Spanish Individual Growth & Development Indicators: IGDIs-Español

Technical Manual

Alisha Wackerle-Hollman, Ph.D.
Lillian Durán, Ph.D.
Michael C. Rodriguez, Ph.D.
Stephanie Brunner, M.A.
Terry Kohlmeier, Ph.D.
Chase Callard
José Palma

IGDILab 2018

ACKNOWLEDGEMENTS

The IGDIs-E development team expresses our deepest gratitude to the participating preschool programs, children, parents, and teachers who made this research possible. Without their support and participation the development of the IGDIs-E would not have been possible. We also wish to express our gratitude to the dedicated team of graduate students who participated across four years of IGDIs-E research: Stephanie Brunner, Chase Callard, Theresa Kohlmeier, Jose Palma.

This work was supported by grant R305A120449, Research and Development of Individual Growth and Development Indicators-Español (IGDIs-E): Early literacy identification measures for Spanish-English bilingual children, from the Institute of Education Sciences, U.S. Department of Education. The authors would like to thank colleagues who assisted with this project including participating childcare centers and programs in the Minneapolis/St. Paul area, central Florida, rural Utah, California and Idaho. Additionally, the authors express sincere appreciation for the work contributed by the research team at Utah State University who participated in measure design and data collection. However, the opinions and recommendations presented in this paper are those of the authors alone, and no official endorsement from the Institute of Education Sciences should be inferred.

"*Our language is the reflection of ourselves. A language is an exact reflection of the character and growth of its speakers*." Cesar Chavez

Recommended citation:

Wackerle-Hollman, A., Durán, L., Rodriguez, M.C., Brunner, S., Kohlmeier, T., Callard, C., & Palma, J. (2018). Spanish Individual Growth & Development Indicators: IGDIs-Español technical manual. Minneapolis, MN: IGDILab, University of Minnesota. Retrieved from http://hdl.handle.net/11299/201544

IGDILAB

INTRODUCTION

The Individual Growth and Development Indicators-Español (IGDIs-E) are a universal screening measure designed for use with 4-5-year-old Spanish-speaking preschoolers in their year before kindergarten. We define Spanish-English bilingual (SEB) students as those who have some native degree of Spanish, regardless of when acquired, who are now living in the United States and learning English (also regardless of when acquired). As such, SEB students can be Spanish dominant and learn English only when they enter the formal schooling environment or they can be balanced bilinguals building both Spanish and English skills from birth or early in life.

The IGDIs-E measure oral language, phonological awareness, and alphabet knowledge. In the oral language domain, tasks include *Identificación de los Dibujos/Picture Naming* and *Verbos – Expresivo/Expressive Verbs*. First Sounds measures phonological awareness. Receptive Letter Naming and Sound Identification measure alphabet knowledge.

The development of IGDIs-E is based on a regional sample of Spanish-speaking preschoolers including participants from California, Florida, Kansas, Minnesota, and Utah. Children's parents originated from countries in Central America, South America, Mexico, Puerto Rico, and the Dominican Republic. The IGDIs-E are therefore reflective of the abilities of a wide range of young Spanish speakers in the US and can be used with confidence to determine children's instructional needs.

The IGDIs-E are the first measures available publically that have been developed based on the trajectory of development in Spanish and they are not translations of the English Individual Growth and Development Indicators (IGDIs 2.0). The development team carefully considered approaches to measurement that were based on the best available evidence regarding Spanish language and literacy development. The IGDIs-E are designed to complement the IGDIs 2.0 and both can be administered to SEB preschoolers to obtain scores that can guide meaningful instructional decisions in each language.

Throughout this manual we have aligned our presentation of research, development and related psychometric evidence through the most recent *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), with a focus on establishing a sound validity argument to support the interpretation and use argument for IGDIs-E (Kane, 2013). We present the IGDIs-E as a psychometrically sound, theory-based, practical set of measures designed to assess early literacy skills in Spanish for 4-5 year old SEB preschool children.

As such, we have developed a series of assertions about the interpretations and uses of the IGDIs-E measures. In sum, we have compiled a list of 15 interpretations and uses. These assertions are presented here to provide the reader with a scope of the limits and capacities of the measures. When relevant, each interpretation and use that is supported in each section of the manual is identified in a sidebar.

IGDIs-E Assertions regarding Score Interpretations and Uses

1. The IGDIs-E are psychometrically sound and theory-based, using Mark Wilson's constructing measures model and Rasch modeling for scaling.
2. The IGDIs-E are designed to align with general outcome measure standards including: ease of use, related to meaningful, long-term outcomes, quick and efficient to deliver, meaningful score interpretation and inexpensive or easily accessible.
3. IGDIs-E assess three different yet related domains of early literacy in Spanish- *alphabet knowledge, oral language,* and *phonological awareness.*
4. IGDIs-E are developmentally appropriate for SEB 4-5-year-old children.
5. IGDIs-E are a set of screening measures that can accurately identify student skill level in Spanish within the context of differentiated instruction or within a multi-tiered system of support.
6. IGDIs-E are designed to measure change over a school year.
7. IGDIs-E are inclusive of a variety of Spanish dialects and socio-economic backgrounds that are representative of Spanish speaking populations.
8. IGDIs-E are appropriate for use with children who are native Spanish speakers, with exposure to Spanish ranging from Spanish dominant to English dominant.
9. IGDIs-E are complementary to the English IGDIs 2.0/ Literacy + and together they can assess the overall language development of preschool SEBs.
10. Performance standards at each tier level (Tier 1, Tier 2/3) were set based on empirical data, expert review panels, and information from parents and teachers, and revised based on longitudinal analysis of Kindergarten performance.
11. IGDIs-E can be used within classrooms that use a variety of models of language of instruction (from English only, to some Spanish, to a balanced use of both).
12. IGDIs-E have the potential to be used in a variety of early childhood programs (FFN, Head Start, private and public preschools, etc.).
13. IGDIs-E are uniquely designed to attend to how Spanish language develops rather than by translating existing English measures.
14. IGDIs-E scores can be used to inform instructional planning.
15. Item difficulties are stable across seasons.

