation point, comma, apostrophe) • Appositives in subjective or objective case
Word Structures	• Prefixes (i.e., in addition to 4th – 5th grade, en-, em-, fore-, de-, trans-, anti-, ex-, auto-, bio-, mini-, micro-, uni-, etc.) • Suffixes (i.e., in addition to 4th – 5th grade, -ity, -al, -ial, -ion, -ation, -sion, -tion, -ish, -ant, -ent, -hood, -logy, -ology, etc.) • Latin and Greek Roots (i.e., aqua, act, mit, anni/annu/enni, arch, duc/duct, gram/graph, geo, man, nym/onym, phon, rupt, scrib/script, tox, therm, etc.) • Combining forms (e.g., microscope, biology, etc.)
Word Analysis	• Figurative language – similes, metaphors, personification, idioms, etc.
Questions	• Main idea – 25% explicit; 75% implied • Cause/effect or problem/outcome • Inference • Critical judgment (i.e., determining author's point of view and bias, fact and opinion, drawing conclusions, theme, motivation, theme)
Seventh Grade	
Suggested Minimum WPM	< 140 WPM for Low Risk
Word Count	225 < 350
Text Genre	• 25% Narrative • 75% Expository
Suggested Text Structures	• Narrative story • Explanatory / Factual story or report • Classification • Compare / Contrast • Persuasive • Biographical sketch

Sentence Structures	• Variety of simple sentences • Compound sentences (using two independent clauses) • Complex sentences (using dependent clauses and phrases with independent clauses - and using participial phrases – gerunds, infinitives, etc.) • Punctuation (period, question mark, exclamation point, comma, apostrophe, hyphens, semicolons, colons) • Appositives in subjective or objective case
Word Structures	• Prefixes (i.e., in addition to 4th – 6th grade, anti-, ab-, a-, co-, con-, com-, pro-, intra-, mega-, post-, chrono-, etc.) • Suffixes (i.e., in addition to 4th – 6th grade, -ous, -ious, -eous, -ive, -ative, -itive, -ence, -ance, -ic, -ize, -fy, -ify, -age, -some, etc.) • Latin and Greek roots (i.e., chron, temp, aer/aero, cede/ceed, cept/ceive, dict, fract/frag, gen, grat, ject, liber, leg/lect/lig, mater/matr/matri, pater/patr, mot/mob, opt, ped/pod, spect/spec, urb, pop, pend) • Combining forms (e.g., chronology, maternity, etc.)
Word Analysis	• Figurative language – similes, metaphors, personification, idioms, etc.
Questions types	• Main idea – implicit • Cause/effect or problem/outcome • Inference • Critical judgment (i.e., determining author's point of view and bias, drawing conclusions, theme, motivation, theme, symbolism, tone, mood)
Eighth Grade	
Suggested Minimum WPM	< 150 WPM for Low Risk
Word Count	250 < 400
Text Genre	• 25% Narrative • 75% Expository

Suggested Text Structures	• Narrative story • Explanatory / Factual story or report • Classification • Persuasive • Argumentative • Biographical sketch
Sentence Structures	• Variety of simple sentences • Compound sentences (using two independent clauses) • Complex sentences (using dependent clauses and phrases with independent clauses and using participial phrases – gerunds, infinitives, etc.) • Compound-complex sentences (compound forms with dependent clauses and phrases) • Punctuation (period, question mark, exclamation point, comma, apostrophe, hyphens, semicolons, colons) • Appositives in subjective or objective case
Word Structures	• Prefixes (i.e., in addition to 4th – 7th grade, hyper-, hypo-, hyp-, omni-, homo-, hetero-, ultra-, etc.) • Suffixes (i.e., in addition to 4th – 7th grade, -cide, -ery, -ary, -ism, -ium, -tude, etc.) • Latin and Greek roots (i.e., aud, cred, archae/archi, belli, claim/clam, crat/cracy, hemo/hema, luna, mar, mort, path, pel, struc/struct, vis/vid, voc/voke, cogn, loc/loqu) • Combining forms (e.g., pedestrian, credible, etc.)
Word Analysis	• Figurative language – similes, metaphors, personification, idioms, etc.
Questions Types	• Main idea – implicit • Cause/effect or problem/outcome • Inference • Critical judgment (i.e., determining author's point of view and bias, drawing conclusions, theme, motivation, theme, symbolism, tone, mood)

Example Item Types





Concurrent Validity

Concurrent validity evidence was established by computing Pearson Product Moment correlation coefficients between ISIP Advanced Reading subtests and norm referenced external measures with established psychometric properties including: *Gray Oral Reading Test-4* (GORT – 4), *Woodcock-Johnson-3* (WJ-III), *Wechsler Individual Achievement Test-II* (WIAT-II) and the *Peabody Picture Vocabulary Test-IV* (PPVT-IV). Both Fluency and Comprehension subtest of the GORT-4 were administered. The Spelling, Reading Fluency, Vocabulary (pictures and synonyms) subtests of the WJ-III were administered. Spelling, Word Recognition, Pseudoword Decoding subtests were administered from the WIAT-II.

The WIAT-II was standardized using a total sample of 5,586 individuals, with two standardization samples drawn for Pre-K to 12 (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 more than 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 aged 6-18 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).

Sample

To establish concurrent validity evidence, data were collected during the 2010-11 school year from two large north Texas independent school districts. These districts were different from the districts used in the IRT calibration study. Demographics of the study participants are found in Table 3-6

Table 3-6: Student Demographics

	Grade Level											
	4th		5th		6th		7th		8th		All	
	n	%	n	%	n	%	n	%	n	%	n	%
Students	115	20.2	123	21.7	138	24.3	106	18.7	86	15.1	568	100
By School												
A	36	31.3	35	28.5	32	23.2					103	18.1
B	34	30	26	21.1	39	28.3					99	17.4
C	25	21.7	28	22.8	19	13.8					72	12.7
D	20	17.4	34	27.6	24	17.4					78	13.7
E					12	8.7	20	18.9	22	25.6	54	9.5
F							21	19.8	15	17.4	36	6.3
G					12	8.7	24	22.6	25	29.1	61	10.7
H							41	38.7	24	27.9	65	11.4
Gender												
Male	64	55.7	62	50.4	77	55.8	51	48.1	36	41.9	290	51.1
Female	51	44.3	61	49.6	61	44.2	55	51.9	50	58.1	278	48.9
Ethnicity												
White	51	44.3	57	46.3	59	42.8	9	8.5	7	8.1	183	32.2
African American	25	21.7	29	23.6	19	13.8	20	18.9	15	17.4	108	19
Hispanic/Latino	25	21.7	25	20.3	50	36.2	75	70.8	61	70.9	236	41.5
American Indian	0	0	2	1.6	3	2.2	0	0	0	0	5	0.8
Asian	11	9.6	9	7.3	5	3.6	1	0.9	2	2.3	28	4.9
Pacific Islander	1	0.9	0	0	2	1.4	0	0	0	0	3	0.5
Other	2	1.7	1	0.8	0	0	1	0.9	1	1.2	5	0.8
Economically Disadvantaged												
Yes	51	44.3	59	48	68	49.3	83	78.3	75	87.2	336	59.2

