# Istation's Indicators of Progress Español

# Technical Report

Computer Adaptive Testing System for Continuous Progress Monitoring of Reading Growth for Students Pre-K through Grade 3

Istation

Supporting Educators. Empowering Kids.
Changing Lives.

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

www.istation.com

# Table of Contents

# Chapter 1: Introduction

ISIP™, Istation's Indicators of Progress, Español (ISIP Español) 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 Spanish early reading.

The ISIP Español assessment is based on sound standards for educational testing and is guided by the latest publications used internationally through Early Grade Reading Assessment (EGRA) (Sprenger-Charolles et al., 2000). These foundational bases were used to design the framework utilized for item writing and editing.

Designed for students in Pre-Kindergarten through Grade 3, who are receiving language arts reading instruction in Spanish, ISIP Español 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 Español provides growth information in the five critical domains of early reading: phonemic awareness, alphabetic knowledge and skills, vocabulary, fluency, and comprehension. It is designed to (a) identify children at risk for reading difficulties, (b) provide automatic continuous progress monitoring of skills that are predictors of later reading success, and (c) provide immediate and automatic linkage of assessment data to student-learning needs, which facilitates differentiated instruction.

ISIP Español has been designed to automatically provide continuous measurement of Pre-Kindergarten through Grade 3 student progress throughout the school year in all the critical areas of early reading, including phonemic awareness, alphabetic knowledge and skills, fluency, vocabulary, and comprehension. This is accomplished through short tests, or "probes," administered at least monthly, that sample critical areas that predict later performance. Assessments are computer–based, and teachers can arrange for entire classrooms to take assessments as part of scheduled computer lab time or individually as part of a workstation rotation conducted in the classroom. The entire assessment battery for any assessment period requires 40 minutes or less. It is feasible to administer ISIP Español assessments to an entire classroom, an entire school, and even an entire district in a single day - given adequate computer resources. Classroom and individual student results are immediately available to teachers, illustrating each student's past and present performance and skill growth. Teachers are alerted when a particular student is not making adequate progress so that the instructional program can be modified before a pattern of failure becomes established.

# The Need to Improve Testing Practices in Bilingual Classrooms

Districts implementing special language programs are required to designate students' academic plans based on their unique needs. Consequently, school districts are currently addressing academic issues in both Spanish and English, utilizing tests results in both languages when students are enrolled in bilingual education classrooms. The current testing practices have proven to be unfavorable to teachers' instructional time as well as the district- and school-funding needs.

Current national discussions regarding academic services for students whose first language is Spanish have found that academic programs suffer from a dearth of assessments that prove to be non-biased and appropriate (O'Hanlon, 2005; Escamilla, 2006 and the National Center for Latino Child and Family Research, 2009).

Bilingual education programs, including Dual language models, are in great need of improving current testing practices. Improved practices should (a) allow teachers to re-direct time to instructional purposes, (b) target funding to other academic needs, and (c) adopt testing tools that are culturally, linguistically, and cognitively appropriate for programs that follow Spanish Language Arts and Reading standards.

Monitoring students' literacy ability and academic growth in each language is necessary to attend to bilingual students' academic needs. Both versions of ISIP, Early Reading and Español, can provide tools for monitoring literacy development in two languages. These tests were built individually, using different items, and based on separate field-test data.

Obtaining data results that are relevant, reliable, and valid improve assessment practices. To be relevant, data must be available on a timely basis and target important skills that are influenced by instruction. To be

reliable, there must be a reasonable degree of confidence in the student scores. To be valid, the skills assessed must provide information that is related to student performance expectations. Hispanic students who rely on their mother language to excel academically have been found to be impacted negatively by results of invalid, biased, or inadequate commonly used assessment practices. As a result, there has been an over-identification and/or under-identification of Spanish speaking students in special education programs (Espinosa & López, 2007). Identifying students' progress toward literacy in the language of instruction will produce more effective identification of true intervention needs and special education cases.

Roseberry-McKibbin and O'Hanlon (2005) reviewed surveys completed by public school speech-language pathologists on service delivery for non-native English speakers from 1990 through 2001 and found that there was a dearth of assessments that proved to be both unbiased and appropriate. Test items were generally outside of the students' cultural knowledge; and therefore unfamiliar to speakers of other languages, resulting in students' inability to demonstrate the skill being tested. Additionally, test norms based on native speakers of English should not be used with individuals whose first language is not English; and those individuals' test results should be interpreted as reflecting, in part, their current level of proficiency rather than their ability, potential, aptitude, etc. (AERA, APA, & National Council on Measurement in Education [NCME]; Standards for Educational and Psychological Testing, 1999). In order to assess Spanish speaking students, tests must be adequately tested for cultural relevance and proper Spanish terminology that avoids regionalisms and colloquial terms. At the same time, items must demonstrate internal consistency when tested and scored with the population they intend to evaluate; in this case, Hispanic students in bilingual education classrooms in the US public education system are targeted.

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

# The Need to Link Spanish Early Reading Assessment to Instructional Planning

Instructional time is utilized more effectively when assessment is linked to instruction. Early reading assessments of Spanish literacy development need to (a) identify students at risk for reading difficulties, students that may need extra instruction or intensive intervention if they are to progress toward grade-level standards in reading by year end; (b) monitor student progress for skill growth on a frequent and ongoing basis and identify students that are falling behind; (c) provide information about students who will be helpful

in planning instruction to meet their needs; and (d) assess whether students achieved grade level reading standards at the end of the school year.

A teacher needs to be able to identify students at risk of reading failure and/or struggling to meet end-of-year grade level expectations. These individualized student data support differentiated instruction; therefore, teachers must first have information about the specific needs of each child.

Linking teacher instruction to the results of assessment is promoted by using formative assessments. Following progress through formative assessments needs to occur often enough that teachers may discover when instruction has not been effective in order to make modifications in a timely manner (Crooks, T., 2001). According to current research, the best examples that follow a formative assessment structure are called "Online Formative Assessment" (Gomersall, 2005; Nicol, D.J. & Macfarlane-Dick, D., 2006). It is also envisioned that computer-based formative assessments will play an increasingly important role in learning, with the increased use of banks of question items for the construction and delivery of dynamic, on-demand assessments (Guide to Assessment, Scottish Qualifications Authority; June 2008).

Research suggests that children with different levels of language proficiency who are also developing literacy skills (whether in one language or two) respond successfully to frequent formative assessments. These assessments' results pinpoint skills as they are emerging and provide the best information as to which readers require additional support in specific reading skills (Gersten et al., 2007).

The purpose of formative assessment can be defined as assessment "for learning, rather than of learning" (Stiggins & Chappuis, 2006, p. 10). Equal educational opportunities for emergent readers should offer the use of formative assessments as a necessity, regardless of language of instruction. Formative assessments provide detailed pictures of the abilities that are measured, in order to make modifications to the instruction that is relevant and, in many cases, critical to students' progress. The primary goals of formative assessment are to guide curriculum and teaching strategies. Districts, teachers, and curriculum developers use data to differentiate classroom instruction while monitoring academic progress. It is important to engage in an ongoing process rather than a single test when using formative assessment. Consistent measures of student progress that involve students in the process enable opportunities for both teachers and students to work together toward common goals. Assessment tools that support self-monitoring contribute to engaging students in self-driven progress practices (McManus, 2008).

A systematic and collaborative process that involves self-monitoring and feedback benefits both teachers and students, because it promotes engagement in meta-cognitive processing that informs learning and increases student achievement (Stiggins & Chappuis, 2006). This type of assessment is most useful when (a) it is conducted periodically, (b) it provides information immediately, (c) it is easy and systematic in administration, and (d) it helps gather a more complete picture of each student, including a range of ability to perform an academic task that varies constantly (Gersten et al., 2007). Computer-based evaluations support all four strengths of formative assessment and allow students to self-monitor their progress.

# Continuous Progress Monitoring

ISIP Español grows out of the model of Continuous Progress Monitoring (CPM) called Curriculum-Based Measurement (CBM). Teachers who monitor their students' progress and use this data to inform instructional planning and decision-making have higher student outcomes than those who do not (Conte & Hintze, 2000; Fuchs, Fuchs, Hamlett, & Ferguson, 1992; Mathes, Fuchs, & Roberts, 1998). These teachers also have a more realistic conception of the capabilities of their students than teachers who do not regularly use student data to inform their decisions (Fuchs, Deno, & Mirkin, 1984; Fuchs, Fuchs, Hamlett, & Stecker, 1991; Mathes et al., 1998).

The collection of sufficient, reliable assessment data on a continuous basis is a daunting task for schools and teachers. Screening and inventory tools for Spanish literacy such as the *Tejas LEE*® (Brookes Publishing Co.) and IDEL®: *Indicadores Dinámicos del Èxito en la Lectura* (Good & Kaminski, 2002) use a benchmark or screen schema, in which testers administer assessments three times a year. More frequent continuous progress monitoring is recommended for all low-performing students, but administration is at the discretion of already overburdened schools and teachers.

Districts currently use CBM models to index student progress over time, which in turn can facilitate teachers' formative evaluation of their teaching effectiveness. Research indicates that CBM can accurately, meaningfully, and sensitively describe such progress (Marston, 1989). This is accomplished through the frequent administration of short, equivalent tests sampling all the skills in the curriculum. A student's past, present, and probable future growth is tracked. When students are not making adequate progress, teachers modify their instructional programs. The educational value of CBM would greatly benefit the outcomes of bilingual education programs, and research demonstrates that instructional programs designed with CBM can result in greater student achievement, enhanced teacher decision making, and improved student awareness of learning (e.g., Fuchs, Fuchs, Hamlett, & Stecker, 1991). Thus, CBM represents a logical model for helping bilingual teachers to identify those students for whom the standard curriculum in place in the classroom is not having the desired effects. Once identified, teachers can intervene before failure has already occurred.

Although proven to be a great tool for classroom teachers, CBM has not been as widely embraced as would be hoped and has hardly been recognized in the field of bilingual education. These assessments, even in their handheld versions, require a significant amount of work to be administered individually to each child. The examiners who implement these assessments must also receive extensive training in both the administration and scoring procedures to uphold the reliability of the assessments and avoid scoring errors. Because these assessments are so labor-intensive, they are expensive for school districts to implement. Bilingual education classrooms, already pressed for time to evaluate students' academic and proficiency needs in two languages, are unable to easily accommodate the requirements of CBM implementation. Therefore, it is difficult for bilingual teachers to be able to use CBM models for continuous progress monitoring and validation of test results.

The introduction of handheld technology has allowed for graphing of student results. Assessments like Tejas LEE (Brookes) can be recorded using palm-pilot devices, but information in this format is often not available on a timely basis for total class or whole school results. Additionally, the time needed for one-on-one administration and the need for additional staff to support classroom teachers during testing periods make it difficult to implement with fidelity and consistency.

Computer applications have been found to be reliable means by which to deliver CBM models by applying similar equivalent test sampling with students over time, using the computer platform to deliver the assessments and a program to collect the data, both immediately and over time.

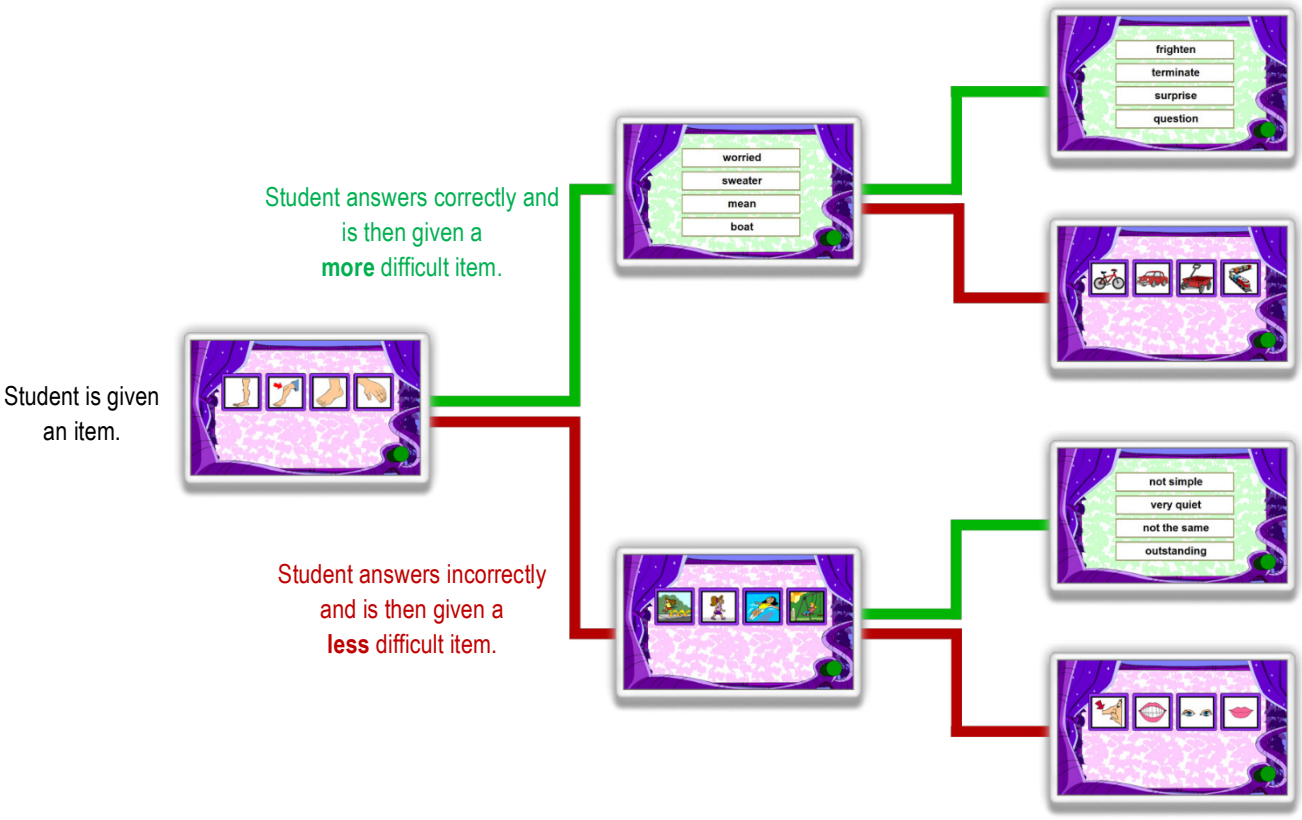
Computerized CBM applications are a logical step in increasing the likelihood that continuous progress monitoring occurs more frequently with monthly or even weekly assessments in both the general education and bilingual education classrooms. Computerized CBM applications have been developed and used 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 children according to risk and recommended instructional levels. Student results are instantly plotted on progress charts with trend lines projecting year-end outcomes based upon growth patterns, eliminating the need for teachers to manually create documentation of results.

# Computer Adaptive Testing

With recent advances in Computer Adaptive Testing (CAT) and computer technology, it is now possible to create CPM assessments that adjust to the actual ability of each child. Thus, CAT replaces the need to create parallel forms. Assessments built on CAT are sometimes referred to as "tailored tests" because the computer selects items for students based on their performance, thus tailoring the assessment to match the performance abilities of the students. This also means that students who are achieving significantly above or below grade expectations can be assessed to more accurately reflect their true abilities.

There are many advantages to using a CAT model rather than a more traditional parallel forms model, as is used in many early-reading instruments. For instance, it is virtually impossible to create alternate forms of any truly parallel assessment. The reliability from form to form will always be somewhat compromised. However, when using a CAT model, it is not necessary for each assessment to be identically difficult to the previous and future assessments. Following a CAT model, each item within the testing battery is assessed to determine how well it discriminates ability among students and how difficult it actually is through a process called Item Response Theory (IRT) work. Once item parameters have been determined, the CAT algorithm can be programmed. Then, using this sophisticated computerized algorithm, the computer selects items based on each student's performance, selecting easier items if previous items are missed and harder items if the student answers correctly. Through this process of selecting items based on student

performance, the computer is able to generate "probes" that have higher reliability than those typically associated with alternate formats and that better reflect each student's true ability.