Organization of this Manual

1. Overview of the Measures
2. Purpose of the Assessment
3. Origins of the IGDIs-E
4. Test Administration, Scoring, and Interpretation
5. Design Principles and Quality Indicators
6. Measurement Model: Construct Map
7. Measurement Model: Item Development
8. Measurement Model: Outcome Space
9. Measurement Model: Rasch Item Scaling
10. Standard Setting
11. Validity Evidence

Overview of the Measures

Oral Language

Identificación de los dibujos/Picture naming. *Identificación de los Dibujos* is an expressive task that requires children to name images of common and culturally-relevant objects, animals, and foods. The administrator presented cards one at a time to the child. Each card displayed one picture for the child to name. The administrator asked the child, "¿Qué es?" (What is this?). If an image had more than one name due to dialectal differences, all possible correct answers appeared on the back of the card. Items for which the child produced a response that matched a response on the back of the card were scored as correct, and all other responses (i.e., anything that did not appear on the back of the card) were scored as incorrect.

Verbos (expresivo)/Expressive verbs. *Verbos (Expresivo)* is an expressive task that requires children to produce a verb that describes the action being portrayed in a picture. Each card contained one image. The administrator presented each card in succession, asking the child, "¿Qué está pasando?" (What is happening?) Although attempts were made in the design process to select images portraying one clear action, multiple possible responses were included for cards whose images solicited multiple verbs. Items were scored as correct when the child produced a verb that was included on the back of the card, and all other responses were scored as incorrect.

Phonological Awareness

Primeros sonidos/First sounds. *Primeros Sonidos* is a PA task designed to measure a child's ability to identify the initial sounds of words independent of meaning. Administrators presented each item by pointing to and labeling each of two or three images, followed by the prompt, "¿Cuál de estos dibujos empieza con ___?", inserting the target sound in the blank. Item targets varied, including single phonemes (e.g., /c/), and initial syllables (e.g., /cha/). Children responded by pointing to or verbally labeling the image. Scores were recorded as correct (identifying the image matching the target sound) or incorrect (identifying any other image).

Claim 3: *IGDIs-E assess three different yet related domains of early literacy in Spanish-alphabet knowledge, oral language,* and *phonological awareness.*

Alphabet Knowledge

Identificación de las letras (receptivo)/Receptive letter identification. *Receptive letter ID* measures a child's ability to select (point to) the letter that corresponds to the letter spoken by the administrator. Prompts include "¿Cuál letra es __?" or "Señala la letra __." with the name of the target letter in the blank. Letters were represented as lowercase or uppercase in a uniform font. Scores were recorded as correct (pointing to the target letter) or incorrect (pointing to any other letter).

Identificación de los sonidos/Sound identification. *Sound ID* was designed to measure a child's ability to match the sound of a target letter with the written form of that letter. Children were expected to point to the target letter after hearing a letter sound from the administrator. The prompt was "Cuál letra hace el sonido ___?" with the blank filled in with the letter sound. Letters were again represented as either lowercase or uppercase on the cards in a uniform font. Scores were recorded as correct (pointing to the target letter) or incorrect (pointing to any other letter).

PURPOSE OF THIS ASSESSMENT
What is the Purpose?

The IGDIs-E were developed to be used as a universal screener within a multi-tiered system of support (MTSS) or Response to Intervention (RTI) model of service delivery. In addition, if a program is not implementing MTSS, the IGDIs-E can be used within a model designed to differentiate instruction and intervention without tiered service delivery. The IGDIs-E should be administered three times a year in general education preschool settings with the goal of identifying children in need of higher levels of support. In the context of an MTSS, the IGDIs-E designate children in need of Tier 2/ 3 instruction.

Language and early literacy development for SEB children. Ample evidence for preschool-age English-speaking monolingual children demonstrates that component domains of early literacy, including oral language, phonological awareness, and alphabet knowledge, are strong predictors of how well children will learn to read (Farver et al., 2007; Anthony & Lonigan, 2004; Lonigan, 2009; Lonigan, Burgess & Anthony, 2000; NELP, 2008; Whitehurst & Lonigan, 1998). Existing English language measures of early literacy describe children's performance in these areas, including *Get Ready to Read!* (GRTR; Whitehurst & Lonigan, 2001), *IGDIs 2.0* (McConnell et al., 2010), and the *Test of Preschool Early Literacy* (TOPEL; Lonigan et al., 2007). Researchers and practitioners now need assessments that perform similarly for SEB children. Complementary evidence is emerging that SEB skill development in Spanish predicts reading achievement in English and thus may be a similarly important foundation (Cárdenas-Hagan, Carlson, & Pollard-Durodola, 2007; Cisero & Royer, 1995).

Contemporary scholars (MacWhinney, 2008) suggest competence in one language may affect performance in the other. MacWhinney's Unified Model of Language Acquisition states whatever *can* transfer from a child's first language (L1) will transfer to their second language (L2), where L1 represents a resource to support L2 acquisition. This model builds on Cummins' Developmental Interdependence Hypothesis (DIH; described on p. 9) in which L2 language acquisition is dependent on the level of skill represented in L1 at the time the new language was introduced. This cross-linguistic transfer will prove important to descriptive and treatment research, but first we must be certain we can measure development and competence in each language. Therefore, it is imperative that we assess SEB children in both languages to better understand each child's abilities in Spanish and English based on the quality and quantity of exposure they have had in each.