	Grade Level											
	4th		5th		6th		7th		8th		All	
No	64	55.7	64	52	70	50.7	23	21.7	11	12.8	232	40.8
Limited English Proficiency												
Yes	11	9.6	5	4.1	19	13.8	20	18.9	64	74.4	119	21
No	104	90.4	118	95.9	119	86.2	86	81.1	22	25.6	449	79
Receiving Special Education Services												
Yes	9	7.8	1	0.8	12	8.7	4	3.8	4	4.7	30	5.3
No	106	92.2	122	99.2	126	91.3	102	96.2	82	95.3	538	94.7
Qualifying Special Education Disability**												
Autism	1	--	--	--	--	--		--		--		--
Emotional Disturbance	1	--	--	--	--	--		--		--		--
Learning Disability	1	--	--	--	7	--		--		--		--
Other Health Impairment	1	--	1	--	2	--		--		--		--
Orthopedic Impairment	--	--	--	--	1	--		--		--		--
Speech Impairment	5	--	--	--	--	--		--		--		--

**No data for 6th, 7th, and 8th graders from schools E, F, G, & H

Research Design

Students were administered the test battery across four sessions, with each session lasting approximately 40 to 45 minutes. We grouped these tests into A, B, C, and D administrations and counterbalanced by session of administration to ensure that any administration order effects were washed out in the end.

Groupings

A. GORT-4

B. WJ-III
   Spelling, Reading Fluency, Vocabulary (pictures and synonyms)

C. WIAT-II
   Spelling, Word Recognition, Pseudoword Decoding

D. ISIP AR
   PPVT-IV, WIAT-II Reading Comprehension

Counterbalancing of tests by sessions is presented below.

Group	Session 1	Session 2	Sessions 3	Session 4
A	A	B-1/2	C	D-1/2
B	B-1/2	C	D-1/2	A
C	C	D-1/2	A	B-1/2
D	D-1/2	A	B-1/2	C

Results

The Pearson Product Moment correlations for ISIP-AR and the External measures are presented in Tables 3-7 through 3-11

Table 3-7: Correlations between External Measures and ISIP Advanced Reading Subtest Scores for Grade 4

	GORT4 FL	GORT4 CMP	PPVT4	WIAT CMP	WIAT WR	WIAT SPL	WIAT PSD	WJIII SPL	WJIII PV	WJIII RVS	ISIP AR OVR	ISIP AR CMP	ISIP AR SPL	ISIP AR VOC	ISIP AR TF
GORT4 FL	1.000[a]	0.392[a]	0.381[a]	0.476[a]	0.716[a]	0.632[a]	0.609[a]	0.655[a]	0.128[a]	0.630[a]	0.707[a]	0.640[b]	0.646[a]	0.471[a]	0.601[d]
GORT4 CMP		1.000[a]	0.515[a]	0.443[a]	0.286[a]	0.175[a]	0.159[a]	0.294[a]	0.412[a]	0.419[a]	0.420[a]	0.471[b]	0.164[a]	0.421[a]	0.408[d]
PPVT4			1.000[a]	0.491[a]	0.502[a]	0.397[a]	0.314[a]	0.403[a]	0.593[a]	0.606[a]	0.567[a]	0.474[b]	0.401[a]	0.623[a]	0.405[d]
WIAT CMP				1.000[a]	0.426[a]	0.418[a]	0.343[a]	0.533[a]	0.279[a]	0.503[a]	0.614[a]	0.540[b]	0.531[a]	0.468[a]	0.480[d]
WIAT WR					1.000[a]	0.718[a]	0.770[a]	0.725[a]	0.274[a]	0.672[a]	0.673[a]	0.515[b]	0.727[a]	0.493[a]	0.508[d]
WIAT SPL						1.000[a]	0.714[a]	0.868[a]	0.165[a]	0.485[a]	0.708[a]	0.497[b]	0.811[a]	0.487[a]	0.582[d]
WIAT PSD							1.000[a]	0.685[a]	0.085[a]	0.498[a]	0.599[a]	0.440[b]	0.682[a]	0.361[a]	0.455[d]
WJIII SPL								1.000[a]	0.112[a]	0.557[a]	0.748[a]	0.543[b]	0.832[a]	0.504[a]	0.564[d]
WJIII PV									1.000[a]	0.288[a]	0.252[a]	0.237[b]	0.151[a]	0.356[a]	0.119[d]
WJIII RVS										1.000[a]	0.655[a]	0.572[b]	0.546[a]	0.584[a]	0.456[d]
ISIP AR OVR											1.000[a]	0.907[b]	0.784[a]	0.762[a]	0.665[d]
ISIP AR CMP												1.000[b]	0.547[b]	0.603[b]	0.653[d]
ISIP AR SPL													1.000[a]	0.435[a]	0.532[d]
ISIP AR VOC														1.000[a]	0.393[d]
ISIP AR TF															1.000[d]

[a]N = 115 [b]N= 114 [c]N= 113 [d]N = 112

Table 3-8: Correlations between External Measures and ISIP Advanced Reading Subtest Scores for Grade 5

	GORT4 FL	GORT4 CMP	PPVT4	WIAT CMP	WIAT WR	WIAT SPL	WIAT PSD	WJIII SPL	WJIII PV	WJIII RVS	ISIP AR OVR2	ISIP AR CMP2	ISIP AR SPL2	ISIP AR VOC2	ISIP AR TF2
GORT4 FL	1.000[a]	0.365[a]	0.332[a]	0.484[a]	0.669[a]	0.557[a]	0.566[a]	0.628[a]	0.314[a]	0.406[a]	0.615[a]	0.519[a]	0.607[a]	0.429[a]	0.631[b]
GORT4 CMP		1.000[a]	0.556[a]	0.444[a]	0.284[a]	0.157[a]	0.112[a]	0.187[a]	0.465[a]	0.328[a]	0.377[a]	0.412[a]	0.168[a]	0.432[a]	0.386[b]
PPVT4			1.000[a]	0.530[a]	0.370[a]	0.280[a]	0.179[a]	0.312[a]	0.618[a]	0.518[a]	0.562[a]	0.502[a]	0.315[a]	0.693[a]	0.365[b]
WIAT CMP				1.000[a]	0.477[a]	0.417[a]	0.274[a]	0.488[a]	0.503[a]	0.425[a]	0.581[a]	0.538[a]	0.449[a]	0.587[a]	0.460[b]
WIAT WR					1.000[a]	0.749[a]	0.709[a]	0.739[a]	0.344[a]	0.443[a]	0.714[a]	0.544[a]	0.753[a]	0.516[a]	0.516[b]
WIAT SPL						1.000[a]	0.638[a]	0.892[a]	0.290[a]	0.402[a]	0.669[a]	0.397[a]	0.835[a]	0.460[a]	0.553[b]
WIAT PSD							1.000[a]	0.637[a]	0.133[a]	0.288[a]	0.531[a]	0.377[a]	0.636[a]	0.305[a]	0.399[b]
WJIII SPL								1.000[a]	0.294[a]	0.449[a]	0.757a	0.494[a]	0.867[a]	0.459[a]	0.594[b]
WJIII PV									1.000[a]	0.363[a]	0.496[a]	0.433[a]	0.323[a]	0.577[a]	0.302[b]
WJIII RVS										1.000[a]	0.481[a]	0.373[a]	0.416[a]	0.518[a]	0.363[b]
ISIP AR OVR2											1.000[a]	0.858[a]	0.808[a]	0.779[a]	0.604[b]
ISIP AR CMP2												1.000[a]	0.472[a]	0.580[a]	0.533[b]
ISIP AR SPL2													1.000[a]	0.512[a]	0.561[b]
ISIP AR VOC2														1.000[a]	0.417[b]
ISIP AR TF2															1.000[b]