# ISIP Español Domains

ISIP Español uses a CAT algorithm that tailors each assessment to the performance abilities of individual children while measuring progress in the critical early reading skill domains.

The specific domains and the order in which the domains and skills are presented in ISIP Español are based on an analysis of the findings and recommendations of the United States National Reading Panel, European and Latin-American research, including the latest publications from *Marco Común Europeo de Referencia Para Las Lenguas: Aprendizaje, Enseñanza, y Evaluación. [Instituto Cervantes, Ministerio de Educación, Cultura y Deporte, España; 2001].* In addition, the following research findings were considered when developing the assessment blueprint for ISIP Español:

> Es evidente que las prácticas educativas orientadas a exponer al niño a experiencias de comunicación, de intercambio comunicativo, de partir de sus experiencias previas, de tener sentido aquello que se trata de descodificar, etc. es algo que está plenamente justificado y que no importa para ello el contexto idiomático. Sin embargo, los hallazgos

más recientes, desde una perspectiva psicolingüística, ponen de manifiesto que todo ello no sería suficiente ya que el proceso cognitivo de asociación grafía-fonema es un elemento imprescindible cuando se aprende a leer en un sistema alfabético *[Enseñanza de la lectura: de la teoría y la investigación a la práctica educativa. Juan E Jiménez & Isabel O'Shanahan; Universidad de La Laguna, España. Marzo 2008].*

English Translation: It is clear that educational practices designed to expose children to experiences of communication, communicative exchange, use of their prior experiences, making sense of what is referred to as decoding, etc. is something that is fully justified and is not dependent upon the linguistic context. However, recent findings from a psychological perspective indicate that all of this would not be sufficient since the cognitive process of **grapheme/sound correspondence** is an essential element when learning to read in an alphabetic system.

Studies have also demonstrated that the performance of reading words for students learning to read in different linguistic contexts (such as English, French, and Portuguese) is systematically higher in Spanish than in other languages. These studies have also found that knowledge of the complex rules of grapheme/sound correspondence occurs earlier in Spanish than in English, French, or Portuguese. Similarly, Spanish-speaking students reach higher levels of word reading earlier, when compared to students who speak other languages. Such a finding indicates that the appropriate use of the phonological process occurs earlier in Spanish than in English, French, or Portuguese. Findings from these studies have been confirmed most recently with research from Université Paris V - René Descartes in France that compared English-, German-, French-, and Spanish-speaking children learning to read. Conclusions from both studies are disclosed below:

On the one hand, when Grapheme-Phoneme Correspondences (GPC) are almost regular, as in Spanish, reliance on the phonological procedure very often leads to the production of the correct word. Thus, in shallow orthographies, reading skills burst out very rapidly. On the other hand, when the number of inconsistent words is significant, as in English, and to a lesser extent in French, reliance on the GPC procedure sometimes leads to a reading error and reading acquisition is slowed down because of some incoherence between sub-lexical and lexical outputs (From Linguistic Description to Psycholinguistic Processing, Liliane Sprenger-Charolles and Danielle Béchennec, CNRS & Université René Descartes, Paris, 2008).

Cuando se ha comparado el rendimiento en lectura de palabras entre niños que aprenden a leer en distintos contextos idiomáticos (v. gr., inglés, francés y portugués) éste es sistemáticamente más alto en español que en otras lenguas Así, el conocimiento de las reglas de CGF complejas es más temprano en español que en inglés, francés y portugués. Igualmente, los niños españoles alcanzan altos rangos de lectura de palabras muy temprano si los comparamos con otras lenguas, lo que indicaría que la utilización adecuada del procedimiento fonológico ocurre más pronto en español que en inglés,

francés y portugués (Enseñanza de la lectura: de la teoría y la investigación a la práctica educativa. Juan E Jiménez & Isabel O'Shanahan; Universidad de La Laguna, España, 2008).

Additionally, ISIP Español domains are parallel to those in the Early Grade Reading Assessment (EGRA) conducted in Latin American countries like Nicaragua and Guatemala. These research studies have consistently reported the following critical areas for Spanish early literacy development, shown below:

| Domain |
|---|
| **CONCIENCIA FONOLÓGICA (CF)**<br>**Phonemic Awareness**<br>La instrucción en CF consiste en enseñar a los niños a segmentar el lenguaje oral en fonemas sin apoyo de las letras del alfabeto.<br>*Phonemic awareness allows students to segment oral language in phonemes without using the letter names.* |
| **CONVERSIÓN GRAFEMA-FONEMA (CGF)**<br>**Grapheme-phoneme conversion**<br>La instrucción de reglas de CGF es una forma de enseñar a leer que enfatiza la adquisición de las correspondencias símbolo-sonido.<br>*Grapheme-phoneme conversion comprises the reading rules to acquire symbol-sound correspondence.* |
| **VOCABULARIO**<br>**Vocabulary**<br>Hay dos tipos de vocabulario: el oral y el escrito. Cuando un lector encuentra una palabra en el texto puede decodificarla, es decir, convertirla en habla.<br>*Vocabulary objectives can be divided in two categories: oral and written. Decoding enables conversion of text into a verbal outcome.* |
| **COMPRENSIÓN**<br>**Comprehension**<br>Las investigaciones sugieren que la comprensión mejora cuando los alumnos son capaces de relacionar las ideas que están representadas en el texto con su propio conocimiento y experiencias, al igual que las representaciones mentales construidas en la memoria.<br>*Comprehension is improved when the students are able to relate ideas from the text to their own background knowledge.* |
| **LECTURA CON FLUIDEZ**<br>**Fluency**<br>La fluidez en la lectura es necesaria para la comprensión. Leer con velocidad, precisión, y entonación respetando los signos de puntuación facilita la comprensión del texto.<br>*Fluency is necessary to develop correct pace, observing punctuation and thus enhancing reading comprehension.* |

Taking into consideration the studies conducted on the transparency of languages, there are common elements for assessment such as grapheme-phoneme correspondence, word reading, level of vocabulary, reading comprehension of both narrative and expository texts, listening comprehension, and fluency, but there are also elements that are critical to the support of reading development in each particular language (O'Shanahan, Jimenez, 2008). In the case of the Spanish language, writing development and orthography are closely tied to reading. This is the reason why reading in Spanish is referred to as lecto-escritura (Ferreiro, Chile 2002; Bazán, Acuña, Vega, Mexico 2008). This term posits an intrinsic relationship between writing and reading.

There are differences, as well as similarities, in emergent reading and writing behaviors of Spanish-speaking children (Escamilla & Coady, 1998). English writing rubrics cannot help to guide instruction in Spanish. Differences in writing development can impact outcomes in grade-level and state-standards-based assessments. Issues that emerged from this research highlight Spanish primary students' development, in which vowels emerge before consonants; primary students move from strings of letters to invented spelling in Spanish earlier than English speakers do. A writing component is critical for the assessment of emergent literacy skills in Spanish-speaking children. ISIP Español domains include a writing component that aims to address the specific needs of children developing Spanish literacy skills based on the principles of Spanish lecto-escritura.

The growing enrollment of Spanish-speaking students in the Texas public education system clearly reveals the need to develop Spanish assessments that prove to be not only linguistically and culturally appropriate but also aligned with language arts standards and delivered efficiently. ISIP Español allows for more equitable educational opportunities for students, particularly for English-language learners who are Spanish-speaking, the largest growing number of ELLs in Texas. This student population requires qualified instructors and special language programs that support efficient and appropriate ways to assess bilingual education methodologies and special language programs geared toward improved academic achievement for Hispanic students (Guiding Principles for Dual Language Education, Center for Applied linguistics, CAL, 2007; The National Task Force on Early Childhood Education for Hispanics/La Comisión Nacional para la Educación de la Niñez Hispana, 2007; Miller, L.S. & Garcia, E. 2008).

The domains selected for the assessment measures of ISIP Español were established using the literature review described in the section above. Additionally, a number of revisions and feedback were solicited from nationally known researchers in the field of bilingual education, including Dr. Kathy Escamilla from the University of Colorado at Boulder; Dr. Barbara Flores from the University of California, San Bernardino; and Dr. William Pulte from Southern Methodist University, Dallas, Texas.

# CONCIENCIA FONOLÓGICA (CF)

## Phonemic Awareness

Items in this domain intend to evaluate the early literacy skills associated with the mechanics of reading that link to recent findings in neuropsychology studies emerging from post-modern views that impact current students' educational experience (Serie Didáctica de la Lengua y de la Literatura: Catalá Agrás G, Molina H, Bareche Monclús, R., & Editorial Graó, Barcelona, 2007).

Phonemic and syllabic awareness is "the ability to notice, think about, and work with the individual sounds in spoken words" (Armbruster, Lehr, & Osborn, 2003, p. 2). A broader term for this concept is phonological awareness.

> Los modos de representación pre-alfabética se suceden en cierto orden: primero varios modos de representación ajena a toda búsqueda de correspondencia entre la pauta sonora de una emisión y la escritura. Luego modos de representación silábicos (con o sin valor sonoro convencional) y modos de representación silábico-alfabético que preceden la aparición de la escritura. Estos niveles están caracterizados por formas de conceptualización que actúan en un sistema asimilador, absorbiendo la información dada (Alfabetización: teoría y práctica Emilia Ferreiro, 2002).
>
> English Translation: Prior to any alphabetic representation, there are identifiable audible emissions of written text that find correspondence, beginning with a single sound, followed by syllabic representations (with or without conventional value), and ending with alphabetical syllabic representations that precede emerging writing. These levels of representation are assimilated in conceptual systems that absorb the information given.

The concepts that describe phonological awareness suggest that before children learn to read print, they must understand that words are made up of speech sounds. The United States' National Reading Panel (NRP, 2000) found that children's ability to read words, to comprehend what they read, and to spell is improved with phonemic awareness. Studies of phonemic awareness conducted with Spanish-speaking children have been used recently to confirm the different levels of phonological awareness that are relevant to the Spanish language (Serrano et al., 2009). These levels comprise identification of phonemes in isolation, beginning and ending sounds and syllables, and intra-syllabic sounds, which impact the development of reading skills in unequal levels of relationship (Serrano et al, 2009).

# CONVERSIÓN GRAFEMA-FONEMA (CGF)

## Grapheme-Phoneme Conversion

Grapheme-phoneme correspondence is the ability to pair sounds (phonemes) with the letters (graphemes) that represent them. The term phonics is also widely used to describe methods used for teaching children to read and decode words (Abadzi, 2006). According to US base studies, children begin learning to read using phonics, usually around the age of five or six (NRP, 2000). In the case of some alphabetic languages such as Spanish, the orthographic representations of sounds are even simpler because there is nearly a one-to-one correspondence between letter patterns and the sounds that represent them. Even though studies conducted with Spanish-speaking children have not been completed in large quantities, as studies with English-speaking students have, the transparency of the language has been widely researched in the field of linguistics (Wimmer, Mayringer, 2001; Ziegler, Perry, Ma-Wyatt, Ladner, & Körne, 2003), placing languages such as Spanish and French on the transparent side of the languages scale and English and German on the opaque side (Seymour et al., 2003).

> En el caso del castellano, diversos trabajos han mostrado la relevancia de la sílaba, señalando que la conciencia silábica se puede usar como un buen indicador de las habilidades lectoras importantes en una ortografía transparente como el castellano, debido a la correspondencia directa entre grafemas y fonemas (Carrillo, 1994; Jiménez & Ortiz, 2000).
>
> English Translation: Studies conducted in Spanish language (terminology using "castellano" refers to Spanish) have demonstrated that syllabic awareness is a good predictor of reading skills, due to the direct influence of a transparent orthography over the grapheme-phoneme correspondence.

# Comunicación Escrita

## Written Communication

The subtests in this domain measure orthography development and dictation. Orthography measures comprise spelling and use of accent marks, while dictation measures a student's ability to follow grammatically correct sentence structures and emergent syntactic skills.

Spelling refers to the ability to determine the fully specified orthographic representations of words in the language. Knowing the spelling of a word makes the representation of it sturdy and accessible for fluent reading (Ehri, 2000; Snow et al., 2005).

Dictation refers to receptive and productive syntactic skills that have been found to be related to reading ability (Scarborough, 1990). These studies found that there are evident discrepancies between the sentences produced by preschoolers who became poor readers and sentences by those who did not.

According to research by Snow, Burns, and Griffin (1998) with young learners, there are three components of first language ability that have been shown to correlate with later reading development. These components include story recall, lexical skills, and syntactic skills. If a student is able to find the relationship of words inside a sentence after hearing it, he or she should also be able to demonstrate it productively.

# Vocabulario

## Vocabulary

Current scientific research overwhelmingly supports the idea that a dearth of vocabulary impedes reading comprehension and a broad vocabulary increases comprehension and facilitates further learning (Hirsch Jr., 2003). Adequate reading comprehension has been correlated to the number of words in a text that a reader already knows. Experts consider that a good reader knows between 90 and 95 percent of the words in a text (Nagy & Scott, 2000).

Oral language vocabulary refers in general to "the words we must know to communicate effectively" (Armbruster, Lehr, & Osborn, 2003). On the other hand, reading vocabulary refers to words that a student needs to know in order to be able to understand what is read. The development of oral language proficiency—both productive (speaking) and receptive (listening)— is key to literacy growth. Furthermore, there is a rich vein of literature that suggests that vocabulary is an important precursor to literacy (see Scarborough, 2005 for a summary of this literature).

Reading vocabulary demands knowledge of words and their relationships, as well as the ability to extract meaning from words in context. The percentage of words that a reader understands when reading a text either causes the reader to miss the gist of the reading or allows the reader to get a good idea of what is being said and; therefore, to make correct inferences in order to determine the meaning of any unfamiliar words (Hirsch Jr., 2003). Bilingual education settings need to take advantage of the vocabulary that students acquire in their native language. First language vocabulary ability has been shown to correlate with later Academic Language development. "Academic language refers to the decontextualized, cognitively challenging language used not only in school, but also in business, politics, science, and journalism, and so forth. In the classroom, it means the ability to understand story problems, write book reports, and read complex texts" (Crawford & Krashen, 2007).

# Comprensión

## Comprehension

**Comprensión auditiva | Listening Comprehension**

Items in this domain intend to evaluate students' listening comprehension proficiency levels as indicators of foundational early literacy skills.

> En todo proceso auditivo, para poder ser asimilada; la información debe ser integrada a un sistema previamente construido (o un sistema en proceso de construcción). No es la información de por sí, la que crea conocimiento, el conocimiento es el producto de la construcción de un sujeto cognoscente (Alfabetización: teoría y práctica Emilia Ferreiro, 2002).
>
> English Translation: Listening proficiency assimilates new information based on existing constructed systems that integrate new information. The information alone does not constitute knowledge. Knowledge is a product constructed through a cognizant subject.

Listening Comprehension refers to the ability to effectively receive auditory input (receptive skills) in order to understand the information that was said. In bilingual education settings, listening skills developed in the native language benefit second language acquisition. In fact, listening often develops before the productive skill of speaking, so students may depend on listening skills while they are silent for an extended period during second language acquisition (Díaz-Rico, 2008; Crawford & Krashen, 2007).

Students developing early reading skills continue to strengthen listening comprehension abilities while they are able to read silently. Reading difficulties have been found to reach the same levels that the listening comprehension disorders do. Even so, these two may not be evident simultaneously or may not be mutually exclusive (Junqué I Plaja et al., 2004).

Latest publications from the United States National Early Reading Panel (NEPL) and the National Institute for Literacy identify listening comprehension as one of the key foundational skills for later reading achievement.

**Comprensión de lectura | Reading Comprehension**

This domain parallels reading comprehension measures, as determined by each state's criterion reference tests, by incorporating the same types of questions. Comprehension questions are aligned to fiction and non-fiction objectives, such as main idea, summarization, drawing conclusions, and predicting as it applies to explicit and implicit cues.