Claim 1: *The IGDIs-E are theory-based.*

Researchers indicate there is overlap in foundational early literacy skills (e.g., phonological awareness, oral language and alphabet knowledge) that are predictive of reading in both English and Spanish. As with English speaking young children, associations have been found between early Spanish oral language, phonological awareness, and alphabet knowledge and later Spanish reading development (Borzone de Manrique & Signorini, 1994; Branum-Martin et al., 2006; Carillo, 1994; Cisero & Royer, 1995; Farver et al., 2007; Gorman & Gillam, 2003; Jiménez González, & García, 1995; Signorini, 1997). Researchers also provide evidence of cross-linguistic transfer of many of these same early literacy skills with higher achievement in Spanish phonological awareness, letter and word knowledge, and print concepts in kindergarten and first grade predicting improved reading achievement in English in third and fourth grades (Lindsey, Manis, & Bailey, 2003; Manis, Lindsey, & Bailey, 2004). From correlational studies, we find SEB children with higher skills in Spanish phonological processing, oral language development, decoding, and print awareness have improved reading skills in English (August, Carlo, Dressler, & Snow, 2005; Cobo-Lewis, Eilers, Pearson, & Umbel, 2002; Dickinson, McCabe, Clark-Chiarelli, & Wolf, 2004; Durgunoğlu, Nagy, & Hancin-Bhatt, 1993; López & Greenfield, 2004; Proctor, August, Carlo, & Snow, 2006). Given the associations identified between early literacy skills in Spanish and later reading achievement in English, the IGDIs-E will provide a basis for identifying intervention candidacy and for evaluating those interventions targeted at building proficiency in Spanish as a foundation for later English (and Spanish) language and literacy competence. Therefore, we first focus our effort on assessing skills that are strong

Claim 13: *IGDIs-E are uniquely designed to attend to how Spanish language develops rather than by translating existing English measures.*

correlates with reading in English and second focus efforts on targeting skills that show promise for predicting both English and Spanish reading outcomes for those children who attend bilingual elementary school programs where they receive reading instruction in both languages.

Phonological awareness. Phonological awareness (PA) is the ability to detect and manipulate sound structure of words independent of their meaning; as such, PA serves as the foundation for later decoding skills (Phillips, Clancy-Menchetti, & Lonigan, 2008; Torgeson & Mathes, 2000), and is one of the strongest determinates of learning to read (Foy & Mann, 2006). Phonological awareness has been tested via a variety of tasks, all of which involve identification and manipulation of the structural units of language and vary in level of difficulty. These tasks include *segmenting* (dividing words into syllables, phonemes, and onset-rime units), *blending* (identifying a word when its phonemes or syllables are spoken with breaks in between), *matching* (identifying a word with the same onset sound as the target word; referred to as sensitivity to alliteration by Carrillo (1994); Durgunoglu et al., 1993); *elision* (deleting syllables and phonemes in order to form new words; Cardenas-Hagan et al., 2007); *rhyming* (identifying a word with the same rime, or ending sound, as the target word; Anthony et al., 2011; Carrillo, 1994); *identification of discrepant sounds* (Kuo & Anderson, 2010); *position segment identification* (stating whether the target phoneme is the initial, medial, or final phoneme), and *segment isolation* (identification of initial or final phoneme or syllable following hearing the target word; related to segmenting; Carrillo, 1994; Cisero & Royer, 1995). Due to strong correlations between PA tasks, phonological awareness can be thought of as a unified underlying construct that manifests in a sequence of skills that are increasingly complex (Anthony et al., 2011; Branum-Martin et al., 2006).

PA is a meta-linguistic skill that appears to transfer across languages because it accesses cognitive skills that support language learning in general (Kuo & Anderson, 2010; Castilla, Restrepo, & Perez-Leroux, 2009). This characteristic of PA is supported by Cummins' (1979) Developmental Interdependence Hypothesis, which is the idea that skills from one language support the development of a second language. Similarities between PA development in two languages suggest that PA abilities can and will transfer between these languages (Cisero & Royer, 1995; Melby-Lervåg & Lervåg, 2011). As previous researchers have highlighted, PA development in Spanish and English is highly similar, and there is significant evidence of a cross-linguistic

Claim 1: *The IGDIs-E are theory-based.*

Claim 13: IGDIs-E are uniquely designed to attend to how Spanish language develops rather than by translating existing English measures.

transfer of PA skills in SEB children. The relevant question when developing the IGDIs-E, then, was how well Spanish PA skills transfer to assist in English literacy development (Gutierrez, Zepeda, & Castro, 2010).

Regarding item development for PA test types, researchers have suggested a series of features potentially best suited for IGDIs-E. Anthony and colleagues (2011) proposed that it may be best to have all items involve manipulation of syllables, but to test syllable manipulation via a variety of tasks (i.e. blending, segmenting, and sound identification of syllables). In other words, it seems effective to use one level of linguistic complexity evaluated by tasks that vary in difficulty. The type of test, rather than the phonological unit, is what will poise some items to have the potential to require higher levels of ability. The use of multiple tasks should be more helpful to researchers and educators in the identification of children with delayed PA abilities.

In sum, the PA items considered for the IGDIs-E involved blending, sound identification, and segmenting of syllables using culturally appropriate and common two-to-three-syllable vocabulary words (Gorman & Gillam, 2003; Anthony et al., 2011; Carrillo, 1994; Branum-Martin et al., 2006). In particular, initial and final sound identification are developmentally appropriate for preschool children and are strong predictors of reading ability in the early elementary grades (Gorman & Gillam, 2003), so it seems necessary to include this type of task when constructing PA measures. Gorman and Gillam (2003) also suggested a few intriguing testing ideas, such as asking children to segment a word into the smallest pieces they can rather than explicitly asking for segmentation of syllables as an attempt to gauge SEB children's skill level to see if it matches the developmental progression outlined in the literature.