[a]N = 123 [b]N= 120

Table 3-9: Correlations between External Measures and ISIP Advanced Reading Subtest Scores for Grade 6

	GORT 4 FL	GORT4 CMP	PPVT 4	WIAT CMP	WIAT WR	WIAT SPL	WIAT PSD	WJIII SPL	WJIII PV	WJIII RVS	ISIP AR OVR	ISIP AR CMP	ISIP AR SPL	ISIP AR VOC	ISIP AR TF
GORT4 FL	1.000[a]	0.292[a]	0.363[a]	0.408[a]	0.739[a]	0.687[a]	0.646[a]	0.720[a]	0.233[a]	0.510[a]	0.659[a]	0.573[b]	0.673[a]	0.451[a]	0.547[c]
GORT4 CMP		1.000[a]	0.534[a]	0.413[a]	0.299[a]	0.262[a]	0.162[a]	0.232[a]	0.453[a]	0.584[a]	0.353[a]	0.395[b]	0.150[a]	0.445[a]	0.210[c]
PPVT4			1.000[a]	0.564[a]	0.436[a]	0.339[a]	0.243[a]	0.285[a]	0.772[a]	0.643[a]	0.552[a]	0.527[b]	0.266[a]	0.664[a]	0.323[c]
WIAT CMP				1.000[a]	0.446[a]	0.423[a]	0.339[a]	0.387[a]	0.427[a]	0.601[a]	0.535[a]	0.565[b]	0.341[a]	0.579[a]	0.429[c]
WIAT WR					1.000[a]	0.766[a]	0.767[a]	0.788[a]	0.351[a]	0.622[a]	0.703[a]	0.543[b]	0.734[a]	0.541[a]	0.498[c]
WIAT SPL						1.000[a]	0.745[a]	0.893[a]	0.249[a]	0.522[a]	0.746[a]	0.497[b]	0.825[a]	0.560[a]	0.479[c]
WIAT PSD							1.000[a]	0.784[a]	0.193[a]	0.505[a]	0.593[a]	0.440[b]	0.713[a]	0.339[a]	0.413[c]
WJIII SPL								1.000[a]	0.186[a]	0.511[a]	0.683[a]	0.416[b]	0.849[a]	0.465[a]	0.503[c]
WJIII PV									1.000[a]	0.593[a]	0.474[a]	0.451[b]	0.204[a]	0.578[a]	0.228[c]
WJIII RVS										1.000[a]	0.615[a]	0.565[b]	0.446[a]	0.629[a]	0.378[c]
ISIP AR OVR											1.000[a]	0.855[b]	0.818[a]	0.456[a]	0.584[c]
ISIP AR CMP												1.000[b]	0.507[b]	0.456[b]	0.541[c]
ISIP AR SPL													1.000[a]	0.456[a]	0.494[c]
ISIP AR VOC														0.456[a]	0.453[c]
ISIP AR TF															1.000[c]

Table 3-10: Correlations between External Measures and ISIP Advanced Reading Subtest Scores for Grade 7 (n = 106)

	GORT4 FL	GORT4 CMP	PPVT4	WIAT CMP	WIAT WR	WIAT SPL	WIAT PSD	WJIII SPL	WJIII PV	WJIII RVS	ISIP AR OVR2	ISIP AR CMP2	ISIP AR SPL2	ISIP AR VOC2	ISIP AR TF2
GORT4 FL	1.000	0.209	0.296	0.440	0.618	0.574	0.408	0.645	0.344	0.326	0.659	0.533	0.690	0.492	0.598
GORT4 CMP		1.000	0.461	0.464	0.234	0.201	0.002	0.108	0.400	0.423	0.432	0.345	0.161	0.585	0.243
PPVT4			1.000	0.476	0.420	0.406	0.080	0.423	0.741	0.771	0.481	0.387	0.298	0.520	0.331
WIAT CMP				1.000	0.384	0.226	0.163	0.278	0.509	0.519	0.596	0.510	0.346	0.706	0.443
WIAT WR					1.000	0.669	0.694	0.707	0.340	0.424	0.487	0.317	0.603	0.401	0.334
WIAT SPL						1.000	0.562	0.866	0.396	0.429	0.509	0.322	0.706	0.331	0.432
WIAT PSD							1.000	0.622	0.000	0.151	0.263	0.080	0.523	0.189	0.149
WJIII SPL								1.000	0.358	0.453	0.541	0.345	0.729	0.385	0.412
WJIII PV									1.000	0.742	0.490	0.342	0.344	0.563	0.403
WJIII RVS										1.000	0.553	0.430	0.373	0.606	0.396
ISIP AR OVR											1.000	0.879	0.734	0.830	0.728
ISIP AR CMP												1.000	0.451	0.622	0.651
ISIP AR SPL													1.000	0.479	0.549
ISIP AR VOC														1.000	0.548
ISIP AR TF															1.000

Table 3-11: Correlations between External Measures and ISIP Advanced Reading Subtest Scores for Grade 8

	GORT4 FL[a]	GORT4 CMP[a]	PPVT4[a]	WIAT CMP[a]	WIAT WR[a]	WIAT SPL[a]	WIAT PSD[a]	WJIII SPL[a]	WJIII PV[a]	WJIII RVS[a]	ISIP AR OVR[a]	ISIP AR CMP[a]	ISIP AR SPL[a]	ISIP AR VOC[a]	ISIP AR TF[a]
GORT4 FL	1.000	0.234	0.411	0.327	0.646	0.600	0.547	0.649	0.289	0.375	0.595	0.424	0.678	0.462	0.575
GORT4 CMP		1.000	0.411	0.254	0.368	0.199	0.123	0.222	0.283	0.271	0.368	0.375	0.229	0.355	0.373
PPVT4			1.000	0.435	0.667	0.469	0.344	0.531	0.773	0.680	0.542	0.426	0.447	0.610	0.553
WIAT CMP				1.000	0.381	0.321	0.157	0.363	0.499	0.494	0.463	0.482	0.332	0.482	0.490
WIAT WR					1.000	0.763	0.688	0.776	0.558	0.556	0.695	0.529	0.734	0.558	0.564
WIAT SPL						1.000	0.556	0.844	0.423	0.438	0.640	0.428	0.761	0.472	0.483
WIAT PSD							1.000	0.546	0.240	0.297	0.447	0.300	0.574	0.308	0.399
WJIII SPL								1.000	0.427	0.459	0.681	0.448	0.811	0.522	0.463
WJIII PV									1.000	0.646	0.422	0.303	0.405	0.487	0.529
WJIII RVS										1.000	0.530	0.405	0.471	0.596	0.543
ISIP AR OVR											1.000	0.859	0.837	0.873	0.677
ISIP AR CMP												1.000	0.542	0.724	0.674
ISIP AR SPL													1.000	0.583	0.554
ISIP AR VOC														1.000	0.605
ISIP AR TF															1.000