Comprehension is defined as the process through which meaning is extracted from the written language. Comprehension measures can be classified in two types: (a) *literal* comprehension, which focuses on the recognition or retrieval of primary details, main ideas, sequences, or cause-effect patterns from the information that is explicit in the text, and (b) *inferential* comprehension, which requires establishing logical connections and relationships among facts in texts, thus allowing readers to deduce events, make generalizations, interpret facts, and relate previous knowledge or personal experiences to information implicit in the text (Camba, 2006).

## Lectura con fluidez

### Fluency

This domain allows students to read silently, indicating the number of words that they are able to read per minute while, at the same time, demonstrating accuracy.

Fluency is the ability to read text correctly and with appropriate pace. Reading with fluency requires accuracy and speed. Therefore, a fluent reader is able to read aloud effortlessly using a natural expression, as if speaking; whereas a reader who has not yet developed fluency reads slowly, word by word, with inconsistency and constant interruption. Accuracy refers to the percentage of words read correctly, and speed is the number of words read per minute. In order to measure fluency, a calculation of the number of words read correctly in one minute yields a fluency rate. Recent studies of fluency outcomes conclude that attention to connections among ideas and between these ideas, as they relate to background knowledge, are more characteristic of fluent readers than non-fluent readers (Armbruster, Lehr, & Osborn, 2003; NRP, 2000).

# ISIP Español Items

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

The United States has not adopted a set of national common Spanish Language Arts and Reading (SLAR) standards. The state of Texas requires the implementation of TEKS for Spanish Language Arts Reading (SLAR) and English as a Second Language (ESL) Elementary Standards under Texas Education Code 128.10. These standards were used to develop a "hybrid" version of Spanish standards combined with selected states and countries (i.e., California, Texas, Puerto Rico, WIDA consortium, Colombia, Mexico and Spain). The combined standards were utilized to determine end of year expectations for Pre-Kindergarten through Grade 3. These standards were revised by national experts who collaborated with istation as members of the ISIP Español Advisory Council. Once the standards and benchmarks were developed, a blueprint of skills to be assessed was determined based on an analysis of existing Spanish literacy assessments. Input and suggestions were then sought from the group of experts in Spanish language content, linguistics, and language acquisition that comprised the Advisory Council. Members of this council are listed below:

**Dr. Iliana Alanis**
Assistant Professor
University of Texas at San Antonio
Department of Interdisciplinary Learning and Teaching

**Dr. Igone Arteagoitia**
Research Associate
Center for Applied Linguistics (CAL)
Washington DC

**Dr. Kathy Escamilla**
Professor of Education
School of Education
University of Colorado at Boulder

**Dr. Gilda Evans**
Retired Assistant Superintendent over the Multi-language Department of Dallas ISD
Current Vice-president of Bilingual Education Association of the Metroplex (BEAM)

**Dr. Eugene E. Garcia**
Professor and Vice President, Education Partnerships
Arizona State University

**Kathleen Leos**
Former Assistant Deputy Secretary and Director to the US Department of Education's Office of English Language Acquisition (OELA) President and Co-Founder of Global Institute for Language and Literacy Development (GILD)

**Dr. William Pulte**
Associate Professor, Director of Bilingual Education Programs
Simmons School of Education
Southern Methodist University

**Dr. Luis Rosado**
Professor – College of Education
Director – Center for Bilingual Education
University of Texas at Arlington

**Lisa Saavedra**
Vice President and Co-Founder of Global Institute for Language and Literacy Development (GILD)
Former Bureau Chief for the Bureau of Academic Achievement through Language Acquisition for the Florida Department of Education

**Dr. Annette Torres-Elias**
Assistant Professor of Education
School of Education
Texas Wesleyan University

A Texas-based editorial firm, Tri-Lin, devoted to assessment, development, and special education services focusing on the Spanish and bilingual educational community, was contracted by Istation to write more than 5,000 items that make up the item bank for forms for ISIP Español.

Items were written by the Spanish test development staff. The items were required to follow specific rules for each domain, based on the ISIP Español assessment blueprint. The multiple-choice answer options were also driven by elimination rules specifications (rules for item creation and answer-choice elimination are available upon request). All items were originally written for use in this assessment; no items were translated or derived from any assessment delivered in English. In addition, all items underwent comprehensive analysis to ensure that no items contained linguistic or cultural bias and that all were age- and grade-level appropriate. Thus, the range of item types was extended to include items with difficulties as low as the end of Pre-K and as high as Grade 5/6. Additionally, items were developed within each domain to represent easy, moderate, and hard items for each grade. This wide range of items make ISIP Español an appropriate measure for the full range of students, including students with special needs or who struggle and students who are high-achieving or gifted. While ultimately the IRT calibration work identified the difficulty of each item, the team was assured of having items representing the full continuum of achievement for each domain.

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

## Accuracy and Fluency

Within ISIP Español, each subtest has both an accuracy component and a fluency component. Assessments that measure a student's accuracy and speed in performing a skill have long been studied by researchers. Such fluency-based assessments have been proven to be efficient, reliable, and valid indicators of reading success (Fuchs et al. 2001; Good, Simmons, & Kame'enui, 2001). Fluency in cognitive processes is seen as a proxy for learning, such that as students learn a skill, the proficiency with which they perform the skill indicates how well they know or have learned the skill. In order to be fluent at higher-level processes of reading connected text, a student will also need to be fluent with foundational skills.

Because each of the subtests has a fluency component, the tests are brief. This makes it feasible to administer the subtests on a large scale with minimal disruption of instructional time. Numerous items are

available for each subtest, making the subtests repeatable throughout the school year with many alternative forms.

## Teacher Friendly

ISIP Español 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 Español as experiencing difficulties as they complete ISIP Español. 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 Español battery or individual skill assessments.

## Child Friendly

ISIP Español is also child friendly. Each assessment session feels to a child like he or she is playing a fast-paced computer game called "A ver cuánto sabes" (Show what you know). In the beginning of the session, an animated owl enters the screen (named Don Buhiermo for Búho and Guillermo) that acts as a game show announcer and invites children to participate by saying, "¡Bienvenidos! En este juego vas a demostrar que ¡si puedes!" (It's time to show that you can do it!) The owl helps the children to understand the game rules, and then the assessment begins. At the end of each assessment, children see an animated graph of their progress. Each activity proceeds in a similar fashion.

# ISIP Español Subtests

ISIP Español measures progress in each critical component of reading instruction in a manner appropriate to the underlying domain.  There are a total of six subtests that align to the critical domains of Spanish reading, as shown in the table below. Of these subtests, four are built using a CAT algorithm, while two use parallel forms. Subtests that tailor items using CAT include Destreza fonológica y fonética, Vocabulario, Comprensión de lectura, and Comunicación escrita. Lectura con fluidez and Comprensión auditiva are designed as parallel forms that measure end of grade level expectations.

| Domain | Subtest |
|---|---|
| CONCIENCIA FONOLÓGICA<br>*Phonemic Awareness* | Destreza fonológica y fonética<br>*Phonemic and Phonological Awareness* |
| CONVERSIÓN GRAFEMA-FONEMA<br>*Grapheme-phoneme conversion* | Comunicación escrita<br>*Written Communication* |
| VOCABULARIO<br>Vocabulary | Vocabulario<br>*Vocabulary* |
| COMPRENSIÓN<br>*Comprehension* | Comprensión auditiva<br>*Listening Comprehension*<br><br>Comprensión de lectura<br>*Reading Comprehension* |
| LECTURA CON FLUIDEZ<br>*Fluency* | Lectura con fluidez<br>*Text Fluency* |

# ISIP Español Administration Format

ISIP Español is presented to students using a game-like format. Students are never told that they are being given a test. Instead, they are told that they are playing a game called "A ver cuánto sabes" (Show What You Know).


The first time a student takes ISIP Español, the computer will administer items that are defaulted based on the student's grade, unless the default setting is changed intentionally, as may be appropriate in special education settings. From the very first item, however, the CAT engine immediately begins to tailor the test to

the individual student. As a result, students will only be administered subtests that are appropriate for their performance abilities. Within a classroom, students may have some variation in the exact subtest they are administered. However, scores reflect these differences (explained below). For example, students whose performance scores indicate that they are not yet reading words will not be asked to read connected text. Likewise, students whose performance scores indicate that they read connected text fluently, and with comprehension, will not be asked to complete letter knowledge and phonemic awareness tasks.

Listening Comprehension is administered only in Pre-K and Kindergarten. In Grade 1, Text Fluency is administered only after students obtain a high enough overall reading score to suggest that they can handle the task. Lectura con fluidez is administered to all students, beginning in Grade 2.

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

| Grade | Subtest |
|---|---|
| Pre-Kindergarten | Destreza fonológica y fonética<br>Vocabulario<br>Comprensión auditiva |
| Kindergarten | Destreza fonológica y fonética<br>Vocabulario<br>Comprensión auditiva<br>Comunicación escrita |
| 1st Grade | Destreza fonológica y fonética<br>Vocabulario<br>Comprensión de lectura<br>Lectura con fluidez<br>Comunicación escrita |
| 2nd and 3rd Grade | Destreza fonológica y fonética<br>Vocabulario<br>Comprensión de lectura<br>Lectura con fluidez<br>Comunicación escrita |

## Rationale for Subtest Defaults by Grade

Children acquire the skills that they need to become proficient readers during the first years of school. These skills may be introduced and monitored separately, but teachers need to practice integrating these skills during daily reading routines as often and quickly as possible. Critical early reading skills are emphasized according to the grade level and the developmental stage of the child; however, daily practice of identified critical domains such as phonemic awareness, vocabulary, comprehension, and fluency is highly desirable (Vaughn & Linan-Thompson, 2004).

Based on research findings from the National Reading Panel (NRP, 2000), instruction in phonemic awareness is emphasized in Kindergarten and is recommended for about 15 minutes a day (Vaughn & Linan-Thompson, 2004). The teaching of phonemic awareness has expanded through many countries (particularly Spanish-speaking countries) by introducing instructional methodologies that allow students to manipulate phonemic and syllabic sounds of spoken words. Such methods involve teaching and practicing blending, segmentation of sounds, and identification of sounds that represent sounds in speech with or without the use of letters (NRP, 2000).

Early literacy instruction that incorporates decoding and word study provides a strong foundation for emergent literacy. Therefore, related skills such as grapheme-phoneme correspondence and print awareness can be introduced as early as Kindergarten.

Based on transparency of language studies that place Spanish as a language with shallow orthography (Wimmer & Mayringer, 2001; Ziegler, Perry, Ma-Wyatt, Ladner, & Körne, 2003; Sprenger-Charolles, Béchennec, 2008), children learn these skills rather quickly, and it is important to integrate these emergent reading skills with reading comprehension questioning. Taylor et al., (2002) found that children in first grade grew more in comprehension and fluency when their teachers asked more high-level questions.

Once students acquire a solid foundation in word recognition and decoding, fluency instruction should be emphasized. English literacy studies have shown that this usually begins during the second semester of first grade (Vaughn & Linan-Thompson, 2004). For fluency instruction in Spanish, the transparency of the language must be taken into consideration. Fluency development may begin as early as Kindergarten, based on the fact that Spanish phoneme combinations can be represented in only 45 variations (Sprenger-Charolles, Béchennec, 2008). Spanish grapho-phonemic conversions also make writing development equally accessible and rapidly attained. Tasks that require knowledge of grapheme-phoneme and symbol sound correspondence can be evaluated through students' spelling and dictation performance. Studies which examine the productive and receptive syntactic skills of Kindergarteners also show correlations with success in reading (Ballantyne, Sanderman, D'Emilio, & McLaughlin, 2008). Research has also shown that learning to spell and learning to read rely on much of the same underlying knowledge, such as the relationships between letters, letter units, and sounds. Knowing the spelling of a word makes the representation of it sturdy and accessible for fluent reading (Ehri, 2000 ; Snow et al., 2005).

Teaching vocabulary supports reading comprehension and increases speed, thus improving fluency. First and second grade vocabulary teaching impacts reading development, and the instructional methodology must incorporate familiar language in order to take advantage of the word superiority effect (Cattell, 1986). Students need to be able to identify with and connect to the reading material in order to be interested in it. Linguistic considerations are important, but cultural relevance may be a determining factor in assessing students' reading success (Skiba, Simmons, Ritter, Kohler, & Wu, 2003). When stories are interesting and written in simple language, they are very likely to encourage struggling students to persevere (Abadzi, 2006).

Successful reading development is also associated with small-group instruction. Hierarchical Linear Modeling (HLM) has been used to analyze classroom variables and compare outcomes. These observations confirm that teachers who engage often in small-group instruction have students who demonstrate more gains in fluency, comprehension, and vocabulary.

The subtests and domains of ISIP Español have been developed based on the sequence and frequency of the critical areas of reading identified in Kindergarten through Grade 3. Additionally, ISIP Español is equipped with research-based downloadable teacher resources that support formative assessment and small-group instruction. ISIP Español items were written for Hispanic students in bilingual education programs whose linguistic and cultural aspects were taken into consideration.

Measures for each subtest are described below:

**Beginning Sound**
Beginning Sound is a measure of phonemic awareness that assesses a child's ability to recognize the initial sound in an orally presented word. This skill is tested in Kindergarten and Grade 1, as aligned to the Spanish Language Arts and Reading Texas Essential Knowledge and Skills (SLAR TEKS) standards, and the resulting score is factored in with other skills under the same domain: Destreza fonológica y fonética.

**Blending**
Blending is a measure of phonemic awareness that assesses a student's ability to blend syllables and phonemes that make up spoken words. This skill is tested in Kindergarten and Grade 1, as aligned to SLAR standards, and the resulting score is factored in with other skills under the same domain: *Destreza fonológica y fonética.*

**Letter Sound**
Letter Sound is a measure of alphabetic principle that assesses how many letter sounds a student can correctly identify. Item selections for this portion of the assessment represent a combination of both upper and lower case letters, including vowels and consonants. This skill is assessed in Kindergarten and Grade 1, as aligned to SLAR standards, and the resulting score is factored in with other skills under the same domain: Destreza fonológica y fonética.

**Symbol Sound**
Symbol Sound is a measure of symbol conversion based on auditory input that combines letter units (syllables), as opposed to single phonemes (letters). Item selections for this portion of the assessment include the following syllable types: opened (CV), closed (CVC), consonant combination (CCV), and vowel combination (VV), and items are presented as they apply to each grade level expectation in Kindergarten and Grade 1. The resulting score is factored in with other skills under the same domain: Destreza fonológica y fonética.

### Vocabulary

Vocabulary is a measure of a student's knowledge of two types of vocabulary words: (1) oral vocabulary, or "common" words, which are primarily used in daily social interactions according to each developmental age (grade-level appropriate) and (2) academic vocabulary, or "meaning" words, which are frequently encountered in text but not typically used in daily conversation (Beck, McKeown, & Kucan, 2002 ). In particular, this second evaluation target contains items that were developed to assess students' knowledge of specific Spanish language elements that support understanding of meaning, such as word association, word derivatives, word roots (prefixes/suffixes), and synonyms. These two types of vocabulary words are evaluated separately in all grade levels, Kindergarten through Grade 3. The two scores (oral and academic vocabulary) are combined into a single score and reported as a vocabulary result in this domain: Vocabulario.

### Listening Comprehension

Listening Comprehension is a measure of a student's ability to listen and retain enough information in his or her working memory to be able to recall simple facts. This skill is tested in Kindergarten and Grade 1, as aligned to SLAR standards. It presents one narrative passage and one expository passage, followed by short answer questions that use a similar pattern to the reading comprehension battery. The resulting score is reported independently under the domain with the same name: Comprensión auditiva.

### Comprehension

Comprehension is a measure of a student's ability to read and understand grade level appropriate narrative and expository passages. According to the NRP research (NRP, 2000), text comprehension is a complex cognitive process that incorporates all foundational reading areas, including vocabulary and fluency. In order to assess a student's ability to comprehend the passages, this type of evaluation requires an intentional and thoughtful interaction between the student and the computer screen where the passages are presented. Students in Kindergarten are able to listen to the answer choices associated with a picture before they select their answer. Students in Grades 1 through Grade 3 are able to apply reading comprehension strategies to enhance understanding. The passage appears on the screen, and the student prompts the computer to begin the questions. Once the questions begin, the passage moves to the left side of the screen, and each question changes after 50 seconds to avoid inactivity. The questioning design is similar to the multiple-choice patterns used in state criterion referenced tests that combine explicit and implicit answers as they apply to grade-level requirements aligned to SLAR standards. The resulting score is reported independently under the domain with the same name: Comprensión de lectura.