Oral language. Oral language (OL) is a child's use of expressive and understanding of receptive vocabulary to share ideas (Priest et al., 2001). Vocabulary development and the acquisition of meaning of words is a key area of oral language, and a strong contributor to reading comprehension in early as well as upper elementary grades (Dickinson, & Tabors, 2001; Snow, Burns & Griffin, 1998; Biemiller, 2005, Beck, McKeown & Kucan, 2002).

The association between Spanish and English OL development is less clear than that of PA; whereas some researchers have found small associations between OL in Spanish and later reading comprehension and word reading efficiency in English (Miller et al., 2006; Proctor et al., 2006), these associations are modest and not as robust as those found for PA (Mancilla-Martinez & Lesaux, 2010). There is however, evidence that measures of oral language administered in English and Spanish to SEB preschoolers are more predictive of reading in English in kindergarten than measures administered in only one language (Hammer, Lawrence & Miccio, 2007). This provides evidence that measurement of Spanish OL in addition to English could be important to increase criterion-related predictive validity evidence.

There are two main categories of Spanish OL measurement in the literature: decontextualized standardized measurement (e.g. picture naming) and contextualized naturalistic measurement (i.e. naturalistic speaking and listening tasks; Miller et al., 2006). Some argue the context of standardized testing may limit SEB children's performance because these contexts and ways of eliciting language may be unfamiliar (Miller et al., 2006; Peña et al., 2003; Peña & Halle, 2011). Thus, measures that integrate decontextualized and contextualized items may be best.

Tasks beyond picture naming or recognition may also need to be developed that access different types of semantic skills. Peña and colleagues (2003) found that item types, including categorization, characteristic properties, similarities and differences had easy to medium levels of difficulty for the preschool and early school age range. These findings indicate that a range of OL tasks may provide opportunities for more accurate identification of SEB children who need intervention and as such, IGDIs-E include three measures of oral language: *Identificación de los Dibujos/Picture Naming*, *Verbos – Expresivo/Expressive Verbs* and the Storybook narrative task.

Alphabet knowledge. *Alphabet knowledge (AK) is the recognition and production of letter* names and sounds (McBride-Chang, 1999). Evidence indicates that along with PA, AK is one of the strongest longitudinal predictors of future reading success in young children (Adams, 1990; Ball & Blachman, 1991; Bradley & Bryant, 1983; NELP, 2008). Studies have also found associations between AK in Spanish and English and AK in Spanish and later reading in English (e g. Dickinson et al., 2004).

Because there is significant overlap in AK skills for both languages, AK IGDIs-E measures were designed stemming from initial English formats. The summary of research presented here contributes evidence for specific design features within IGDIs-E that attend to Spanish phonological awareness, oral language and alphabet knowledge development.

The Origins of Spanish-IGDIs: Why is it important to measure Spanish language development?

Theoretical and Empirical Rationale for the IGDIs-E

Understanding and preventing academic failure in SEB[1] children is perhaps one of the most persistent, challenging, and controversial issues facing U.S. schools. This issue is pronounced in the elementary and secondary levels (Fry & Gonzales, 2008; Grigg, Daane, Jin, & Campbell, 2003; National Center for Education Statistics, 2003), but also occurs in discussions about services and supports for preschool-aged children (Garcia & Miller, 2008; Garcia & Jensen, 2009). Latinos specifically are overrepresented in the category of Learning Disabilities (LD) with fifty-six percent of all Latino students in special education programs identified as LD with reading problems as their primary concern (Zehler, et al., 2003).

Early intervention is a critical component of reducing disproportionate representation of diverse students within special education (Donovan & Cross, 2002). However, early childhood programs that seek to promote children's later literacy performance have few tools available that help them determine whether an SEB child is on a path toward success. A critical aspect of addressing this challenge is *developing improved measures of language and literacy development of SEBs that meet rigorous psychometric standards for scientific and clinical use*. Such measures require special attention to the developmental course of language and literacy development among SEB young children through carefully designed, rigorously implemented research. Simultaneously, measures need to be sensitive to the practical demands of conducting assessment in community early education settings. Further, simple translation of existing English measures is not sufficient, because it does not take into account differences in the content of each language, differences in language structure, rates of development and the manner and context in which other languages are acquired. IGDIs-E were designed to respond to this need.

Claim 1: *The IGDIs-E are theory-based.*

Claim 9: *IGDIs-E are complementary to the English IGDIs 2.0/ Literacy + and together they can assess the overall language development of preschool SEBs.*

Claim 13: *IGDIs-E are uniquely designed to attend to how Spanish language develops rather than by translating existing English measures.*

Claim 14: *IGDIs-E scores can be used to inform instructional planning.*

[1] A variety of labels and terms are available in the literature for children of interest here, including English-Language Learners, Dual-Language Learners, and Spanish-English Bilinguals. In many ways, these terms are (for the purposes of this project) functionally equivalent; we are interested in children who have some proficiency in Spanish language and literacy as well as instructional goals in English language and literacy development. For ease of presentation, we will refer to these children as Spanish-English Bilingual or SEB.

IGDIs-E are designed to be used descriptively and within MTSS to identify students who are candidates for Tier 2 (targeted) and Tier 3 (intensive) intervention. *IGDIs-E accurately and efficiently measure important developmental domains of early literacy, including alphabet knowledge, phonological awareness and oral language in Spanish to provide knowledge of early skills and abilities in Spanish* to complement information about English performance (IGDIs 2.0).

Research and development of these measures was built on the innovative work at the Center for Response to Intervention in Early Childhood (CRTIEC) where English IGDIs (IGDIs 2.0) were developed and refined based on Wilson's (2005) measurement framework, Rasch item response modeling and features of general outcome measurement (GOM; Rodriguez, 2011).