[a]N = 86 [b]N = 78

Discussion

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 as in the overall reading ability scores. In addition, ISIP Advanced Reading produced stable scores over time. These outstanding results could stem from a number of converging reasons. First, the authorship team took great care in constructing the ISIP Advanced Reading item pool. They utilized the most up-to-date findings in reading research as a basis for the item types and content they represent. Also, ISIP Advanced Reading is an engaging and adaptive computer-based assessment program. Items are presented to students at their ability using high-quality computer animation. Students feel they are "playing a game" rather than "taking another test," which likely results in less off-task behavior during assessment, producing more consistent results.

In considering the concurrent validity correlations between ISIP Advanced Reading and the external measures, it is important to keep in mind that the nature of each of these measures was somewhat different, and thus near-perfect correlations were not expected. With the exception of spelling, all ISIP Advanced Reading items required students to select a correct answer from among choices on the computer screen; whereas all of the external subtests required students provide one answer verbally to an examiner. 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.

Spelling was most similar in administration across all three tests (ISIP Advanced Reading, WJ-II, and WIAT-II), and not surprisingly, the highest correlations across grades were observed for these three measures to each other. ISIP Advanced Reading spelling also correlated well with measures related to word analysis including WIAT-II Word Recognition, and Pseduoword reading, strengthening our theory that Spelling is really a measure word analysis.

The ISIP Advanced Reading Text Fluency measure and the GORT-4 fluency measures had large correlations across grades. Because these two measures measure somewhat different aspects of fluency, we did not expect near perfect correlations. The GORT-4 fluency measures are strictly an oral reading speed measure of text and is not linked to specific grade level expectations. The ISIP Advanced Reading Text fluency measure is a measure of silent reading fluency that incorporates comprehension monitoring. Importantly, correlations for ISIP Advanced Reading Text Fluency were more highly correlated to our measures of comprehension than the GORT-4 Fluency, giving credibility to our theory that the MAZE task incorporated with ISIP Advanced Reading text fluency taps into both speed of text processing and comprehension monitoring. The underlying theory of fluency for ISIP Advanced Reading is that a fluent reader reads both with speed and processing of meaning. The ISIP Advanced Reading appears to be measuring fluency in accordance with this theory.

Similarly, the ISIP Advanced Reading measure of vocabulary correlated reasonably well to the external measures of vocabulary, and about as well as the external measures of vocabulary correlated to each other. Given that the way in which ISIP Advanced Reading approached the assessment of vocabulary was different than the external measures, the correlations observed are in line with what we expected. It is important to remember that ISIP Advanced Reading incorporated four types of items, none of which were similar in format to the item types found on the PPVT-4, the WJIII-Picture vocabulary or the WJIII Synonyms subtests. Picture items on ISIP Advanced Reading required students to identify the word for four choices best illustrating a word (for example ligament). This was most similar to the WJIII Picture vocabulary, except the WJIII subtest provided no choices. ISIP Advanced Reading Synonym items were quite dissimilar from WJIII Synonym items in that ISIP Advanced Reading required students to choice the best synonym to the target word from among four choices and required no decoding ability. The WJIII Synonyms were confounded (in our opinion) by the ability to decode the target, to produce a synonym. None of the external measures incorporated definition items, or anything remotely similar to the ISIP Advanced Reading contextual items. Because our theory of vocabulary incorporates that ability to infer new vocabulary from context, we include these items and believe that perhaps ISIP Advanced Reading represents a more sophisticated measure of vocabulary measures that will not be expected to correlate extremely highly to measures of vocabulary that measure only less sophisticated vocabulary knowledge.

The ISIP Advanced Reading comprehension and the underlying theory of comprehension is very different from the external measures of comprehension. First, The ISIP Advanced Reading assessment of comprehension is designed to reduce the influence of decoding on comprehension. We achieved this by selecting the passages students read according to their theta score grade placement for spelling. Thus, text is matched to reader in terms of decoding, allowing a better estimation of actual ability to comprehend text without the possible confound of inadequate decoding impeding comprehension. Neither the GORT-4 nor WIAT-II has a mechanism to control for inadequate decoding leaving those measures confounded with decoding. We chose not to compare ISIP Advanced Reading to the popular WJ-III Passage Comprehension because it is well-known that this particular measure is highly confounded with decoding ability. Further, in constructing the ISIP Advanced Reading items, care was taken to ensure that being able to answer the questions correctly were dependent on actual reading of the passage. The GORT-4 is known to suffer from passage independence. Thus, we were not surprised that ISIP Advanced Reading and GORT-4 correlations were not strong, but neither were the correlations between the GORT-4 comprehension measure and the WIAT-II comprehension measure. The ISIP Advanced Reading correlations with WIAT-II Comprehension were more positively correlated than those of the GORT-4. We did not expect these measures to be extremely correlated because the testlet nature of the ISIP Advanced Reading compared to the one-item nature of the WIAT-II. Likewise, the underlying theory of comprehension between the two tests was fairly different.

Conclusion

Evidence of concurrent validity, can be found in the numerous strong, positive relationships to external measures of reading constructs. Further evidence of the validity of ISIP Advanced Reading is that subtest scores correlated more highly to external measures designed to tap into the same underlying construct and not as well to external measures measuring different underlying constructs. Further, ISIP Advanced Reading measures generally correlate to external measures similarly as the external measures correlate to each other. Taken together, the evidence supports the claim that ISIP Advanced Reading produces reliable and valid data for measuring key areas of reading for students in the middle grades (i.e. Grades 4-8).

Chapter 4: 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 Advanced Reading test, we are interested in comparing students to a national sample of students who have taken the ISIP Advanced 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 Advanced Reading users during the 2014-2015 school year in grades 4-8. The state distributions for the sample are found in Table 4-1.