### Fluency

Fluency is a measure of a student's ability to read fluently with comprehension. This subtest is constructed in a very different manner than others, using grade-level, culturally-appropriate passages. Each of these passages was carefully written to conform to specific word level features, follow linear story grammar structure, and have readability according to a commonly accepted readability formula for end-of-grade level expectations as it applies to Spanish fluency. (It uses the middle to mid-high ranges of the Spanish literacy fluency chart, in Table 3). In order to assess text reading on the computer, a maze task is utilized in which

every seventh word is left blank, with a three-word menu of choices to complete the sentence. This task has been shown to be highly correlated to measures of both fluency and comprehension, and it has high reliability and concurrent validity (Espin, Deno, Maruyama, & Cohen, 1989; Fuchs & Fuchs, 1990; Jenkins, Pious, & Jewell 1990; Shinn, Good, Knurson, Tilly, & Collins, 1992). As opposed to fluency measures for Grade 1, in which the teacher relies on oral/observable measures, students in Grade 2 and 3 would be more accurately assessed using tools that register receptive reading skills. Fluency is tested in Grade 2 and 3, and the resulting score is reported independently under the domain with the same name: Lectura con fluidez.

**Spelling and Dictation**

Spelling and Dictation is a measure designed to determine if students are developing fully specified orthographic representations of words. Items were designed following different rules, depending on the grade level assessed. For Grade 1 students, a word is given and an array of syllables appears on the screen, with which the student spells the word. Grade 2 and 3students use individual letters to spell the words. Items for the dictation subtest follow a similar functionality. Students choose, from a word bank, the necessary word to complete a sentence. These items have been carefully constructed to move from easier to harder, using a sequence of difficulty aligned to SLAR standards. Additionally, students in Grade 1 through Grade 3 are required to select correctly spelled words that exemplify commonly used accented patterns for word categories such as palabras llanas, graves, agudas, and esdrújulas. The scores for each subtest are factored together into a single score under the domain Comunicación escrita.

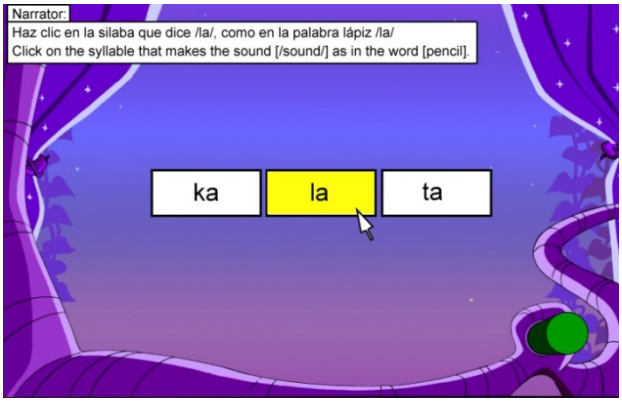
# Description of Each Subtest

## Destreza fonológica y fonética

The Destreza fonológica y fonética subtest is comprised of several types of items:

**Conversión grafema-fonema** items measure the students' ability to identify symbols that correspond to specific sounds of the Spanish language: letras *(letters),* sílabas *(syllables),* combinaciones vocálicas *(vowel combinations),* grupos consonánticos *(consonant clusters),* and palabras *(words).* The computer presents items representing various upper- and lower- case letter combinations. Four boxes appear on the screen, and only one choice contains the correct answer for each item. The narrator asks students to click on a particular grapheme (letter, syllable, etc.) that represents a sound produced orally by the narrator.

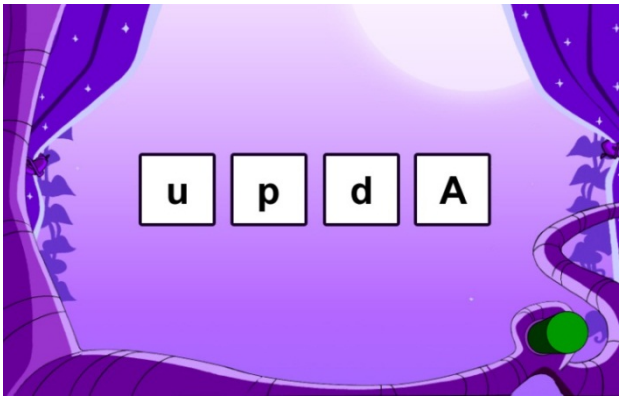
Screenshot examples:

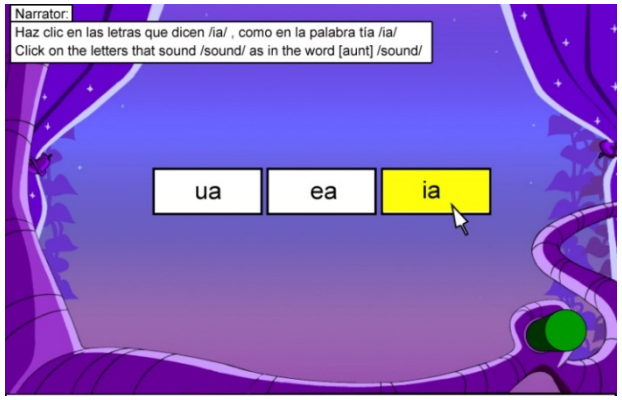




**Conciencia fonética y silábica** (*phonemic and syllabic awareness*) items measure the students' ability to identify single sounds (letter or syllable) in grade-level appropriate words. The level of difficulty adapts with the student response. Students identify beginning sounds and use syllables or letters to find words.

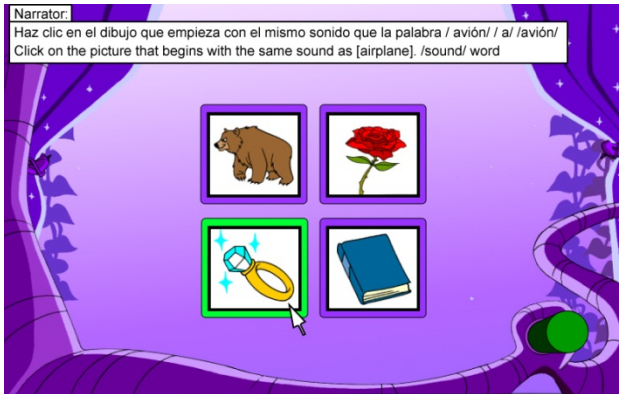
Screenshot examples:




**Unión de sonidos** *(blending)* and **sonido inicial** *(beginning sounds)* are items presented independently. First students find the beginning sound of a word following the narrator instructions. The name of each picture is given as these appear on the screen. Each box is highlighted while students are asked to click on the picture that has the same beginning sound as the sound is produced orally by the narrator. For blending items, a box appears in the middle of the screen containing an animated side view of a head that pronounces the sounds. Once the word is said by pronouncing each phoneme or syllable, the student is asked to click on the picture that shows the word that has been spoken using only sounds.

Screenshot examples:

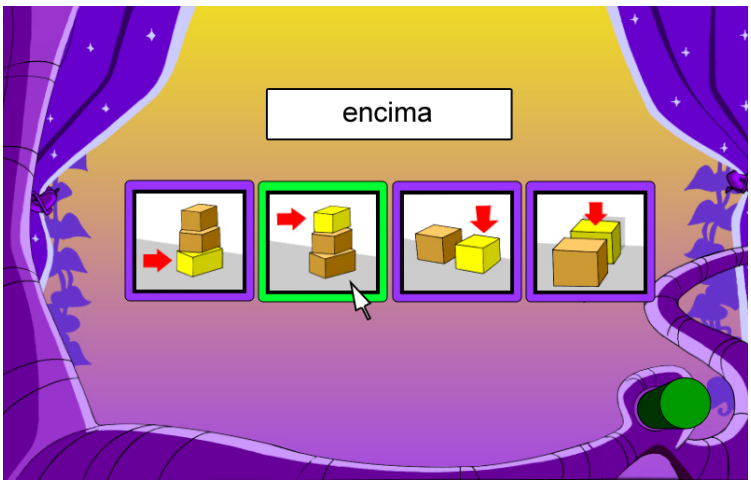

## Comprensión auditiva

**Comprensión auditiva** (*Listening Comprehension*) is a subtest used to evaluate children's ability to listen, understand, and answer questions related to a story that is presented orally. In this activity, a picture related to a short story appears on the screen. The narrator reads aloud with no text present on the screen. The narrator then asks the student a question related to the story. From the four pictures that appear on the screen, the student chooses the one that best answers the question.

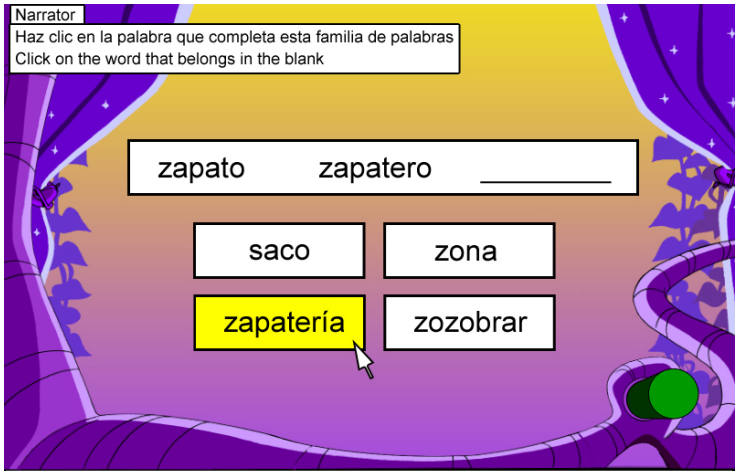
## Vocabulario

The Vocabulario subtest is comprised of several types of items:

**Vocabulario de lenguaje oral** (*Oral Vocabulary*) items measure a student's vocabulary knowledge. In this subtest, four pictures appear on the screen. The narrator asks the student to identify the picture that best illustrates the word spoken orally.

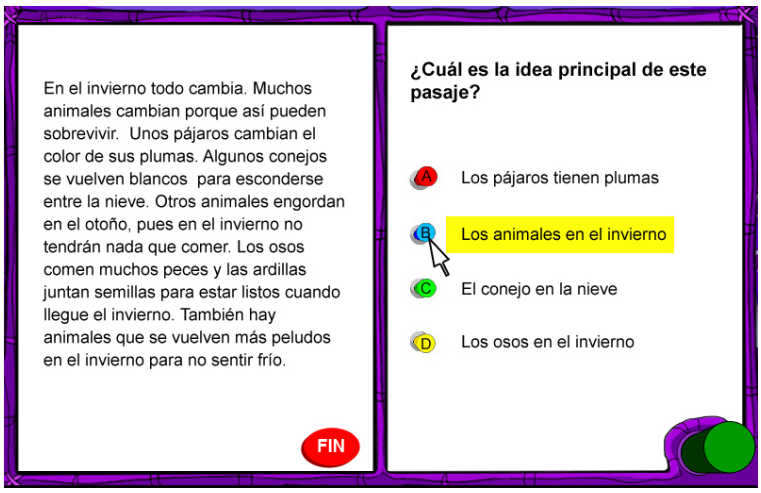



**Vocabulario para lectura y escritura** (*Reading Vocabulary*). For these items, a combination of word strategies (i.e., knowledge of roots, prefixes, and suffixes) is assessed using both pictures and words that appear in sets of four on the screen. The questions spoken by the narrator cover word knowledge such as familias de palabras, clasificación de palabras, sinónimos, etc. (*synonyms, classification of words, derivatives, etc.*) After instructions are given, the student is asked to identify each word accordingly.



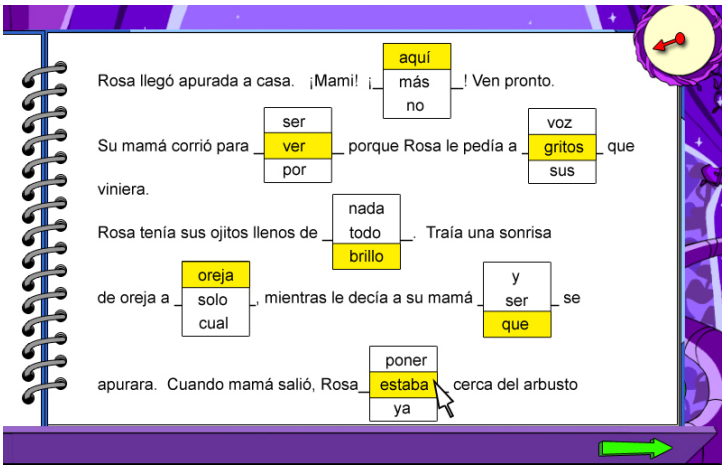

## Comprensión de lectura

**Comprensión de lectura** (*Reading Comprehension).* In this subtest, students are assessed on entendimiento de lo leído (*their ability to read and understand sentences and texts*). This is accomplished using evidential/inferential question patterns to evaluate both narrative and expository texts. The item types that measure reading comprehension and thinking skills are linked to criterion-referenced tests. In this task, a passage appears on the screen. The student indicates when he or she is finished reading by clicking on a green button. After this button is clicked, questions populate on the right side of the screen. The student is able to read the text as often as needed while choosing an answer from among four choices. Kindergarten students select from pictures that represent each answer choice, and these are read by the narrator and repeated as needed.




## Lectura con fluidez

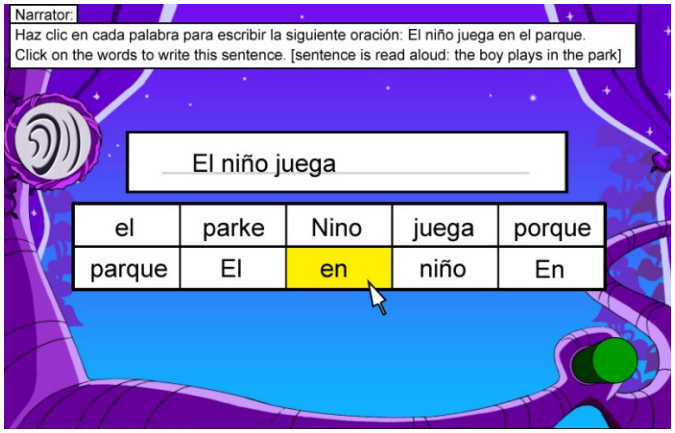
**Lectura con fluidez** *(Text Fluency)* is a subtest constructed in a very different manner than the other subtests. Students are assessed on their skill in reading text with meaning in a specified period of time. In order to assess text reading on the computer, a maze task is utilized in which every seventh word of a grade-level story is left blank from the text. The student is given three choices for each blank from which to choose the word that makes the most sense to complete the sentence. It is the student's job to read the text and select the correct maze responses in two and one-half minutes. This task has been shown to be highly correlated to measures of both leer un texto and tener precisión *(fluency and accuracy).*



# Comunicación escrita

**Ortografía y acentuación de palabras** *(Spelling and accent marks)* is a subtest that determines if students are developing fully specified orthographic representations of words. For each item, an array of letters appears on the screen and the computer asks the student to spell a specific word using those letters and their proper tildes (*accent marks).* In Grade 1, the same objective is achieved using syllables to write the word said by the narrator. The student then spells the word by clicking on each letter/ syllable. As each letter/syllable is selected, the word is formed on a line that is directly above the letter array.


**Dictado** (*Dictation*) the items in this subtest are designed to determine if students are using correct sentence structure: sujeto + verbo + predicado (*subject* + *verb* + *predicate*). For each item, an array of words appears on the screen and the computer asks the student to put together a specific sentence using the words available.