It is important to note that IGDIs-E are not merely literal translations of English measures. Rather, IGDIs-E reflect pertinent features of Spanish early literacy skill development and the acquisition of bilingualism in 4- to 5-year-old SEB children. Development of IGDIs-E included attention to linguistic, metric, cultural and functional factors that influence performance and alignment with early literacy constructs present in English early literacy development (as measured with English IGDIs 2; Peña, 2007). The process of development utilized expertise within early childhood community and school-based programs, research experts and content resources to produce a psychometrically rigorous model that has broad utility within a variety of settings, and allows educators to accurately assess SEB children's early literacy skill level *and* identify children who may need additional intervention prior to formal reading instruction. This provides the foundation of a successful MTSS model: identification.

IGDIs-E research and validation has centered on **three research aims**:

(a) develop, evaluate, and prepare measures that reflect contemporary criteria for assessment of Spanish language and early literacy to complement existing English-language measures,

(b) develop a set of items that can be used to accurately identify students who are appropriate candidates for tiered intervention, and

(c) establish conceptual and psychometric qualities, comparative scoring, and growth evidence for IGDIs-E and IGDIs 2.0 assessments.

Background & Context: Reading Readiness of Spanish Speakers in the US

Over the past decade the percentage of Latino children who attend US schools has increased dramatically, approaching 20% in the year 2000 (Garcia & Jensen, 2009). Approximately three in four young Latino children live in homes where at least some Spanish is spoken regularly (Garcia & Jensen, 2009). As many as 85% of Latino children are not proficient readers by 4th grade (NAEP, 2008; Grigg et al., 2003; National Center for Education Statistics, 2003), indicating an alarming disproportion of SEB students with limited literacy. Similarly, SEB preschoolers demonstrate low performance on most measures of language and literacy development (Garcia & Jensen, 2009; Páez, Tabors, & López, 2007).

Claim 1:
The IGDIs-E are theory-based.

Claim 13:
IGDIs-E are uniquely designed to attend to how Spanish language develops rather than by translating existing English measures.

Successfully learning to read is one of the most prominent indicators of academic success (National Reading Panel, 2000). A large and diverse body of research indicates children who are not successful readers early on are less likely to be academically and socially successful (Cunnigham & Stanovich, 1997; Dickinson & Neuman, 2006; Juel, 2006; Lyon, 1996; Reynolds & Temple, 1998; Snow, Burns & Griffin, 1998). Although a variety of child and environmental characteristics may be associated with delays in early reading proficiency (Snow, Burns, & Griffin, 1998), beginning one's educational career with limited English proficiency is a robust predictor of later reading delays and difficulties (Vaughn et al., 2006; Garcia & Jensen, 2009).

To date, however, it has been difficult to parse out the various parameters of dual language acquisition, given the significant variability in the input these children receive in English and Spanish across home and school language environments (Hammer, Miccio, & Rodriguez, 2004). This variability makes it difficult to determine what typical development in language and literacy should be for this population in each language. Additionally, once SEB children reach preschool age they are often enrolled in predominantly English-speaking preschool environments which can cause dramatic and rapid shifts in language performance (Anderson, 2004). Being bilingual in and of itself should not be a risk factor for language delay or lower reading achievement (Paradis, 2010; Gillam, Peña, Bedore, Bohman & Mendez-Perez, 2013). However, the subtractive bilingual contexts of school programs in the US in addition to other risk factors such as higher rates of poverty have created significant risk for poor reading and academic

achievement (Goldenberg, 2008; Rolstad, Mahoney, & Glass, 2005; Slavin & Cheung, 2005).

In sum, researchers provide limited knowledge regarding how trajectories of early literacy and language development in SEB children predict later literacy achievement. Early childhood programs that seek to promote children's later literacy performance have few tools to help them determine whether an SEB child is on a path toward success. As such, current practice in early assessment and intervention is not sufficient.

Claim 1:
The IGDIs-E are theory-based.

Thus, it is imperative that assessments are provided that accurately and validly measure SEB children's early literacy skills in English and Spanish. Such information will allow for appropriate instructional decisions, including determining candidacy for tiered early intervention.

Recent theoretical and empirical work in both measurement and Spanish-English bilingualism provides clear criteria to guide development of measures. Language and early literacy measure development for young SEB preschoolers must rest on a careful analysis of children's skills in each language, and consideration of the relation in status and development between languages (Cummins, 1979; Hammer, Lawrence, & Miccio, 2007; MacWhinney, 2008). For bilinguals, measures in English and Spanish must be purposefully designed for use together. Such measures should reflect common approaches for considering linguistic and developmental factors and complementary approaches to scoring and scaling performance. Attention to cultural variations and relevance of items and measures, are also critical for development of quality measures (Peña, 2007). Finally, measure development and evaluation should rest on a solid understanding and analysis of the essential features of assessment (Wilson, 2005; Rodriguez, 2011).

Claim 13:
IGDIs-E are uniquely designed to attend to how Spanish language develops rather than by translating existing English measures.

There are few adequate measures that address the skills of the SEB preschooler, and no measures, beyond IGDIs-E, that have been systematically and intentionally developed and evaluated for use with SEB students in keeping with these criteria. To better understand factors affecting SEB students' reading development, and to provide teachers and policy makers with tools to help prevent reading problems in this group, new measures, such as IGDIs-E, are critically needed.

Understanding Contributing Variables: Language Exposure and Language of Instruction

Language exposure. In the early childhood years, language proficiency is significantly impacted by the current level of language exposure children have to the various languages used across their natural environments. Current language exposure includes what languages the child hears and speaks across all natural settings and all communicative partners. Specifically, SEBs current exposure to English and Spanish has been found to effect children's ability levels in each language (Bedore et al., 2012; Bedore, Peña, Griffin & Hixon, 2016; Goodrich, Lonigan & Farver, 2013; Quiroz, Snow & Zhao, 2010). Although the language proficiency of young SEBs is important in understanding how to improve their early language and literacy outcomes, few studies have examined how to empirically evaluate language exposure and then use this information to examine performance trends on assessments in ways that can inform instruction. Therefore it is important for researchers who are interested in both assessment and intervention issues with SEBs to carefully consider the quantity and quality of language exposure children have to more fully understand their ability levels in each language to more accurately interpret performance on assessments in each language and to more effectively target instruction.