Table 4-1: State Distributions & Demographics For ISIP Advanced Reading Norming Sample

	Grade				
	4th	5th	6th	7th	8th
	Frequency (%)	Frequency (%)	Frequency (%)	Frequency (%)	Frequency (%)
Gender					
Female	55,219 (34.3)	48,462 (34.7)	22,033 (37.7)	16,755 (37.6)	13,998 (37.6)
Male	58,200 (36.1)	51,229 (36.7)	23,321 (39.9)	18,197 (40.9)	15,421 (41.5)
Special Education					
No	69,954 (43.4)	59,446 (42.6)	27,772 (47.5)	21,387 (48.1)	18,593 (50.0)
Yes	7,203 (4.5)	6,481 (4.6)	3,180 (5.4)	2,669 (6.0)	2,164 (5.8)
State					
Alabama	1,271 (0.8)	1,147 (0.8)	799 (1.4)	734 (1.6)	1,013 (2.7)
Arizona	45 (0.0)	11 (0.0)	2 (0.0)	17 (0.0)	6 (0.0)
California	1,185 (0.7)	1,083 (0.8)	326 (0.6)	28 (0.1)	24 (0.1)
Colorado	273 (0.2)	171 (0.1)	53 (0.1)	4 (0.0)	7 (0.0)
Connecticut	1 (0.0)	1 (0.0)	1 (0.0)	-	-
District of Columbia	-	-	-	-	-
Florida	4,787 (3.0)	3,184 (2.3)	276 (0.5)	-	-
Georgia	1,415 (0.9)	892 (0.6)	128 (0.2)	151 (0.3)	101 (0.3)
Illinois	455 (0.3)	451 (0.3)	344 (0.6)	380 (0.9)	205 (0.6)
Indiana	182 (0.1)	171 (0.0)	55 (0.1)	2 (0.0)	
Maryland	152 (0.1)	55 (0.0)	4 (0.0)	1 (0.0)	3 (0.0)
Montana	781 (0.5)	749 (0.5)	699 (1.2)	1,038 (2.3)	1,022 (2.7)
Massachusetts	-	4 (0.0)	4 (0.0)	5 (0.0)	2 (0.0)
Missouri	254 (0.2)	149 (0.1)	-	-	
North Carolina	1,019 (0.6)	625 (0.4)	13 (0.0)	80 (0.2)	4 (0.0)
North Dakota	113 (0.1)	86 (0.1)	71 (0.1)	16 (0.0)	5 (0.0)
New Jersey	722 (0.4)	732 (0.5)	5 (0.0)	4 (0.0)	6 (0.0)
New Mexico	41 (0.0)	23 (0.0)	11 (0.0)	-	
New York	9 (0.0)	676 (0.5)	-	-	
Ohio	-	-	19 (0.0)	-	
Oregon	3 (0.0)	-	-	-	
Pennsylvania	252 (0.2)	267 (0.5)	110 (0.2)	96 (0.2)	49 (0.1)
Rhode Island	1 (0.0)	-	-	-	
South Carolina	1,072 (0.7)	696 (0.5)	16 (0.0)	-	
South Dakota	30 (0.0)	23 (0.0)	17 (0.0)	-	
Tennessee	5,642 (3.5)	5,076 (3.6)	3,153 (5.4)	3,127 (7.0)	2,318 (6.2)
Texas	138,601 (86.0)	121,203 (86.8)	52,033 (89.0)	38,679 (86.9)	32,381 (87.0)
Utah	82 (0.1)	16 (0.0)	59 (0.1)	-	
Virginia	2,013 (1.2)	1,652 (1.2)	209 (0.4)	62 (0.1)	25 (0.1)
West Virginia	91 (0.1)	23 (0.0)	27 (0.0)	73 (0.2)	26 (0.1)

Sample

We last updated the ISIP Advanced Reading Instructional Tier Goals in August 2011. Since that time, there has been substantial growth in the number of students using the ISIP Advanced 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 Advanced 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. In total, the ISIP Advanced Reading scores from 323,505 students were considered to establish norms. This sample used in establishing the Instructional Tier Goals for the ISIP Advanced Reading Overall ability score, as well as all subtests within ISIP Advanced Reading.

Accounting for Sample Bias

Since the ISIP Advanced Reading assessment population includes many students who struggle with reading, additional analysis were completed to account for possible bias for all grades. ISIP has a history of moderate to high correlations with the State of Texas Assessment of Academic Readiness (STAAR) in reading (Patarapichayatham, Fahle, & Roden 2013) across all grades assessed.

Table 4-2: Correlations between ISIP and STAAR

Grade	Measure	ISIP_BOY	ISIP_MOY	ISIP_EOY
4th	ISIP_MOY	.876**		
	ISIP_EOY	.841**	.879**	
	STAAR	.730**	.743**	.735**
5th	ISIP_MOY	.889**		
	ISIP_EOY	.866**	.904**	
	STAAR	.706**	.708**	.706**
6th	ISIP_MOY	.869**		
	ISIP_EOY	.917**	.845**	
	STAAR	.721**	.709**	.752**
7th	ISIP_MOY	.783**		
	ISIP_EOY	.758**	.756**	
	STAAR	.632**	.596**	.549**
8th	ISIP_MOY	.795**		
	ISIP_EOY	.825**	.825**	
	STAAR	.545**	.601**	.628**

** Correlation is significant at the 0.01 level (2-tailed)

STAAR results from the 2014-2015 school year were acquired from a partnering school district in north Texas. There were approximately 4,000 students per grade included in the sample, see Table 4-3.

The analysis examined the mean STAAR scores for all students as compared to the mean STAAR scores of only those students participating in ISIP Advanced Reading. The normalized differences between these two samples of STAAR scores, expressed as a z-scores, measured the bias in each grade data. The analysis revealed insignificant bias in grades 4 and 5, and an increasing bias in grades 6 - 8. The normalized differences were 0.01, 0.01, 0.14, 0.73. and 0.85 for Grades 4th-8th respectively. These results are consistent with anecdotal evidence that ISIP is implemented only with struggling readers at higher grades.

To account for bias in the norming sample, mean ISIP scores for each grade were adjusted proportionally by the normalized difference (z-score) found in the bias analysis.

Table 4-3: Demographics of 2014-2015 STAAR results

	Grade				
	4th	5th	6th	7th	8th
	Frequency (%)	Frequency (%)	Frequency (%)	Frequency (%)	Frequency (%)
Gender					
Male	2,186 (52.9)	2,150 (52.6)	2,110 (51.3)	2,176 (51.8)	2,253 (51.8)
Female	1,949 (47.1)	1,936 (47.4)	2,004 (48.7)	2,022 (48.2)	2,100 (48.2)
Economic Disadvantage					
No	1,336 (32.3)	1,314 (32.2)	1,395 (33.9)	1,414 (33.7)	1,491 (34.3)
Yes	2,799 (67.7)	2,772 (67.8)	2,719 (66.1)	2,784 (66.3)	2,862 (48.2)
Race					
Hispanic	2,124 (51.4)	2,087 (51.1)	2,117 (51.5)	2,135 (50.9)	2,171 (49.9)
White	857 (20.7)	844 (20.7)	800 (19.4)	831 (19.8)	921 (21.2)
Black	719 (17.4)	708 (17.3)	713 (17.3	783 (18.7)	811 (18.6)
Asian	341 (8.2)	344 (8.4)	371 (9.0)	355 (8.5)	338 (7.8)
Multi	67 (1.6)	91 (2.2)	86 (2.1)	72 (1.7)	86 (2.0)
Am Ind	23 (0.6)	10 (0.2)	25 (0.6)	18 (0.4)	23 (0.5)

Computing Norms

Istation's norms are time-referenced to account for expected growth of students over the course of a semester. The ISIP Advanced 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 Advanced Reading scores of Overall Reading Ability, Word Analysis, Vocabulary, Comprehension, and Text Fluency were used to develop the updated Instructional Tier Goals.