# The ISIP Espanol Link to Instructional Planning

ISIP Español provides continuous assessment results that can be used in recursive assessment-instructional decision loops. Initially ISIP Español identifies students in need of support. If validation of student results is needed re-administering the assessments can increase the reliability of the scores. The technology underlying ISIP Español delivers real-time reports on student progress immediately upon assessment completion. This data facilitates the evaluation of curriculum and instructional plans. 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 Español automatically alerts teachers to children in need of instructional support through the "Priority Report." Students are grouped according to instructional level and skill need. Links are provided to teacher-directed plans for each instructional level and skill category. There are downloadable lessons and materials appropriate for each level of instruction.

A complete history of Priority Report notifications, including those from the current year and all prior years, is maintained for each child. On the report, teachers may acknowledge that suggested interventions have been provided. A record of these interventions is maintained with the student history as an Intervention Audit Trail. This history can be used for special education Individual Education Plans (IEPs) and in Response to Intervention (RTI) or other models of instruction that require modifications of 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. These downloadable, printable lessons support small-roup instruction through scripted lessons. These teacher-directed lessons are based on student individualized needs per the Priority Report. As the lessons are taught, teachers document intervention delivery on the Priority Report. This provides a visual reference of teacher intervention and its effectiveness. The ease of identification of skill needs and readily available lessons facilitates intervention and puts instructional time back in the classroom.

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

A year-to-year history of ISIP Español results is available. Administrators, principals, and teachers may use their reports to evaluate and modify curriculum. Interventions, AYP progress, effectiveness of professional development, and personnel performance may also be correlated to the growth model depicted from the reports.

# Chapter 2: IRT Calibration and the CAT Algorithm of ISIP Español

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 2010-2011 school year, data were collected from Kingdergarten to Grade 3 students in six states. However, most of the testing was in Texas elementary schools. The testing was conducted in 37 school districts, covering 228 schools. Among those, 30 schools districts and 217 schools were in Texas. Table 2-1 shows number of studetns and the demographics of participating students.

Table 2-1: Demographics for Participating Students

| Demographic | n | % |
|---|---|---|
| Total Number of Students | 3,895 | |
| Gender | | |
| Male | 1,818 | 46.48 |
| Female | 1,729 | 44.39 |
| Missing/Unidentified | 348 | 9.13 |
| Enrolled in Special Ed. | | |
| Yes | 241 | 6.19 |
| No | 3,210 | 82.41 |
| Missing/Unidentified | 444 | 11.40 |
| Economic Disadvantage | | |
| Yes | 2,841 | 72.94 |
| No | 610 | 15.66 |
| Missing/Unidentified | 444 | 11.40 |
| English Proficiency | | |
| Non-English Speaker | 1,421 | 36.48 |
| Fluent English Speaker | 2 | 0.05 |
| Limited English Speaker | 2,034 | 63.47 |

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-minutes sessions on the ISIP Español program.

It was unrealistic to administer all the items to each student participating in the study. Therefore, items were divided into grade-specific subpools. Each participant was administered all of the items in the subpool for their grade level. Originally, 2,751 items were tried out. Table 2-2 shows the numbers of items in each grade subpool, not including the 10% overlap items.

Table 2-2: Items Used in Study

| | Grade | | | |
|---|---|---|---|---|
| | K | 1 | 2 | 3 |
| Comprensión de lectura | 155 | 135 | 136 | 201 |
| Escritura | - | 104 | 104 | 169 |
| Fonología y fonética | 419 | 471 | - | - |
| Vocabulario | 164 | 170 | 295 | 228 |

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.

# 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 $\theta_i$ 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 ($\theta$) 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.

Among 2,751 items, 2,419 items were within the desired range of the item difficulty (-3.50, 3.50). 331 items fell below the desired range of the item discrimination (greater than 0.50). Therefore, 2,088 items were used for the ISIP Español item pool.

Overall, most items are in good quality in terms of item discriminations and item difficulties. The reliability was computed from IRT perspective by using this formula; $\rho^2 = 1 - [\mathrm{SE}(\theta)]^2$, where $\theta$ is the student ability. It is 0.850, indicating that ISIP Español is very reliable. The standard error of measurement (SEM) was also computed from IRT point of view. Since the ISIP Español scale score is $(20 * \theta) + 200$, $SEM(\theta) = 20 * \mathrm{SE}(\theta)$. It is 7.748.

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

## Stopping Criteria

ISIP Español has a stopping criterion based on minimizing the standard error of the ability estimate.

### Production Assessment

Item types were grouped according to key reading domains for the production assessment. Each grade level is given the same set of subtests.

These subtests are administered sequentially and treated as independent CAT tests. Items are selected from the full, non-truncated, item pool for each subtest, so students are allowed to demonstrate their ability

regardless of their gradelevel. Each subtest has its own ability estimate and standard error, with no crossing between the subtests. After all subtests are complete, an overall ability score is 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 2,088 Grades K-3 items with reliable parameter estimates aligned on a common scale, with the majority of items ranging from 140 to 289 in difficulty.

After completing this study, which included determining an appropriate IRT model, calibrating the items, and constructing the CAT algorithm, the ISIP Español assessment went into full production starting in the 2012-2013 school year.

# Chapter 3: Reliability and Validity of ISIP Espanol for Grades K–3

## ISIP Español Validity Framework

The Istation ISIP Español assessment is designed to be a criterion-referenced assessment to measure specific skills in early Spanish literacy. It has been developed to be used for formative purposes and progress monitoring. These purposes will be supported to the degree to which the criterion-referencing is supported by the evidence. Aspects of the development of ISIP Español and proposed claims from test scores are described in previous chapters of this document.

To support further development of the assessment and begin the validation process, this chapter describes the following:

1. A summary of the validity as argument framework employed to provide validity evidence
2. A summary of the proposed claims from ISIP Español
3. A description of the ISIP Español pilot validity study, preliminary results, and ongoing analyses

## Validity and Validation

Current definitions of validity vary across fields. However, in educational testing, most agree with the framework described in the *Standards for Educational and Psychological Testing* (hereafter referred to as *Testing Standards*; AERA, APA, NCME, 1999). “Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (AERA, APA, NCME, 1999, p. 9). The *Testing Standards* describes validation as the process of gathering evidence to achieve these goals, including evidence related to

- The construct
- Test content
- Response processes
- Internal structure
- Relations to other variables
- Intended and unintended consequences

In all cases, validation is an ongoing process, and the most important sources of validity evidence are those that are most closely related to the immediate inferences and proposed claims we make regarding test results. What evidence do we need to support the intended meaning of ISIP Español results?

## Validity as Argument

The argument approach to validation (Kane, 1992, 2006a, 2006b) is a contemporary approach that does not rely on our ability to define a construct or specify relations between the construct of interest and other important constructs. The heart of this approach is to make explicit arguments regarding proposed interpretations and uses. This is accomplished through an interpretive argument that specifies the inferences and assumptions leading from the test scores to the interpretations and decisions generated (Kane, 2006a). The validation process must evaluate and articulate the interpretive argument, specifying the reasoning from the score to the intended conclusions and the plausibility of the associated inferences and assumptions. The validity argument provides not only an evaluation of the proposed interpretations and uses of scores, but also alternative interpretations.

The forms of validity evidence described by the *Testing Standards* can be used in the validity argument framework; these include claims, intended inferences, and assumptions. These forms of evidence can be gathered to support the validity argument. These conceptualizations of validation are complimentary, providing the strongest approach to securing evidence to support that a measure is appropriate, meaningful, and useful.

## The Interpretive Argument

The first component of this process is clarifying the interpretive argument. This component frames the validation efforts by identifying the issues that need to be addressed. As Kane (2006b) describes, the interpretive argument provides three critical elements: (a) a framework to allow for the test development process to accommodate important assumptions and requirements that can be met in the design process, (b) a framework to clarify the validity argument by identifying the inferences and assumptions requiring evaluation, and (c) a framework for evaluating the validity argument by specifying the questions that need to be addressed. There are four important elements of the interpretive argument:

1. The conclusions and decisions to be made from test scores
2. The inferences and assumptions leading from test scores to the conclusions and decisions
3. The potential competing interpretations
4. Evaluative evidence for and against the proposed interpretive argument

# Proposed Claims of ISIP Español

An important tool in the specification of the interpretive argument is the clarification of the conclusions and decisions to be made and the intended inferences and associated assumptions. This includes identifying the proposed claims we hope to make based on ISIP Español results. The following list of proposed claims is derived from the evidence provided by the functionality and development of the English version of ISIP and potential reconditioning for use with reading objectives in other languages. These considerations include the following: (a) the primary objective of this version was to determine how to accurately measure,

on the computer, early reading skills known to be predictive of later reading success (Mathes, & Torgesen, 1996-1999); (b) the functionality of prototype tasks created for each subtest using a technology-based platform; and (c) the possibility to compare technology-based assessments to paper and pencil evaluations of a similar construct (ISIP was compared against CTOPP, TOWRE, and DIBELS ORFA). A number of these claims have been addressed throughout the development of ISIP Español, and some claims were addressed during the 2010-2011 pilot validation study. We recognize that validation is an ongoing process, and therefore we will continue to evaluate results in 2014 and beyond, as described in Chapter 6 of this document.

## Claims Regarding Spanish Early Literacy Evaluation

1. To measure achievement of selected Spanish language arts standards (domains), focusing on reading.
2. To measure whether specific early literacy skills in Spanish (subtests) have been achieved.
3. To measure students' knowledge, skills, and performance level in the domains of Spanish reading that apply to each grade level, including:
    - Phonemic awareness and grapheme/sound correspondence
    - Oral language and listening comprehension
    - Vocabulary
    - Reading comprehension thinking skills
    - Written communication
    - Text fluency
4. To determine the progress students make in Spanish reading instructional programs (also described as progress monitoring of Spanish foundational reading skills).

The assumptions regarding the degree to which domains are appropriately defined and to which the item types are appropriate measures of the knowledge, skills, and abilities targeted in each domain are included in Chapter 1 of this document.

There is an assumption regarding the representativeness of the models, based on Spanish Language Arts and Reading (SLAR) standards, which suggests that they provide an appropriate context for the intended purposes of ISIP Español.

More specifically, there is an assumption that standards and domain definitions, based on SLAR standards from selected states and countries (i.e., California, Texas, WIDA consortium, Puerto Rico, Colombia, Mexico, and Spain), are appropriate for the proposed purposes and the intended population. This research-based evidence is thoroughly presented and reviewed in the description, domains and item development sections in Chapter 1 of this document.

### Claims Regarding Conclusions and Decisions

1. To report domain scores with sufficient precision to warrant independent scoring and reporting of distinct aspects of Spanish literacy.
2. To be functional for the purposes of formative assessment.
3. To determine whether students are progressing toward end-of-year expectations and achieving selected skills for each domain.
4. To provide information on how groups of students are progressing toward achieving grade level expectations in each domain.

These claims depend on the plausibility of the first four claims and associated assumptions regarding content, described above. In addition, there are assumptions about the appropriateness of performance standards (end-of-year expectations and achieving expected levels of performance).

# The ISIP Español Pilot Validity Study

Istation is completing a validity and reliability study during the 2010-2011 school year. The scope of the validity and *reliability study is aimed at answering the following questions: (1) Do the scores produced by* ISIP Español show evidence of reliability, including internal consistency, and alternate form reliability? (2) Do the scores produced by ISIP Español show evidence of validity, including concurrent validity and predictive validity? (3) Do the scores produced by ISIP Español show evidence of accurate classification, as established by ROC analysis? The development of a technical report will be completed once the validity study is finished. The first part of the study, which includes content validity analysis, has been completed as described in this chapter.

In addition to the validity and reliability study, a concordance analysis also has been conducted, whereby the results of students assessed on ISIP Español were compared to results on other external measures obtained by the same group of students. The external measures selected for this analysis include Evaluación del desarrollo de la lectura (EDL2, Pearson), *Téjas* LEE (Brookes) for Grades Kindergarten through Grade 2, and Texas Assessment of Knowledge and Skills in Reading (TAKS Reading, TEA). This part of the research will not be completed until after results are obtained from 2011 administration of the TAKS and all data have been analyzed.

# Core Elements of the Validity Study, Evidence, and Analyses

It is important to recognize that validation is an ongoing process. Validity is not something that is present or absent; validity is the accumulation of evidence that supports ongoing and current (and new) interpretations and applications of test results. The analyses will be described in terms of the inferences that are important for a standardized measure of observable attributes, in this case ISIP Español. These four inferences (scoring, generalization, extrapolation, and implication) will provide the framework that encompasses all elements of the current validation design, analyses, and anticipated results (see Technical Note on Content-Related Validity Evidence, following this section).

**Scoring (supporting the inference from observations of performance to an observed score)**

Since the test is a standardized, objectively-scored instrument, the inference regarding scoring that supports the content-validity argument of ISIP Español is relatively easy to achieve. [Claim 5]

- Key confirmation
- Internal consistency (reliability and score accuracy), item discrimination

***Generalization*** **(moving from an observed score on a sample of tasks to an expected score on the universe of generalization)**

Since the test is based on relatively narrow subdomains with relatively homogenous items, this inference is fairly direct. [Claim 1]

- Alternate form correlations
- Documentation of construct representation
- Item discrimination

***Extrapolation*** **(moving from the universe score to the target score)**

Since the test is based on a broad range of language arts and reading standards, the degree to which the domain is covered by the represented skills is more challenging to support. [Claims 2, 3, 6]

- Correlations with criterion variables in same target domain (EDL; *Téjas* LEE, administered 3 times during year; TAKS, the spring state standards-based assessment)
- Documentation of content coverage, given review of existing standards

***Implication*** **(the translation of the estimated target score into a description of knowledge, skills, or ability level).**

The test is designed for multiple purposes, so the clarity of the attribute being measured and how these can be described is critical. [Claims 4, 6, 7, and 8]

- Performance level descriptors
- Instrument development process
- Agreement among users and relevant content experts
- Standard error of measurement, score precision, and test information functions

## Analysis Methods

To support the analyses of test scores from the ISIP Español pilot study, based on the relevant inferences described above, the following analyses of pilot data were conducted or are planned:

- Winsteps was used to provide classical test theory item statistics, including classical item difficulty (item p-values, proportion correct) and discrimination (point-biserial item-total correlations).
- Winsteps was used to provide a Rasch analysis of item performance, including item fit measures.
- With the scores on the 13 forms administered to a subsample within grade, inter-form correlations and form difficulty (means) and variability (variance) were examined.
- To the extent possible from the larger sample, Mplus will be used in subsequent studies to conduct confirmatory factor analysis to test the degree to which forms are unidimensional (on forms with large enough samples) versus multidimensional, based on the reading subdomains.
- Finally, once criterion measures are administered and scored, correlations between ISIP Español forms and criterion measures will be assessed (with attention to intended alignment of constructs across measures). This will be accomplished after TAKS results are analyzed.

## Study Sample and Form Design

The sample was obtained through careful selection of students from El Paso Independent School District to include a full range of ability levels among Spanish speaking students in grades K through 3, including 219 students with valid responses, as shown in Table 3-1.

Table 3-1: Validation Sample Size by Grade and Percentage Female

| Grade | *n* | % Female |
|---|---|---|
| K | 52 | 56% |
| 1 | 56 | 43% |
| 2 | 52 | 46% |
| 3 | 59 | 54% |

For each grade level, 13 forms were constructed to cover parallel content. The 13 forms were administered online within three to four weeks, in random order across students.

Table 3-2: Skill Areas Assessed by Grade

| Grade | Skill Area | | | | |
|---|---|---|---|---|---|
| | Listening Comp | Phonological Awareness | Reading Comp | Vocabulary | Reading Fluency |
| K | ✓ | ✓ | ✓ | ✓ | |
| 1 | | ✓ | ✓ | ✓ | |
| 2 | | | ✓ | ✓ | ✓ |
| 3 | | | ✓ | ✓ | ✓ |

A total of 3,832 items were used across skill areas on the 13 forms.

### Scoring

Reading Fluency was assessed by a 90-second timed maze task and was scored with an algorithm that accounts for (a) number of tasks completed within the 90-second limit and (b) accuracy of responses. All other skill areas were scored in terms of percent correct.