A second approach for understanding language development in SEBs uses the current level of language exposure to L1 and L2 as a variable in predicting rates of child level performance (Hoff, 2010; Oller & Eilers, 2002; Scheele, Leseman & Mayo, 2010). Current level of language exposure models require that caregivers of SEBs document to what degree the child is exposed to L1 and L2, where exposure is defined as some combination of the quantity of language spoken to the child (i.e. what the child hears, or input) as well as the quantity of language the child speaks (i.e. what the child says, or output) in L1 and L2 over the course of a defined period of time (e.g., a week). This concept is supported by the Usage-Based Theory of language acquisition and Emergentism (O'Grady, Lee & Kwak, 2009; Tomasello, 2005), which in its simplest form posits that children learn language by using language and therefore it logically follows that any fluctuation in overall exposure to and use of any language will impact a child's language development in that language (Lieven & Tomasello, 2008).

Claim 1:
The IGDIs-E are theory-based.

Claim 13:
IGDIs-E are uniquely designed to attend to how Spanish language develops rather than by translating existing English measures.

Level of exposure measurement. Various models exist for assessing such levels of exposure. For example, the Bilingual Input/Output Survey (BIOS; Peña, Gutérrez-Clellen, Iglesias, Goldstein & Bedore, 2014) designed as part of the Bilingual English Spanish Assessment (BESA) provides a detailed analysis of hourly breakdown of exposure levels. The Preschool Language Scale-5 Home Communication Questionnaire (Zimmerman, Steiner & Pond, 2014) provides a broad indicator of the languages used in various environments (e.g. community, home, school etc.); and the Language Diary method developed by De Houwer and Bornstein (2003), asks the caregiver(s) to keep a record of the children's language exposure every day of the week for a period of seven weeks, with each day broken into 30 minute blocks. The diary provides a highly detailed description of children's bilingual experiences and offers information beyond what can be gathered by retrospective caregiver report. Another recently developed measure the Language Exposure Assessment Tool (LEAT; DeAnda, Bosch, Poulin-Dubois, Zesiger, & Friend, 2016) uses a digital format to examine input and output and has been found to demonstrate high internal consistency and with criterion-related validity evidence. Finally, the Language Exposure Evaluation Report (LEER; Durán & Wackerle-Hollman, 2015) is a survey that features a summarized breakdown of languages spoken to the child and languages the child speaks by days of the week.

The questions asked on language exposure questionnaires vary, but all generally target similar information. For example, the BIOS evaluates language exposure by requiring parents to chronicle what language is heard and spoken hour by hour, for two example days of the week –one weekday and one weekend day, through completion of a form or in an interview. The survey includes a series of hourly questions that chronicle who is participating in the conversation, when the child is awake, what language the child is speaking in and what language(s) the child is hearing. This sequence is repeated until all time is accounted for, from the time the child awakes to the time the child goes to bed in hourly increments (Peña et al., 2014). This approach yields a rich source of incremental data on the ebb and flow of language exposure with the added benefit of offering information on who is participating in the communicative interaction with the child.

Claim 1:
The IGDIs-E are theory-based.

Claim 13:
IGDIs-E were uniquely designed to attend to how Spanish language develops rather than by translating existing English measures.

The LEER uses another approach by providing users with a time-block matrix where parents check off what languages are heard and spoken in each block for week day and weekend segments (see Appendix A). This approach allows the parents to report what the child speaks and hears throughout the week. However, this approach also faces the challenge of not providing any information about who is speaking to the child and limits responses to 4-hour time blocks rather than 1-hour blocks. Lastly, the LEAT's calculation of relative language exposure requires caregivers to provide a daily account of who uses which language(s) with their child across for each hour of the day in a one week period. Unlike other parent reports this survey also includes questions about the dialect of the language spoken and whether or not the person is a native speaker. The LEAT has been found to be related the children's' vocabulary size in each of their languages (DeAnda et al., 2016). In all three approaches, the time periods where a child is in childcare or a preschool environment may be more accurately depicted if the caregivers solicit information from the childcare or preschool provider, or if the childcare or preschool provider completes the survey for the relevant time blocks.

The influence of level of exposure on early language and performance. Conceptually, current level of exposure may be a more robust approach than age of acquisition because it examines the current state of the child's language by assessing what they speak (output) and hear (input) rather than relying on the age at which L2 was first introduced. Logically, the greater the length of time between age of acquisition and the outcomes of interests, the greater the opportunity for covariates to contribute to or influence how L2 has developed. Language exposure includes both home and school environments. Oller and Eilers (2002) support the notion that level of exposure in school, specifically opportunities for input and output as a function of bilingual education, impact student language and literacy performance in L1 and L2. In their study they examined the role of type of educational setting: immersion classrooms in contrast to transitional bilingual education. Results indicated SEBs who were in the immersion setting, where exposure in English was maximized, had higher English scores than students in transitional classrooms.

Level of exposure in home environments has been found to be useful in categorizing children into different language exposure categories for analysis (e.g. Bedore et al., 2012; Bedore, Peña, Griffin, & Hixon, 2016; Ruiz-Felter, Cooperson, Bedore & Peña, 2016). Specifically, Bedore and colleagues (2012) provide a conceptual model for categorization by dividing average input and output level of exposure in each language. They developed five categories: Functional Monolingual English (input and output >80% English), Bilingual English Dominant (input and output >60%, <80% English), Balanced Bilingual (input and output >40%, <80% English and Spanish), Bilingual Spanish Dominant (>60%, <80% Spanish) and Functional Monolingual Spanish (input and output >80% Spanish).