To compute these norms, means and standard deviations were computed from the three assessment points collected, adjusted for bias, 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 Advanced 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

Alexander, D. A., Gray, D. B., & Lyon, G. R. (1993). Future directions for scientific research in learning disabilities. In G. R. Lyon, D. B. Gray, J. F. Kavanagh, & N. A. Krasnegor (Eds.), *Better understanding learning disabilities: Perspectives on classification, identification and assessment* (pp. 343-350). Baltimore: Paul H. Brookes Publishing Co.

Alliance for Excellent Education. (2006). Who's counted? Who's counting? Understanding high school graduation rates. Washington, DC.: Author.

Alonzo, J., Ketterlin-Geller, L. R., & Tindal, G. (2007). Curriculum based assessment. In L. Florian (Ed.), *Handbook of special education*. Thousand Oaks, CA: Sage Publications.

Archer, Gleason, & Vachon. (2003). Decoding and fluency: Foundation skills for struggling older readers. *Learning Disability Quarterly*, *(26)2*, 89-102.

Baker, D. W., Parker, R. M., Williams, M. V., Clark, W. S., & Nurss, J. (1997). The relationship of patient reading ability to self-reported health and use of health services. *American Journal of Public Health*, 87(6), 1027–1030.

Baker, D. W., Parker, R. M., Williams, M. V., Pitkin, K., Parikh, N. S., Coates, W., & Mwalimu, I. (1996). The health experience of patients with low literacy. Archives of Family Medicine, 5, 329–334.

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

Bock, R. D., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-444.

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

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

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

Brown-Chidsey, R., Davis, L., & Maya, C. (2003). Source of variance in curriculum-based measures of silent reading, *Psychology in the Schools, 40*, 363-376.

Bryant, D. P., Vaughn, S., Linan-Thompson, S., Ugel, N., Hamff, A., & Hougen, M. (2000). Reading outcomes for students with and without reading disabilities in general education middle-school content area classes. *Learning Disability Quarterly, 23*(4), 238-252.

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

Cain, K., & Oakhill, J. V. (1999). Inference making and its relation to comprehension failure. *Reading and Writing: An Interdisciplinary Journal, 11,* 489-504.

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

Cain, K., Oakhill, J. V., Barnes, M. A., & Bryant, P. E. (2001). Comprehension skill, inference making ability and their relation to knowledge. *Memory and Cognition, 29,* 850-859.

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

Conte, K. L., & Hintze, J. M. (2000). The effects of performance feedback and goal setting on oral reading fluency curriculum-based measurement. *Diagnostique, 25*(2), 85.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297_334.

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

Dale, E., & O'Rourke, J. (1981). *The living word vocabulary.* Chicago: World Book—Childcraft International, Inc.

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

de Ayala, R. J. (2009). The theory and practice of item response theory. New York: Guilford.

Deno, S. L. (2003) Developments in curriculum-based measurements. *The Journal of Special Education,* 37(3), 184-192.

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

DHHS (2000). Healthy People 2010. Department of Health and Human Services. Washington, D.C. Government Printing Office.

DiGangi, S. A., Jannasch-Pennell, A., Yu, C. H., & Mudiam, S. V. (1999). Curriculum- based measurement and computer-based assessment. Society for Computers in Psychology.

Ehri, L. C. (1997). Learning to read and learning to spell are one and the same, almost. In C. A. Perfetti, L. Rieben, & M. Fayol (Eds.), *Learning to Spell* (pp. 237–269). Hillsdale, NJ: Erlbaum.

Ehri, L. C. (1998). Grapheme-phoneme knowledge is essential for learning to read words in English. In J. L. Metsala, & L. C. Ehri (Eds.), *Word recognition in beginning literacy* (pp. 3–40). Mahwah, NJ: Erlbaum.

Ehri, L. C. (2000). Learning to read and learning to spell: Two sides of a coin. *Topics of Language Disorders, 20*(3), 19-36.

Ehri, L. C. (2005). Learning to read words: Theory, findings, and issues. *Scientific Studies of Reading, 9*(2), 167-188.

Ehri, L. C., & Wilce, L. (1987). Does learning to spell help beginning readers learn to read words? *Reading Research Quarterly*, *22*, 47–65.

Encarta Dictionary. http://encarta.msn.com/

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

Everyday Learning Corporation (2000). *Everyday mathematics: Vocabulary by grade level.* Retrieved from http://www.jamerson-es.pinellas.k12.fl.us/docs/vocab.pdf.

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

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

Foorman, B., Santi, K., & Berger, L. (2007). *Scaling assessment-driven instruction using the Internet and handheld computers*. (pp 68-90). In B. Schneider & S. K.McDonald. *Scale-up in education.* Plymouth, UK: Rowman & Littlefield Publishers.

Fry, E.B., Kress, J.K. (2006). *The Reading Teacher's Book of Lists*. San Francisco: John Wiley & Sons, Inc.

Fuchs, D., Fuchs, L., & McMaster, K. (2003). Monitoring children who do not respond to generally effective instruction. In L. Swanson, K. Harris, & S. Graham (Eds.), *Handbook of learning disabilities* (pp. 431-449). New York: Guilford.

Fuchs, L. S. & Fuchs, D. (2008). Best practices in progress monitoring reading and mathematics at the elementary grades. In J. Grimes & A. Thomas (Eds.), *Best practices in school psychology* (pp. 2147-2164). Bethesda, MD: National Association of School Psychologists.

Fuchs, L. S. (1986). Monitoring progress among mildly handicapped pupils: Review of current practice and research. *Remedial and Special Education*, *7*(5), 5-12.

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

Fuchs, L. S., & Fuchs, D. (1991). Curriculum-based measurement: Current applications and future directions. In L. S. Fuchs & D. Fuchs (Eds.), *Applications of curriculum-based measurement*: *Preventing school failure*, 35, 6-12.

Fuchs, L. S., & Fuchs, D. (1992). Identifying a measure for monitoring student reading progress. *School Psychology Review*, 21, 45-58.

Fuchs, L. S., & Fuchs, D. (1999). Monitoring student progress toward the development of reading competence: A review of three forms of classroom-based assessment. *School Psychology Review, 28*, 659-671.

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., 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.

Gamino, J. F., & Chapman, S. B. (in press). Reasoning in children with attention deficit hyperactivity disorder*:* A review of current research*.* Center for Brain Health: The University of Texas at Dallas.

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

Glencoe (2000). *Grammar and composition handbook: Middle school.* New York: Glencoe McGraw-Hill.

Good, R. H., & Kaminski, R. A. (2002). Dynamic indicators of basic early literacy skills (6th Ed.) Eugene, OR: Institute for the Development of Educational Achievement.

Good, R. H., III, Simmons, D. C., & Kame'enui, E. J. (2001). The importance of 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*(3), 257-288.

Graham, S. (2000). Should the natural learning approach replace traditional spelling instruction? *Journal of Educational Psychology, 92,* 235-247.