# Score Reliability (Inferences Regarding Scoring, Implication)

Coefficient alpha is a typical form of reliability, which, under specific assumptions of the parallel measurement model, provides an estimate of item internal consistency. These coefficients are presented here for preliminary consideration only. In this context, results are promising. Test score reliability is a form of validity evidence, as it informs the precision of scores and supports related inferences.

## Rasch Model Reliabilities of the Item Pools

Another index of reliability comes from the Rasch analyses of each measure. In order to accomplish the Rasch analyses, all items across the 13 forms were combined to improve estimation of item functioning. This model considers the 13 forms to be samples of a larger pool of items, which is consistent with the future intent to create online adaptive forms. These estimates of reliability are based on (1) an estimate of true variance (model-based score variance), (2) a measurement error variance (based on the theoretical definition of reliability as the ratio between true-score variance), and (3) an observed-score variance (the proportion of observed variance that is true). Rasch analysis required larger samples, so the results presented in Table 3-3 should be interpreted strictly as preliminary.

Person Reliability, which is similar to traditional test-score reliability, indicates the capacity of the sample to generate a stable ordering of person abilities based on their test scores. Low person reliabilities among a pool of items suggest a high degree of randomness in responses (guessing).

Item Reliability, which has no traditional equivalent, indicates the capacity of the sample to generate a stable ordering of item difficulties.

Table 3-3: Person Reliability and Item Reliability

| Skill Area | Grade | *n* | # Items | Person Reliability | Item Reliability |
|---|---|---|---|---|---|
| Listening Comprehension | Kinder | 52 | 104 | .90 | .80 |
| Reading | Kinder | 52 | 104 | .48 | .88 |
| Comprehension | 1 | 56 | 130 | .83 | .74 |
| | 2 | 52 | 195 | .88 | .66 |
| | 3 | 59 | 260 | .89 | .77 |
| Reading | 2 | 52 | 585 | .92 | |
| Fluency | 3 | 59 | 503 | .94 | *[1] |
| Phonological | Kinder | 52 | 351 | .96 | .83 |
| Awareness | 1 | 55 | 298 | .96 | .78 |
| Vocabulary | Kinder | 52 | 155 | .88 | .85 |
| | 1 | 55 | 194 | .92 | .91 |
| | 2 | 52 | 364 | .96 | .89 |
| | 3 | 59 | 310 | .92 | .90 |

These reliabilities are not associated with individual form-based scores. They provide an index of the measurement quality of the pool of items in each area, based on this specific sample. They might also be interpreted as a potential score reliability, based on the current pools of items.

## Within Skill-Level Analyses across Forms (Inferences Regarding Generalization)

Each skill area was assessed by 13 forms designed to cover parallel content. There are multiple indicators with respect to evaluating the degree to which forms are parallel or similar in means, variances, and total scores. As an initial set of analyses, the correlations between forms for each Skill area by Grade, are provided in Table 3-4. However, as the plans for ISIP Español include computer adaptive testing (CAT)

[1] Item reliability for Fluency needs to be measured differently because this is the only timed subtest in the battery and student responded to ⅓ - ½ of the total available items.

administration, the degree to which forms are parallel will become less important. All items will be placed on the same common scale, making items within domains exchangeable. The results reported in Table 3-4 include the analyses across all 13 forms.

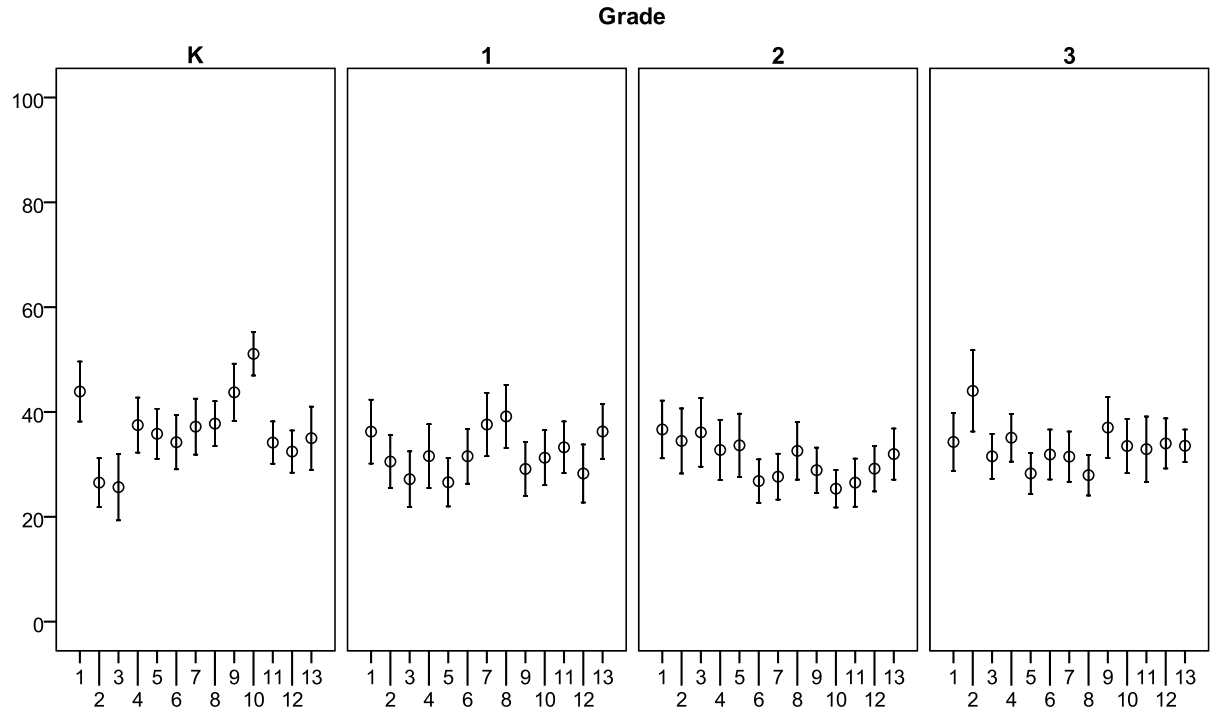
Another indicator of the association among forms (or the degree to which forms are measuring a similar construct) is the correlation among form scores and the average across all forms (excluding the target form score). This analysis is more rigorous since it requires a student to respond to all 13 forms, and consequently the sample has been reduced.

Table 3-4: Analyses across Forms, Including Students Responding to All 13 Forms
Reading Comprehension Corrected Correlations between Form and Average Score by Grade

| Reading Comp Form | Grade K N=19 | Grade 1 N=18 | Grade 2 N=27 | Grade 3 N=29 |
|---|---|---|---|---|
| 1 | .428 | .421 | .522 | .696 |
| 2 | .278 | .221 | .628 | .270 |
| 3 | .387 | .459 | .638 | .472 |
| 4 | - | .474 | .631 | .661 |
| 5 | .436 | .623 | .435 | .726 |
| 6 | .032 | .348 | .503 | .836 |
| 7 | .619 | .646 | .682 | .756 |
| 8 | .427 | .670 | .698 | .766 |
| 9 | .166 | .484 | .669 | .854 |
| 10 | .553 | .301 | .110 | .782 |
| 11 | .243 | .456 | .679 | .834 |
| 12 | - | .482 | .458 | .744 |
| 13 | .506 | .643 | .790 | .410 |

An example interpretation from this table is: The correlation between the Reading Comprehension Form 1 score and the average Reading Comprehension score across all other forms in Grade 3 (excluding Form 1) is .696.

Figure 5-A: Reading Comprehension Form Mean 95% Confidence Intervals by Grade




In this figure, the consistency in difficulty of forms can be seen graphically. Forms in Grade K show some degree of variability, compared to forms in other grade levels. In Grades 1 and 2, forms are within a narrow range of difficulty. In Grade 3, form 2 appears to be easier than some of the others.

The variability observed in the Reading Comprehension scores across forms in Kindergarten is likely due to the small number of items in each form (there are only 8 Reading Comprehension questions in each Kindergarten form). In measuring Reading Comprehension among Kindergarten students, results can be challenging to interpret due to students' ages and possible differences in instruction methods. Since the pilot was administered early in the fall, this may also indicate that Reading Comprehension could be more accurately measured beginning later in spring, not immediately at the beginning of the school year in the fall.

Reading Comprehension (among other grade levels) correlates quite strongly and fairly stably across forms, suggesting that the measures in each domain are strong across grades.

As is expected, Listening Comprehension scores are much more stable across forms for Kindergarten (See Table 3-5). These results coincide with instructional methodology and standards objectives for Kindergarten students in Texas.

Table 3-5: Listening Comprehension Corrected Correlations between Form and Average Score

| Form | Grade K N=18 |
|---|---|
| 1 | .532 |
| 2 | .711 |
| 3 | .108 |
| 4 | .696 |
| 5 | .417 |
| 6 | .753 |
| 7 | .619 |
| 8 | .747 |
| 9 | .824 |
| 10 | .793 |
| 11 | .783 |
| 12 | .786 |
| 13 | .768 |

Figure 5-B: Listening Comprehension Form Mean 95% Confidence Intervals by Grade

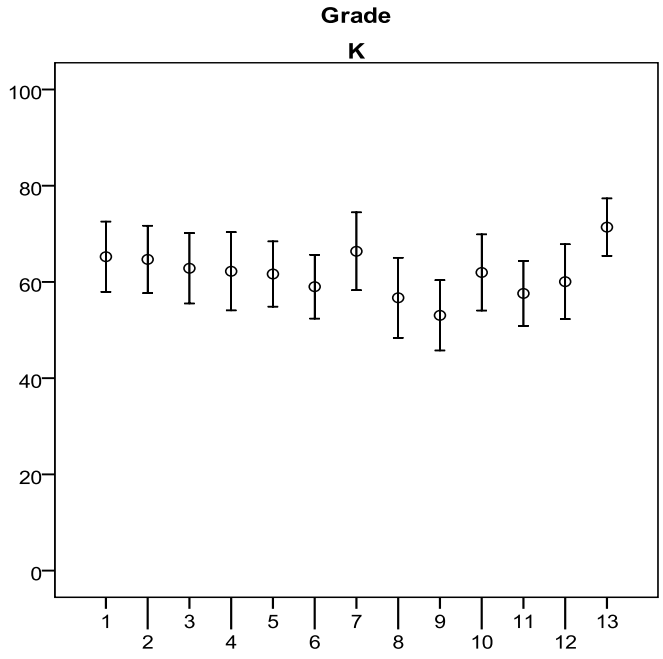


Table 3-6: Reading Fluency Corrected Correlations between Form and Average Score by Grade

| Form | Grade 2 N=27 | Grade 3 N=29 |
|---|---|---|
| 1 | .780 | .759 |
| 2 | .840 | .829 |
| 3 | .838 | .799 |
| 4 | .671 | .779 |
| 5 | .881 | .816 |
| 6 | .893 | .791 |
| 7 | .824 | .675 |
| 8 | .882 | .840 |
| 9 | .839 | .775 |
| 10 | .795 | .817 |
| 11 | .790 | .678 |
| 12 | .445 | .864 |
| 13 | .588 | .664 |

Figure 5-3: Reading Fluency Form Mean 95% Confidence Intervals by Grade

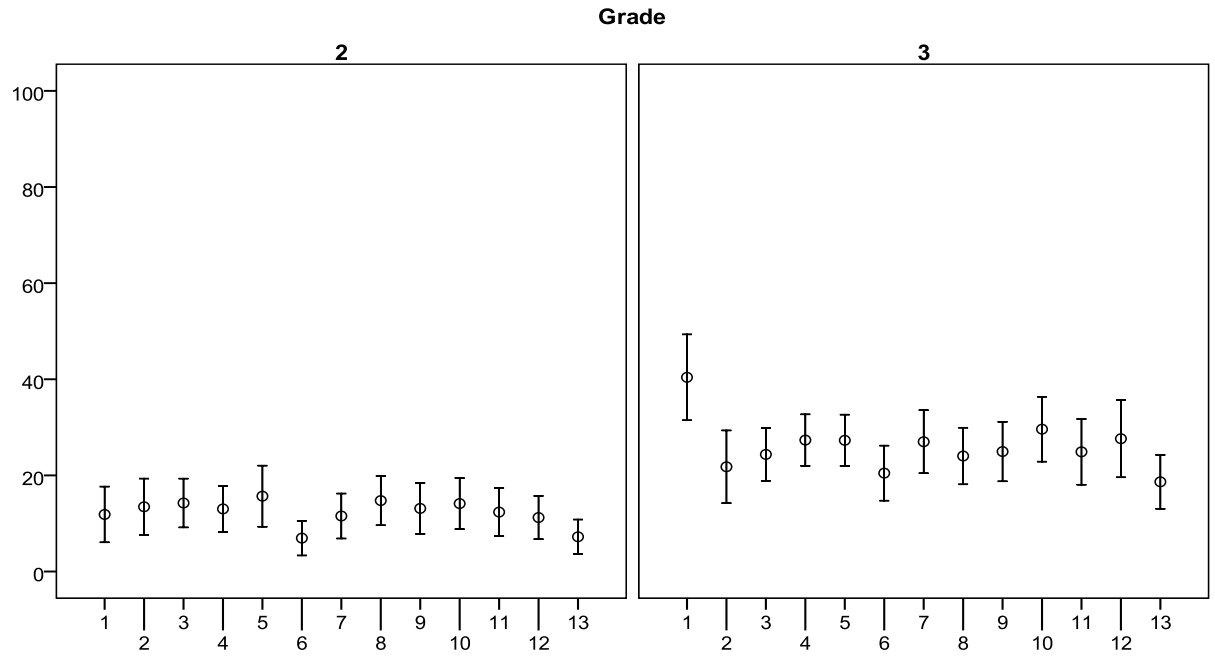


Table 3-7: Phonological Awareness Corrected Correlations between Form and Average Score by Grade

| Form | Grade K N=18 | Grade 1 N=19 |
|---|---|---|
| 1 | .913 | .860 |
| 2 | .964 | .898 |
| 3 | .765 | .932 |
| 4 | .739 | .929 |
| 5 | .727 | .855 |
| 6 | .684 | .825 |
| 7 | .916 | .786 |
| 8 | .890 | .891 |
| 9 | .876 | .821 |
| 10 | .882 | .801 |
| 11 | .921 | .885 |
| 12 | .824 | .867 |
| 13 | .928 | .876 |