This model provides a robust foundation for exploring language exposure; however, it is not without limitations. For example, this model focuses categorization on the mean values of input and output, which reduces the impact of individual differences.

Identification of language exposure clusters. Current work on language clusters represent important progress in considering performance trends of students in such membership groups, however, these approaches have not included an approach to predict or evaluate the meaningfulness, appropriateness, or usefulness of group membership through cluster analysis, factor analysis, or classification consistency. Without strengthening group membership categories through empirical clusters or factors, inferences and claims related to how we interpret data from each category are weakened. As such, it is important to examine how language exposure, through input and output organizes in clusters or factors. A cluster analysis allows individual data with similar patterns (based on relevant characteristics or variables) to be joined together consecutively until all data are accounted for. Each step within the cluster analysis is sequential, so clusters formed earlier in the process cannot be split later in the process, allowing for homogenous cases to cluster together tightly (Kaufmann & Rousseeuw, 2009).

We completed confirmatory analyses to demonstrate the degree to which group membership could be identified in the data. In order to understand how the questions on the LEER could be used to classify a child's dominant language, we employed two statistical techniques: multiple correspondence analysis (MCA) (Abdi & Valentin, 2007; Husson & Josse, 2014) and agglomerative hierarchical cluster analysis (AHCA) (Blashfield, 1976; Hansen & Jaumard, 1997). Here we provide a brief overview of these methods in the context of our results. All analyses were performed in R (v.3.2.4, R Core Team, 2015) using the R packages dplyr (Wickham & Francois, 2015) and FactoMineR (Husson, Josse, Le, & Mazet, 2015).

MCA is a method for reducing multidimensional categorical data, such as the categorical answers to the questions in the present questionnaire. In this method, multidimensional data can be reduced to projections onto a low-dimensional space by identifying associations between categorical variables (Abdi & Valentin, 2007; Husson & Josse, 2014).

Claim 1:
The IGDIs-E are theory-based.

Claim 8:
IGDIs-E are appropriate for use with children who are native Spanish speakers, with exposure to Spanish ranging from Spanish dominant to English dominant.

Claim 13:
IGDIs-E are uniquely designed to attend to how Spanish language develops rather than by translating existing English measures.

For example, each response set on the LEER provided categorical responses to 23 questions about language use and exposure. By employing MCA, the associations between responses can be estimated to localize similar patterns of responses in a low-dimensional space. Each dimension explains a portion of the variance in the overall set of responses (Abdi & Valentin, 2007; Husson & Josse, 2014). It is the categorical version of principal component analysis, helping to identify a smaller structure in complex data.

We performed an MCA on the 587 respondents who responded to all questions on the family survey. The first two dimensions explained substantially more unique variance than subsequent dimensions and are thus considered (see Figure 1). As would be expected, languages are clustered together. In the first dimension, responses of *Spanish* are separated from responses of *Both* and *English*. In the second dimension, responses of *English* are separated primarily from responses of *Both*, where *Spanish* and *Both* are relatively closer (see Figure 2).

Figure 1. Scree plot of explained variance following the multiple correspondence analysis.

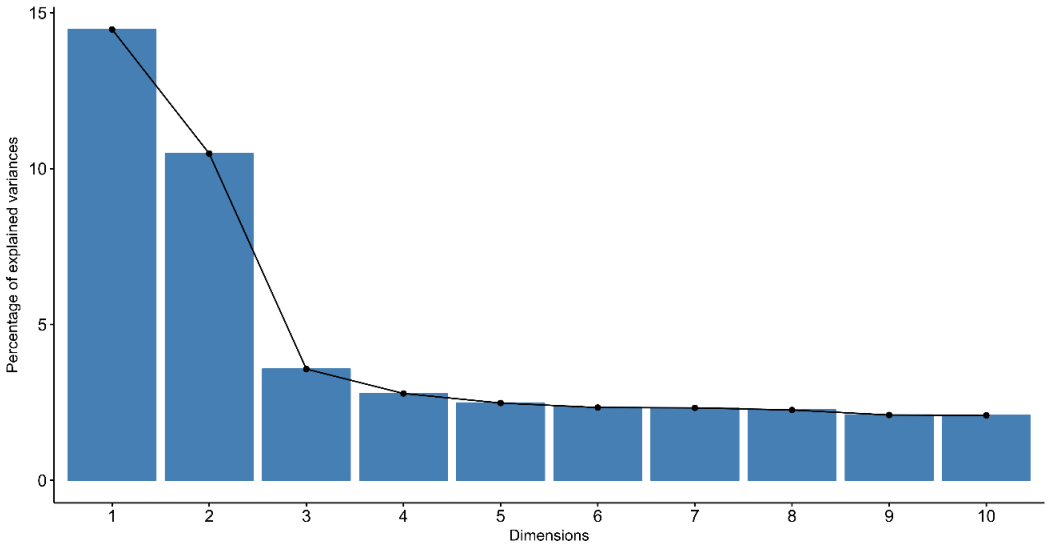
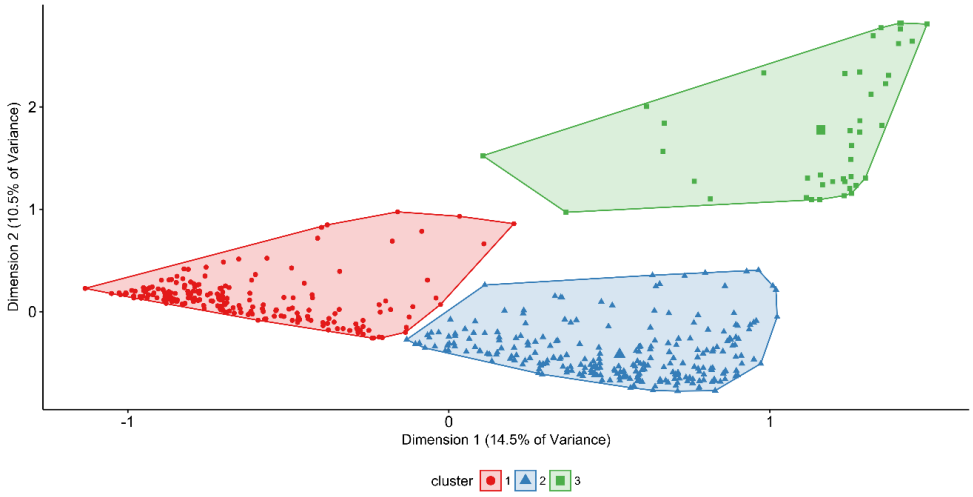


Figure 2. Language clusters.