Graham, S., Harris, K. R., & Chorzempa, B. F. (2002). Contribution of spelling instruction to the spelling, writing, and reading of poor spellers. *Journal of Educational Psychology, 94*(4), 669-686.

Harcourt School Publishers (n.d.). Retrieved from http://harcourtschool.com/.

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

Hiebert, E. A., Valencia, S.W., & Afflerbach, P. P. (1994). *Definitions and perspectives*. In S. H. Valencia, E. H. Hiebert P. P. Afflerbach (Eds.) *Authentic reading assessment: Practices and possibilities*. Newark: IRA, 6-21.

Hoover, H. D., Dunbar, S. B., & Frisbie, D. A. (2007). Iowa tests of basic skills (ITBS). Rolling Meadows, IL: Riverside Publishing.

Houghton Mifflin Harcourt (2008). *Spelling and vocabulary.* Retrieved from http://www.eduplace.com/rdg/hmsv/index.html

Irwin, J. W. (1991). Teaching reading comprehension processes (2nd Ed.) Englewood Cliffs, NJ: Prentice Hall.

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

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

Johnson-Laird, P. N. (1983). Mental models. Cambridge, UK: Cambridge University Press.

Joshi, R. M., Treiman, R., Carreker, S., & Moats, L. C. (2008). How words cast their spell: Spelling is an integral part of learning the language, not a matter of memorization. *American Educator,* 6-16, 42.

Kalinowski, K. E. (2009). ISIP early reading reliability and validity evidence. Dallas, TX: Istation. http://www1.istation.com/research/pdfs/isip_rr.pdf.

Kaminski, R. A. & Good III, R. H. (1996). Toward a technology for assessing basic early literacy skills. *School Psychology Review.* 25(2), 215-227.

Keenan, J. M., & Betjemann, R. S. (2006). Comprehending the Gray Oral Reading Test without reading it: Why comprehension tests should not include passage independent items. *Scientific Studies of Reading, 10*(4), 363-380.

Keenan, J. M., Betjemann, R.S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading, 12*(3), 281-300.

Kim, D., de Ayala, R. J., Ferdous, A. A., & Nering, M. L. (2007, April). Assessing relative performance of local item dependence (LID) indexes. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Kintsch, W. (1998). Comprehension: Paradigms in cognition. New York: Cambridge University Press.

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing*, 21, 202-206.

Lee, J., Grigg, W. S., & Donohue, P. (2007). The Nation's Report Card: Reading 2007. National Center for Education Statistics, Institute for Education Sciences, U.S. Department of Education, Washington, D.C.

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

Linacre, J. M. (2000). Computer-adaptive testing: A methodology whose time has come. Chicago, IL: MESA Press (Memorandum No. 69).

Linacre, J. M., & Wright, B. D. (2000). WINSTEPS v. 3.00: Rasch item analysis computer program manual. Chicago, IL: MESA Press.

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

Lyon, G. R. & Chhabra, V. (l996). The current state of science and the future of specific reading disability. *Mental Retardation and Developmental Disabilities Research Reviews*, 2, 2-9.

Lyon, G. R. & Gray, D. B. (l992). NICHD Learning Disability Research Centers. *Learning Disabilities: A Multidisciplinary Journal*, 4, 3-4.

Lyon, G. R. (1999b). In celebration of science in the study of reading development, reading difficulties, and reading instruction: The NICHD perspective. *Issues in Education: Contributions from Educational Psychology, 5*, 85-115.

Lyon, G. R. (2002a). Reading development, reading difficulties, and reading instruction: Educational and public health issues. *Journal of School Psychology*, 40, 3-6.

Lyon, G. R. (2004). Early childhood education. U.S. House of Representatives Appropriations Committee. Washington, DC: Congressional Printing Office

Lyon, G.R. (April 1998). *Overview of NICHD reading and literacy initiatives. U.S. Senate Committee on Labor and Human Resources, United States Congress. Washington, D.C. Congressional Printing Office.*

Lyon, G. R. (July 1999a). How research Can inform the re-authorization of Title I of the Elementary and Secondary Education Act. U.S. House of Representatives Committee on Education and the Workforce, United States Congress. Washington, D.C.: Congressional Printing Office.

Lyon, G. R. (July 2003). The critical need for evidence-based comprehensive and effective early childhood programs. U.S. Senate Health, Education, Labor and Pensions Committee. Washington, DC: Congressional Printing Office.

Lyon, G. R. (June, 2002b). Learning disabilities and early intervention strategies. U.S. House of Representatives Committee on Education and the Workforce – Subcommittee on Education Reform. Washington, D.C.: Congressional printing Office.

Lyon, G. R. (l995). Research initiatives in learning disabilities contributions from scientists supported by the National Institute of Child Health and Human Development. *Journal of Child Neurology*, 10, 120-127.

Lyon, G. R. (March 2001). Measuring success: Using assessments and accountability to raise student achievement. U.S. House of Representatives Committee on Education and the Workforce, United States Congress. Washington, D.C.: Congressional Printing Office.

Lyon, G. R. (May 2000). Education research and evaluation and student achievement: Quality counts. U.S. House of Representatives Committee on Education and the Workforce, United States Congress. Washington, D.C.: Congressional Printing Office.

Lyon, G. R. (October 1999). Education Research: Is what we don't know hurting our children. U.S. House of Representatives Science Committee: Subcommittee on Basic Research, United States Congress. Washington, D.C.: Congressional Printing Office.

Lyon, G. R. (September 1997). NICHD research findings in learning disabilities. U.S. House of Representatives Committee on Education and the Workforce, United States Congress. Washington, D.C.: Congressional Printing Office.

Lyon, G. R. & Moats, L. C. (1997). Critical conceptual and methodological considerations in reading intervention research. *Journal of Learning Disabilities*, 30, 578-588.

Lyon, G. R. & Kalinowski, K. (2008). Concurrent validity of ISIP compared to DIBLES: Data from Chambersberg, PA. Unpublished manuscript. Dallas,TX: Istation.

MacGinitie, W. H., MacGinite, R. K., Maria, K., Dryer, L. G., & Hughes, K. E. (2002). Gates-Macginitie Reading Tests, 4th Edition (GMRT-R). Rolling Meadows, IL: Riverside.

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

Manyak, P. (2007). Character trait vocabulary: A schoolwide approach [Electronic version]. *Reading Teacher,* 60(6), 574-577.

Marzano, R.J., Kendall, J.S., & Paynter, D.E. (2005). *Appendix: A list of essential words by grade level.* Retrieved from http:// www.tec.leon.k12.fl.us/Vocabulary Project/Vocabulary Project Word List.pdf.

Mathes, P. G. (2007). ISIP concurrent and predicative validity. Dallas, TX: Istation. http://www1.istation.com/research/pdfs/isipcv.pdf.

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

Mathes, P. G., Torgesen, J., & Herron, J. (2007). Istation's indicators of progress (ISIP) [Computer program]. Richardson, TX: Istation.

Merriam-Webster Online. http://www.merriam-webster.com/.