Figure 5-4: Phonological Awareness Form Mean 95% Confidence Intervals by Grade

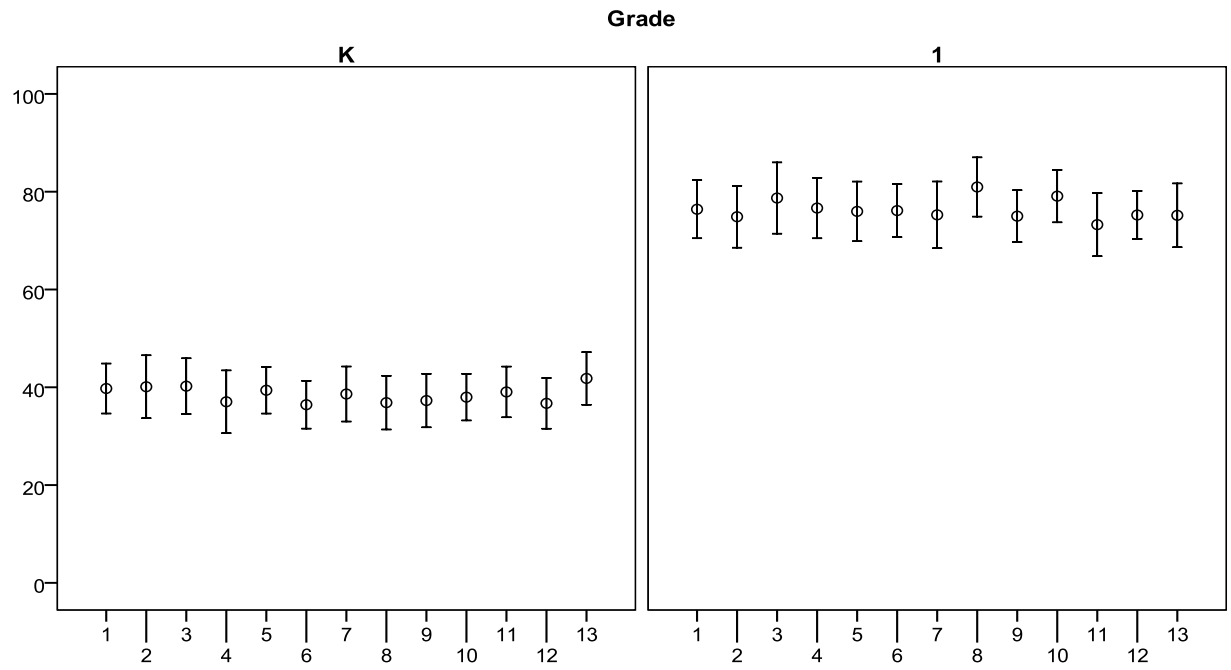
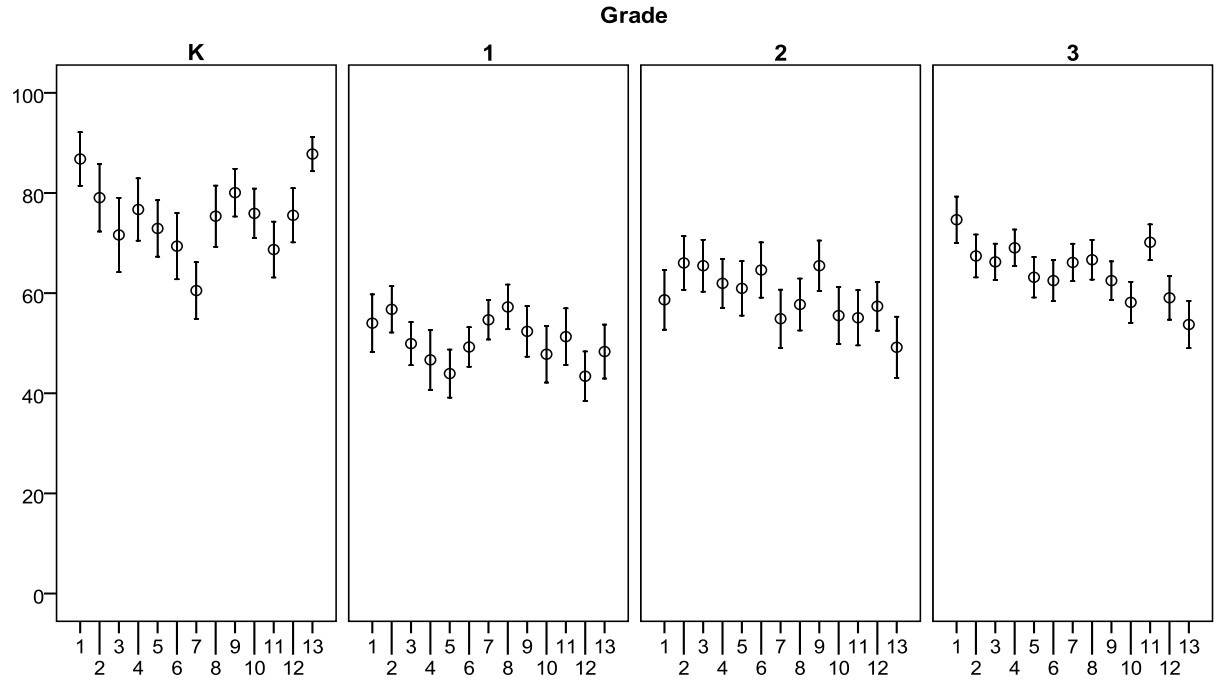


Table 3-8: Vocabulary Corrected Correlations between Form and Average Score by Grade

| Form | Grade K N=18 | Grade 1 N=18 | Grade 2 N=27 | Grade 3 N=30 |
|---|---|---|---|---|
| 1 | .748 | .892 | .693 | .716 |
| 2 | .718 | .584 | .749 | .564 |
| 3 | .665 | .787 | .739 | .661 |
| 4 | .454 | .820 | .806 | .768 |
| 5 | .131 | .637 | .815 | .799 |
| 6 | .728 | .654 | .718 | .779 |
| 7 | .486 | .662 | .785 | .812 |
| 8 | .588 | .430 | .782 | .870 |
| 9 | .438 | .723 | .788 | .676 |
| 10 | .715 | .566 | .757 | .770 |
| 11 | .564 | .905 | .731 | .671 |
| 12 | .864 | .521 | .663 | .691 |
| 13 | .634 | .791 | .889 | .824 |

Figure 5-5: Vocabulary Form Mean 95% Confidence Intervals by Grade



# Between Skill-Level Analyses across Forms (Inferences Regarding Generalizations)

Each of the 13 forms designed to cover parallel content were constructed to include grade-relevant skill areas. Each skill area is designed to assess an important component of reading readiness or reading itself, depending on grade level. As an initial set of analyses, the correlations between skill areas within forms by Grade are summarized in Table 3-9. In addition, an average skill area score was computed for each student to evaluate correlations between average performances among the skills assessed by grade, as shown in Table 3-10. Below, the average inter-skill correlations across forms are reported.

Table 3-9: Between Skill Correlations across Forms, by Grade

| | Grade K | Grade 1 | Grade 2 | Grade 3 |
|---|---|---|---|---|
| Mean Correlation | .206 | .363 | .444 | .402 |
| Median Correlation | .197 | .348 | .489 | .386 |
| SD of Correlations | .145 | .151 | .166 | .151 |
| Minimum Correlation | .000 | .099 | .065 | .027 |
| Maximum Correlation | .491 | .662 | .692 | .631 |

The average inter-skill correlation across forms and grades is approximately .35. This indicates relatively unique skills, with approximately 12% shared variance. The skill areas appear to be measuring unique areas. The maximum correlation between any two skill areas across the forms is .69, indicating less than 50% shared variance.

Based on these findings, it is appropriate to use scores from ISIP Español to report skills independently, as intended. Research regarding Spanish-English bilingual development and assessment will drive further validity studies of ISIP Español domains and their inter-skill correlation to English reading (See Chapter 5). As reviewed by the Center for Early Education and Development at the University of Minnesota, English literacy development has similar relations between skill development in domains of oral language development, phonological awareness, and Spanish reading development (Farver et al., 2007; Gorman & Gillam, 2003; Signorini, 1997). Regarding national goals to improve the English proficiency of all children, a growing body of research provides evidence that Spanish-English Bilingual children's performances on Spanish early literacy measures predict later reading success (Cárdenas-Hagan, Carlson, & Pollard-Durodola, 2007; Cisero & Royer, 1995). Researchers have provided evidence of cross-linguistic transfer of early literacy skills, with higher achievement in Spanish phonological awareness, letter and word knowledge, print concepts, and sentence memory. Cross-linguistic transfer predicts improved reading achievement in English in Kindergarten and Grades 1, 3, and 4 (Lindsey, Manis, & Bailey, 2003; Manis, Lindsey, & Bailey, 2004). Early development of native oral vocabulary may also be related to improving English reading comprehension in elementary grades (Proctor, August, Carlo, & Snow, 2006). These

findings are consistent with the small to moderate correlations found with ISIP Español domains, such that unique information from each domain is relevant in understanding child development of readings skills in Spanish and English and their interrelations.

Average skill scores were computed by taking the average across forms for each student. The average skill scores across forms are much more stable, as they each comprise all of the items administered within a skill area. These average skill scores are correlated and summarized here.

The independent scoring and reporting of ISIP Español is an appropriate way to measure these skills.

Table 3-10: Correlations between Average Skill Scores across Forms, by Grade

| | Grade K | Grade 1 | Grade 2 | Grade 3 |
|---|---|---|---|---|
| Mean Correlation | .456 | .678 | .688 | .638 |
| Median Correlation | .448 | .644 | .682 | .663 |
| SD of Correlations | .164 | .094 | .050 | .051 |
| Minimum Correlation | .275 | .606 | .641 | .580 |
| Maximum Correlation | .744 | .784 | .741 | .672 |

Generally, all of the skill areas resulted in higher inter-skill correlations by using average scores across forms. This is largely expected, due to the higher stability of average scores.

Among Kindergarten students, Phonological Awareness resulted in the lowest correlations among the skill areas. This reinforces the idea that Phonological Awareness is a unique skill, compared to the others.

Association between Skills

- The skill areas appear to be relatively independent and are appropriately reported separately.
- The average inter-skill correlations across forms and grades are highest in Grade 2 (about .44) and Grade 3 (.40), with lower correlations in Kindergarten (about .21). Correlations of average skill scores across forms are higher (.46 to .69).
- The correlations among average skill scores (scores combined across forms) were consistently higher, ranging from .46 (Kindergarten) to .69 (Grade 2). This suggests a moderate association among the skills being measured.

# Item-Level Analysis (Inferences Regarding Scoring, Generalization, Implication)

The data obtained at the item level, although it was collected from a small sample, provides initial information about item functioning. The item responses were analyzed using the Rasch model software Winsteps, which provides (1) classical test statistics of item difficulty (*p*-value), (2) discrimination (point-biserial correlation), and (3) Rasch item statistics (including item fit). Ideally, Infit and Outfit (z-scores) should be within -2 and +2; point-biserials should be positive; p-values should range between .2 and 1.0.

To maximize the information available, all items were analyzed concurrently within skill areas (items across all forms were analyzed simultaneously) by grade. Because of the timed nature of the fluency items, they were not included here.

As can be seen in the following tables (3-11–3-18), few items resulted in poor quality statistics (poor fit or discrimination) across skill area—generally less than 6%. No areas stood out in terms of Rasch item fit. With respect to item discrimination, Reading Comprehension resulted in a high number of items in the poorly functioning range: 43% of Kindergarten items, 21% of Grade 1 items, 21% of Grade 2 items, and 16% of Grade 3 items. In part, the Reading Comprehension result is due to the variable performance of items and forms in Kindergarten. This is also likely a result of having a fewer items in this skill area (only 8 in Kindergarten and 10 in Grade 1).

Regarding the Reading Fluency items, approximately 15–20 of the 50+ items were answered by more than 20 students in Grade 2 and approximately 11–17 of the 50+ items were answered by more than 20 students in Grade 3 across forms. Of these items, 16% in Grade 2 and 8% in Grade 3 had negative discrimination values, indicating potentially poorly fitting items.

Table 3-11: Percent of Items with Infit or Outfit z-Values Larger than 2.0, by Skill Area — Kindergarten

| Form | Skill Area | | | |
|---|---|---|---|---|
| | Listening Comp N=8 | Reading Comp N=8 | Phonological Awareness N=27 | Vocabulary N=12 |
| 1 | 0% | 0% | 0% | 0% |
| 2 | 13% | 0% | 4% | 8% |
| 3 | 13% | 0% | 4% | 8% |
| 4 | 0% | 0% | 0% | 8% |
| 5 | 25% | 0% | 0% | 0% |
| 6 | 0% | 0% | 19% | 8% |
| 7 | 0% | 0% | 4% | 0% |
| 8 | 13% | 0% | 4% | 0% |
| 9 | 0% | 0% | 0% | 17% |
| 10 | 0% | 0% | 11% | 0% |
| 11 | 13% | 0% | 0% | 0% |
| 12 | 0% | 0% | 0% | 0% |
| 13 | 0% | 0% | 0% | 0% |
| **Total** | 6% | 0% | 3% | 4% |

Table 3-12: Percent of Items with Infit or Outfit z-Values Larger than 2.0, by Skill Area — Grade 1

| Form | Skill Area | | |
|---|---|---|---|
| | Reading Comp N=10 | Phonological Awareness N=23 | Vocabulary N=15 |
| 1 | 0% | 9% | 7% |
| 2 | 10% | 9% | 0% |
| 3 | 0% | 0% | 0% |
| 4 | 0% | 4% | 0% |
| 5 | 0% | 4% | 0% |
| 6 | 0% | 9% | 7% |
| 7 | 0% | 4% | 13% |
| 8 | 0% | 0% | 13% |
| 9 | 0% | 17% | 0% |
| 10 | 0% | 0% | 13% |
| 11 | 0% | 4% | 0% |
| 12 | 0% | 13% | 0% |
| 13 | 0% | 0% | 7% |
| **Total** | 1% | 6% | 5% |

Table 3-13: Percent of Items with Infit or Outfit z-Values Larger than 2.0, by Skill Area — Grade 2

| Form | Skill Area | |
|---|---|---|
| | Reading Comp N=15 | Vocabulary N=28 |
| 1 | 0% | 11% |
| 2 | 7% | 7% |
| 3 | 7% | 4% |
| 4 | 0% | 7% |
| 5 | 0% | 4% |
| 6 | 0% | 4% |
| 7 | 0% | 4% |
| 8 | 0% | 11% |
| 9 | 0% | 4% |
| 10 | 0% | 7% |
| 11 | 0% | 7% |
| 12 | 0% | 4% |
| 13 | 0% | 7% |
| **Total** | 1% | 6% |

Table 3-14: Percent of Items with Infit or Outfit z-Values Larger than 2.0, by Skill Area — Grade 3

| Form | Skill Area | |
|---|---|---|
| | Reading Comp N=20 | Vocabulary N=24 |
| 1 | 5% | 4% |
| 2 | 0% | 0% |
| 3 | 0% | 0% |
| 4 | 5% | 0% |
| 5 | 5% | 4% |
| 6 | 0% | 0% |
| 7 | 5% | 0% |
| 8 | 0% | 4% |
| 9 | 0% | 0% |
| 10 | 0% | 4% |
| 11 | 0% | 0% |
| 12 | 0% | 0% |
| 13 | 0% | 0% |
| **Total** | 2% | 1% |

Table 3-15: Proportion of Point-Biserial Correlations (Item Discrimination) Less than 0, by Skill Area — Kindergarten

| Form | Skill Area | | | |
|---|---|---|---|---|
| | Listening Comp N=8 | Reading Comp N=8 | Phonological Awareness N=27 | Vocabulary N=12 |
| 1 | 0% | 25% | 0% | 25% |
| 2 | 0% | 50% | 4% | 0% |
| 3 | 13% | 38% | 0% | 8% |
| 4 | 0% | 75% | 0% | 8% |
| 5 | 0% | 50% | 19% | 33% |
| 6 | 13% | 50% | 15% | 0% |
| 7 | 0% | 25% | 7% | 17% |
| 8 | 0% | 25% | 4% | 0% |
| 9 | 0% | 63% | 4% | 8% |
| 10 | 0% | 25% | 4% | 8% |
| 11 | 13% | 63% | 4% | 8% |
| 12 | 0% | 50% | 7% | 0% |
| 13 | 0% | 25% | 0% | 0% |
| **Total** | 3% | 43% | 5% | 9% |

Table 3-16: Proportion of Point-Biserial Correlations (Item Discrimination) Less than 0, by Skill Area — Grade 1

| Form | Skill Area | | |
|---|---|---|---|
| | Reading Comp N=10 | Phonological Awareness N=23 | Vocabulary N=15 |
| 1 | 30% | 4% | 13% |
| 2 | 40% | 13% | 40% |
| 3 | 0% | 4% | 20% |
| 4 | 30% | 0% | 0% |
| 5 | 0% | 0% | 7% |
| 6 | 20% | 13% | 33% |
| 7 | 20% | 0% | 33% |
| 8 | 10% | 4% | 13% |
| 9 | 30% | 13% | 0% |
| 10 | 10% | 0% | 0% |
| 11 | 30% | 0% | 7% |
| 12 | 20% | 4% | 7% |
| 13 | 30% | 4% | 13% |
| **Total** | 21% | 5% | 14% |

Table 3-17: Proportion of Point-Biserial Correlations (Item Discrimination) Less than 0, by Skill Area — Grade 2

| Form | Skill Area | |
|---|---|---|
| | Reading Comp N=15 | Vocabulary N=28 |
| 1 | 20% | 11% |
| 2 | 13% | 7% |
| 3 | 13% | 4% |
| 4 | 13% | 11% |
| 5 | 20% | 4% |
| 6 | 27% | 0% |
| 7 | 20% | 4% |
| 8 | 20% | 7% |
| 9 | 27% | 7% |
| 10 | 33% | 4% |
| 11 | 20% | 7% |
| 12 | 20% | 7% |
| 13 | 27% | 7% |
| **Total** | 21% | 16% |

Table 3-18: Proportion of Point-Biserial Correlations (Item Discrimination) Less than 0, by Skill Area — Grade 3

| Form | Skill Area | |
|---|---|---|
| | Reading Comp N=20 | Vocabulary N=24 |
| 1 | 10% | 17% |
| 2 | 30% | 25% |
| 3 | 25% | 17% |
| 4 | 20% | 21% |
| 5 | 10% | 17% |
| 6 | 10% | 4% |
| 7 | 15% | 13% |
| 8 | 10% | 8% |
| 9 | 5% | 13% |
| 10 | 15% | 21% |
| 11 | 5% | 8% |
| 12 | 15% | 13% |
| 13 | 40% | 8% |
| **Total** | 16% | 14% |

## Correlations with External Measures (Inferences Regarding Extrapolation)

In addition to completing ISIP Español assessments, students completed beginning of school year evaluation using two external measures: *Evaluación del Desarrollo de la Lectura* (EDL2, Pearson) and *Tejas LEE* (Brookes). Students in Kindergarten through Grade 2 will also complete mid-year and year-end evaluations using the same external measures. Students in Grade 3 will also complete the State of Texas Assessment of Academic Readiness (STAAR Reading, TEA) in the spring of 2015 (these data will also be analyzed at a later date). Correlations with these external measures provide evidence of association with similar measures (traditionally referred to as concurrent criterion-related validity evidence). Correlations with the STAAR provide predictive criterion-related validity evidence, as those scores are obtained several months later. At this time, scores from beginning of school year (BOY) EDL2 and Tejas LEE have been obtained and correlated with each skill area by form and grade.

Decisions and suggestions to be made based on available data include the following:

- These data provide promising results for future administration in a computer adaptive testing system. The larger data set being gathered across multiple districts will be used to estimate item parameters to support computer adaptive testing development.
- These data indicate that a number of items in each domain will be reviewed, potentially revised, or removed from further administration and the computer adaptive testing model.
- Data suggest that students in Kindergarten might be better prepared for administration of Reading Comprehension items during spring semester, rather than beginning Kindergarten.

These data indicate that the small to moderate correlations between domain areas support separate reporting of domain scores. These domains are relatively independent, providing unique information about separate skills.

# 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 Español test, we are interested in comparing students to a national sample of students. 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, 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 all enrolled ISIP Español users during the 2011-2012 school year. In the case of ISIP Español, we felt that demographic considerations were moot, in that most of the students taking the test would be of a similar demographic, and it is difficult to say what constitutes a representative sample of Spanish speaking students in the United States. As such, all users of the program were considered in determining the norms for the test.

Table 4-1. *Demographics* for ISIP Español Reading *Norming Sample*.

| | Grade | | | | | |
|---|---|---|---|---|---|---|
| | PK-3 | PK | K | 1 | 2 | 3 |
| Gender | | | | | | |
| Male | 51.2 | 50.4 | 51.5 | 51.6 | 51.0 | 50.8 |
| Female | 48.8 | 49.6 | 48.5 | 48.4 | 49.0 | 49.2 |
| Race | | | | | | |
| African American | 2.6 | 2.6 | 4.3 | 2.8 | 1.7 | 1.1 |
| American Indian/Native | 11.6 | 8.6 | 12.8 | 12.4 | 9.3 | 13.0 |
| Asian | 0.3 | 0.2 | 0.5 | 0.3 | 0.2 | 0.1 |
| Other | 6.0 | 6.1 | 6.3 | 5.8 | 6.0 | 5.5 |
| Pacific Islander | 0.6 | 0.5 | 0.7 | 0.7 | 0.6 | 0.6 |
| No Answer | 13.7 | 14.3 | 13.7 | 14.1 | 13.2 | 13.7 |
| White | 65.2 | 67.8 | 61.7 | 63.9 | 68.9 | 66.1 |
| Economically Disadvantaged | | | | | | |
| Yes | 57.6 | 70.8 | 59.5 | 54.7 | 53.9 | 57.8 |
| No/NA | 42.4 | 29.2 | 40.5 | 45.3 | 46.1 | 42.2 |
| Ethnicity | | | | | | |
| Hispanic | 76.0 | 79.5 | 74.7 | 75.4 | 75.8 | 77.2 |
| Non-Hispanic | 5.2 | 4.3 | 8.5 | 5.5 | 3.8 | 2.5 |
| No answer | 19.1 | 16.8 | 17.1 | 19.4 | 20.6 | 20.6 |

*Note*: Each category is percent of total responding.

# 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 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.

# Computing Norms

Istation's norms are time-referenced to account for expected growth of students over the course of a semester. The ISIP Español 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. To compute these norms, the ISIP Español test was given to the students in the sample described above, once per month throughout a school year. Because of the test design, including computer-adaptive subtests, retakes of the test 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 Español is expected to go up. Norming tables for each of the ISIP Español subtests, as well as Overall, 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 testtaking. 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.

# References

Abadzi, H., Crouch, L., Echegaray, M., Pasco, C., & Sampe, J. (June 1, 2005). Monitoring basic skills acquisition through rapid learning assessments: A case study from Peru. Prospects: *Quarterly Review of Comparative Education*, 35, 2, 137-156.

AERA, APA, NCME. (1999). *Standards for educational psychological testing*. Washington DC: AERA.

Agrás G, Molina H, & Bareche Monclús, R., (2007). Serie Didáctica de la Lengua y de la Literatura. Editorial Graó.

Armbruster, B. B., Lehr, F., & Osborn, J., National Institute for Literacy (U.S.), & RMC Research Corporation. (2006). A child becomes a reader: Proven ideas from research for parents: kindergarten through grade 3. Washington, D.C.: National Institute for Literacy.

August, D., & Shanahan, T. (Eds.). (2006). Developing literacy in second-language learners: Report of the National Literacy Panel in Language-Minority Children and Youth. Studies in second language acquisition, 30, 1, 116-117.

Ballantyne, K. G., Sanderman, A. R., McLaughlin, N., & National Clearinghouse for English Language Acquisition & Language Instruction Educational Programs. (2008). *Dual language learners in the early years: Getting ready to succeed in school*. Washington, D.C: National Clearinghouse for English Language Acquisition.

Bazán, A. R., Acuña L., & Vega, F. Y., (2002). Efectos de un método para la enseñanza de la lectura y la escritura en el primer grado de primaria. Instituto Tecnológico de Sonora, Departamento de Psicología y Educación, México.

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

Bravo, M. V., Silva, M., Razmilic, T., & Swartz, S. (January 1, 2005). Educando juntos: a long-term intervention to promote literacy learning in low-performing primary schools in Chile. *Early Years, 25,* 2, 97-111.

Broer, M., Castillo, M., Malespín, M., & Gómez, P. (2009). Report on the Results of the EGRA 2008 Pilot Assessment, Oficina de Desarrollo Económico, Agricultura y Comercio (EGAT/ED) MIN ED.

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

Cárdenas-Hagan, E., Carlson, C. D., & Pollard-Durodola, S. D. (2007). The cross-linguistic transfer of early literacy skills: The role of initial L1 and L2 skills and language of instruction. L*anguage, Speech, and Hearing in the Schools*, 249-259.

Carrillo, M. (1994): Development of phonological awareness and reading acquisition. Reading and Writing: *An Interdisciplinary Journal* 6, 279-298.

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

Cisero, C. A. & Royer, J.M. (1995). The development and cross-language transfer of phonological awareness. *Contemporary Educational Psychology*, 20, 275-303.

Crawford, J., & Krashen, S. D. (2007). *English learners in American classrooms: 101 questions, 101 answers*. New York: Scholastic.

Cronbach, L.J. (1988). Five perspectives of the validity argument. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp. 3-18). Hillsdale, NJ: Lawrence Erlbaum.

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

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

Deno, S. (January 1, 2003). Developments in curriculum-based measurement. *The Journal of Special Education*, 37, 3, 184-192.

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

DeThorne, L., Petrill, S., Deater-Deckard, K., Thompson, L., & Schatschneider, C. (January 1, 2006). Reading skills in early readers. *Journal of Learning Disabilities*, 39, 1, 48-55.

Díaz-Rico, L. T., & Díaz-Rico, L. T. (2008). *Strategies for teaching English learners*. Boston: Pearson/Allyn and Bacon.

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

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

Escamilla, K. (November 1, 2006). Semilingualism Applied to the literacy behaviors of Spanish-speaking emerging bilinguals: Bi-illiteracy or emerging bi-literacy. Teachers College Record, 108, 11, 2329-2353.

Escamilla, Kathy, (2000). Bilingual means two: Assessment issues, early literacy and Spanish-speaking children in a research symposium on high standards in reading for students from diverse language groups: Research, practice & policy (pp. 100-128). Washington, DC: U.S. Office of Bilingual Education and Minority Affairs.

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

Espinosa, Linda M., & Lopez, Michael. (2007). Assessment considerations for young English language learners across different levels of accountability Philadelphia: National Early Childhood Accountability Task Force.

Farver, J. M., Nakamoto, J., & Lonigan, C. J. (2007). Assessing preschoolers' emergent literacy skills in English and Spanish with the Get Ready to Read! screening tool. *Annals of Dyslexia*, 57, 161–178. doi: 10.1007/s11881-007-0007-9.

Ferreiro, E. (1997). Alfabetización: Teoría y práctica. México: Siglo Veintiuno Editores, Chile.

Ferreiro, E. (January 1, 2009). Ensayos e investigaciones - La desestabilización de las escrituras silábicas: alternancias y desorden con pertinencia. *Lectura y vida : Revista Latinoamericana de lectura*, 30, 2, 6.

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

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

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

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

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

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

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

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

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

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

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

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

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

Garcia, E. E., & Miller, L. S. (August 01, 2008). Findings and recommendations of the national task force on early childhood education for Hispanics. *Child Development Perspectives, 2,* 2, 53-58.

Gersten, R. M., & National Center for Education Evaluation and Regional Assistance (U.S.). (2007). Effective literacy and English language instruction for English learners in the elementary grades. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

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

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

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

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

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

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

Gorman, B., & Gillam, R. (2003). Phonological awareness in Spanish: A tutorial for speech-language pathologists. Communication *Disorders Quarterly*, 25(1), 13-22.

Hirsch, E. D. J. (June 06, 2003). Reading Comprehension Requires Knowledge--of Words and the World. *American Educator, 27,* 1, 10.

Howard, E. R., Lindholm-Leary, K. J., & Center for Applied Linguistics. (2007). *Guiding principles for dual language education*. Washington, D.C.: Center for Applied Linguistics.

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

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

Jiménez González, J. E., & Ortiz, R. (2000). Metalinguisitc awareness and reading acquisition in the Spanish language. *The Spanish Journal of Psychology*, 3, 37-46.

Jiménez, G., & O'shanahan, I. (2008). *Evolución de la escritura de palabras de ortografía arbitraria en lengua española*. 20.

Junqué, P.C., Bruna, R.O., & Mataró, S.M. (2004). *Neuropsicología del lenguaje: Funcionamiento normal y patológico rehabilitación*. Barcelona: Masson.

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

Kane, M.T. (1992). An argument based approach to validity. *Psychological Bulletin, 112*, 527-535.

Kane, M.T. (2006a). Content-related validity evidence. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development* (pp. 131-154). Mahwah, NJ: Lawrence Erlbaum.

Kane, M.T. (2006b). Validation. In R.L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17-64). Westport, CT: American Council on Education/Praeger.

Kudo, I., Bazan, J., & Banco Mundial (2009). Measuring beginner reading skills: An empirical evaluation of alternative instruments and their potential use for policymaking and accountability in Peru. Washington, D.C: World Bank. Main Report. Tempe, AZ: Author. Retrieved from September 23, 2008.

Lindsey, K. A., Manis, F. R., & Bailey, C. E. (2003). Prediction of first-grade reading in Spanish-speaking English-language learners. *Journal of Educational Psychology, 95,* 482-494. doi: 10.1037/0022-0663.95.3.482.

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

Manis, F. R., Lindsey, K. A., & Bailey, C. E. (2004). Development of reading in grades K-2 in Spanish-speaking English-language learners. *Learning Disabilities Research and Practice, 19,* 214-224.

Marco común europeo de referencia para las lenguas: enseñanza, aprendizaje y evaluación: Propuestas para la enseñanza de ELE. (2005). Madrid: Sociedad General Española de Librería.

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

Mathes, P. G., Fuchs, D., & Roberts, P. H., (1998). The impact of Curriculum-Based Measurement on Transenvironmental Programming. *The Journal of Learning Disabilities*, 31(6), 615-624.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-104). New York: American Council on Education and Macmillan.

National Task Force on Early Childhood Education for Hispanics/La Comisión Nacional para la Educación de la Niñez Hispana. (2007). Para nuestros niños. Expanding and Improving Early Education for Hispanics:

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

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

Pearson, P. D. (2002). Handbook of reading research. Mahwah, N. J: L. Erlbaum Associates.

Proctor, C. P., August, D., Carlo, M.S., & Snow, C. (2006). The intriguing role of Spanish language vocabulary knowledge in predicting English reading comprehension. *Journal of Educational Psychology, 98,* 159-169. doi:10.1037/0022-0663.98.1.159

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

Roseberry-McKibbin, C., Brice, A., & O'Hanlon, L. (January 01, 2005). Serving English language learners in public school settings: A national survey. *Language, Speech, and Hearing Services in Schools*, 36, 1, 48-61.

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

Scarborough, H. S., & Dobrich, W. (January 01, 1990). Development of children with early language delay. *Journal of Speech and Hearing Research, 33,* 1, 70-83.

Serrano, F., Defior, S. & Jiménez, G., (2009). *Evolución de la relación entre conciencia fonológica y lenguaje escrito en niños españoles de primer curso de Educación Primaria*. Universidad de Granada y Ministerio Español de Ciencia y Tecnología.

Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, 94, 143-174.

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

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

Shaywitz, S.E. (1996) Dyslexia, *Scientific American*.

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

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

Signorini, A. (1997). Word reading in Spanish: A comparison between skilled and less skilled beginning readers. *Applied Psycholinguistics*, 18, 319-344.

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

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

Snow, C. E., Carlo, M. S., August, D., Mclaughlin, B., Dressler, C., Lippman, D. N., Lively, T. J., & White, C. E. (January 01, 2004). Closing the gap: Addressing the vocabulary needs of English-language learners in bilingual and mainstream classrooms (Cerrando la brecha: Acerca de las nesecidades de vocabulario de aprendices de inglés en aulas bilingües y comunes). *Reading Research Quarterly, 39,* 2, 188-215.

Sprenger-Charolles, L., Carré, R., & Demonet, J. F. Serniclaes, W. (January 01, 2001). Perceptual discrimination of speech sounds in developmental dyslexia. *Journal of Speech, Language, and Hearing Research*: Jslhr, 44, 2, 384-99.

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

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

Stiggins, R. J., & Assessment Training Institute. (2006). Grading & reporting in standards-based schools. Portland, OR: Assessment Training Institute.

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

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

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

Vaughn, S., & Linan-Thompson, S. (2004). *Research-based methods of reading instruction, grades K-3.* Alexandria, Va: Association for Supervision and Curriculum Development.

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

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

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

Wimmer, H., Mayringer, H., & Landerl, K. (December 01, 2000). The double-deficit hypothesis and difficulties in learning to read a regular orthography. *Journal of Educational Psychology, 92,* 4, 668-80.

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

Ziegler, J. C., Perry, C., Ma-Wyatt, A., Ladner, D., & Schulte-Kör̈ne, G. (January 01, 2003). Developmental dyslexia in different languages: language-specific or universal. *Journal of Experimental Child Psychology, 86,* 3, 169-93.