MetaMetrics. (2004). Lexile analyzer. Durham, NC: MetaMetrics, Inc. Moats, L. C. (2005). How spelling supports reading and why it is more regular and predictable than you may think. *American Educator,* 12-22, 42-43.

Microsoft Word Thesaurus (2007).

Mid-Ohio Educational Service Center (n.d.). *Content area vocabulary by grade level.* Retrieved from http://www1.moesc.k12.oh.us/vocab-contents.htm#ss

Moats, L. C. (2000). From speech to print: Language essentials for teachers. Baltimore: Paul H. Brookes.

Muraki, E. (1992). A generalized partial credit model: Application of the EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

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

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

National Reading Panel. (2000). Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction (NIH Publication No. 00-4679).

Online Etymology Dictionary. http://www.etymonline.com/.

Oakhill, J. V. (1982). Construction processes in skilled and less-skilled comprehenders' memory for sentence. *British Journal of Psychology, 73,* 13-20.

Oakhill, J. V., Hartt, J., & Samols, D. (2005). Levels of comprehension monitoring and working memory in good and poor comprehenders. *Reading and Writing: An Interdisciplinary Journal, 18,* 657-686.

Osborn, J., Lehr, F., & Hiebert, E. H. (2003). A focus on fluency. Honolulu, HI: Pacific Resource for Education and Learning.

Pearson Assessments. (2009). Stanford achievement test series, Tenth edition (SAT10). San Antonio, TX, Author.

Pearson, P. D., Hiebert, E. H., & Kamil, M. L. (2008). Vocabulary assessment: What we know and what we need to learn. *Reading Research Quarterly, 42*, 282-296.

Perfetti, C. A. (1986). Continuities in reading acquisition, reading skill and reading ability. *Remedial and Special Education, 7*(1), 11-21.

Perfetti, C. A. (1994). Psycholinguistics and reading ability. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 849-894). San Diego: Academic Press.

Prior, S. M., & Welling, K. A. (2001). "Read in your head": A Vygotskian analysis of the transition from oral to silent reading. *Reading Psychology, 22*, 1–15.

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. Chicago, IL: University of Chicago Press.

Rasinski, T. V., Padak, N. D., McKeon, C. A., Wilfong, L. G., Friedauer, J. A., & Heim, P. (2005). Is reading fluency a key for successful high school reading? *Journal of Adolescent & Adult Literacy, 49*(1), 22-27.

Read, J. (2007). Second language vocabulary assessment: Current practices and new directions. *International Journal of English Studies,* 7, 105-125.

Reading key: A powerful reading vocabulary program (n.d.). Retrieved from http://readingkey.com/.

Roberts, G., Torgesen, J. K., Boardman, A., & Scammacca, N. (2009). Evidence-based strategies for reading instruction of older students with LD or atrisk. *Learning Disabilities Research &Practice*.

Roid, G. H. (2006). Designing ability tests. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 527-542). Mahway, NJ: Routledge/Erlbaum.

Roid, G. H., & Haladyna, T. M. (1982). A technology for test-item writing. Orlando, FL: Academic Press.

Rueda, R., Saldivar, T., Shapiro, L., Templeton, S., Terry, C.A., Valentino, C., & Wolf, S. (2001). *English*. Boston: Houghton Mifflin.

Scammacca, N., Roberts, G., Vaughn, S., Edmonds, M., Wexler, J., Reutebuch, C. K., et al. (2007). Reading interventions for adolescent struggling readers: A meta-analysis with implications for practice. Portsmouth, NH: RMC Research Corporation, Center on Instruction.

Schumaker, J. B., Deshler, D. D., Alley, G., Warner, M., & Denton, P. (1982). MULTIPASS: A learning strategy for improving reading comprehension. *Learning Disability Quarterly, 5*, 295-340.

Senn, J.A., Skinner, C.A. (2001). *English: Communication skills in the new millennium*. Austin: Barrett Kendall Publishing.

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

Shinn, M., Good, R., (1992). Curriculum-based measurement of oral reading fluency: A confirmatory analysis of its relationship to reading. *School Psychology Review, 21*, 459-480.

Silberglitt, B., Burns, B.M., Madyun, N.H., & Lail., K.E. (2006). Relationship of reading fluency assessment data with state accountability t scores: A longitudinal comparison of grade levels. *Psychology in the Schools*, Vol. 43(5).

Stahl, S. A. (2003). Vocabulary and readability: How knowing word meanings affects comprehension. *Topics in Language Disorders, 23*(3), 241-247.

Stahl, S. A., & Fairbanks, M. M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research, 56*(1), 72-110.

State of Utah Office of Education (2008). *Social studies concept list.* Retrieved from http://www.schools.utah.gov/curr/socialstudies/pdf/WordList.pdf.

Swain, K. D., & Allinder, R. M. (1996). The effects of repeated reading on two types of CBM: Computer maze and oral reading with second-grade students. *Diagnostique, 21*, 51-66.

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

Texas Education Agency. (2003). Texas assessment of knowledge and skills *(TAKS)*. Austin, TX: Author.

Texas Education Agency (1985). *Texas educational assessment of minimum skills: Cumulative vocabulary list.* Texas: Government Publishing Texas State Documents.

Thissen, D. J., Chen, W. H., & Bock, R. D. (2003). *MULTILOG* (Version 7.0) [Computer program]. Mooresville, IN: Scientific Software.

Torgesen, J. K., Houston, D. D., Rissman, L. M., Decker, S. M., Roberts, G., Vaughn, S., et al. (2007). Academic literacy instruction for adolescents: A guidance document from the Center on Instruction. Portsmouth, NH: RMC Research Corporation, Center on Instruction. Van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item-response theory*. New York: Springer.

Weiser, B., & Mathes, P. G. (2009). Does encoding instruction improve the reading and spelling performances of elementary students at risk for reading difficulties? A best-evidence synthesis. Manuscript in review.

Wechsler, D. (2005). Wechsler individual achievement test (2nd ed.). (WIAT-II). San Antonio, TX: Harcourt Assessment.

West Contra Costa Unified School District (n.d.). *Daily mathematics vocabulary from Harcourt math.* Retrieved from http://www.wccusd.k12.ca.us/math/marc/3/.

Williams, J. P. (2003). Teaching text structure to improve reading comprehension. In H. L. Swanson, K. R. Harris, & S. Graham (Eds.), *Handbook of learning disabilities* (pp. 293-305). New York: The Guilford Press.

Wixson, K. K., & Peters, C. W. (1987). Comprehending assessment: Implementing and interactive view of reading. *Educational Psychologist, 22*(3 & 4), 333-356.

Woodcock, R. W. (1991). Woodcock language proficiency battery vised (WLPB-R). Rolling Meadows, IL: Riverside Publishing.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). Woodcock-Johnson III tests of achievement (WJ-III ACH). Rolling Meadows, IL: Riverside Publishing.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.

Ysseldyke, J., & Bolt, D. M. (2007). Effect of technology-enhanced continuous progress monitoring on math achievement. *School Psychology Review, 36*, 453-467.

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

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG (Version 3.0) [Computer program]. Mooresville, IN, Scientific Software.