Note: Cluster 1 is the Spanish-speaking cluster. Cluster 2 is the Bilingual cluster. Cluster 3 is the English-speaking cluster.

Language cluster identification. Although the MCA provides a means for identifying similar responses, it does not inform which individuals are similar to one another. Here we applied agglomerative hierarchical cluster analysis to the MCA-transformed data to localize individuals on this two-dimensional space (Husson et al., 2015). Three clusters emerged in the data (Figure 2). Tables 1 to 4 contain the results of these analyses. In tables 2 to 4, the "Sample" column indicates the percent of children who gave the particular response in the whole sample who belong to the cluster. The "Cluster" column reflects the percent of children in the cluster who provided the identified response. For example, 95% of the children who speak Spanish from 9A - 1P on the weekends (In Sample) are in the Spanish cluster. Of the children in the Spanish cluster (In Cluster), 96% speak Spanish on the weekends from 9A - 1P.

The MCA results indicate three clusters emerge from our data set representing Spanish, English and Both. Results suggest cluster membership is driven primarily by the language spoken and heard by the child on the weekend (Saturday and Sunday) with reduced influence from the language spoken and heard during the week. Although there is some overlap between clusters 1 and 2, the groupings appear generally distinct. Furthermore, we can identify the variables most strongly associated with each cluster.

Table 1. *IGDIs-E Language Exposure Dimensions*

	Eigenvalue	*Variance (%)*	*Cum. Variance (%)*
dim 1	0.54	14.5	14.5
dim 2	0.39	10.5	25.0
dim 3	0.13	3.6	28.5
dim 4	0.10	2.8	31.3
dim 5	0.09	2.5	33.8
dim 6	0.09	2.3	36.1
dim 7	0.09	2.3	38.5
dim 8	0.08	2.3	40.7
dim 9	0.08	2.1	42.8
dim 10	0.08	2.1	44.9
dim 11	0.08	2.1	47.0
dim 12	0.08	2.0	49.0
dim 13	0.07	2.0	51.0
dim 14	0.07	1.9	52.9
dim 15	0.07	1.8	54.7
dim 16	0.07	1.8	56.5
dim 17	0.07	1.7	58.2
dim 18	0.06	1.7	59.9
dim 19	0.06	1.6	61.5
dim 20	0.06	1.5	62.9
dim 21	0.05	1.5	64.4
dim 22	0.05	1.4	65.8
dim 23	0.05	1.4	67.2
dim 24	0.05	1.3	68.5
dim 25	0.05	1.3	69.8

Table 2. *Spanish Speaking Cluster Reporting*

	Sample %	*Cluster %*
SSSpeak_91=SSSpeak_91_Spanish	94.6	96.0
SSSpeak_14=SSSpeak_14_Spanish	95.8	91.2
SSSpeak_4B=SSSpeak_4B_Spanish	94.1	92.7
SSSpeak_A9=SSSpeak_A9_Spanish	87.3	96.0
MFSpeak_4B=MFSpeak_4B_Spanish	92.8	90.1
SSHear_4B=SSHear_4B_Spanish	91.3	84.2
SSHear_91=SSHear_91_Spanish	87.0	88.3
SSHear_14=SSHear_14_Spanish	91.6	79.9
SSHear_A9=SSHear_A9_Spanish	80.6	92.7
MFSpeak_A9=MFSpeak_A9_Spanish	79.7	89.0

Note: Labels are as follows (for example, row 1): SS= Saturday/Sunday, Speak_91= Language child speaks from 9 am to 1 pm, selected as Spanish.

Table 3. *Bilingual speakers cluster reporting*

	Sample %	*Cluster %*
SSSpeak_91=SSSpeak_91_Both	96.2	93.4
SSSpeak_4B=SSSpeak_4B_Both	93.8	94.1
SSSpeak_14=SSSpeak_14_Both	92.6	93.0
SSSpeak_A9=SSSpeak_A9_Both	97.0	84.5
MFSpeak_4B=MFSpeak_4B_Both	91.2	91.9
SSHear_4B=SSHear_4B_Both	80.5	93.0
SSHear_91=SSHear_91_Both	82.2	87.1
SSHear_14=SSHear_14_Both	78.2	91.5
MFSpeak_A9=MFSpeak_A9_Both	89.0	72.0
SSHear_A9=SSHear_A9_Both	85.2	76.8

Note: Labels are as follows (for example, row 1): SS= Saturday/Sunday, Speak_91= Language child speaks from 9 am to 1 pm, selected as Both.

Table 4. *English Speaking Cluster Reporting*

	Sample %	*Cluster %*
SSSpeak_4B=SSSpeak_4B_English	93.5	100
SSSpeak_91=SSSpeak_91_English	91.5	100
SSSpeak_A9=SSSpeak_A9_English	84.3	100
SSSpeak_14=SSSpeak_14_English	78.2	100
MFSpeak_4B=MFSpeak_4B_English	81.6	93.0
MFSpeak_A9=MFSpeak_A9_English	61.9	90.7
MFSpeak_14=MFSpeak_14_English	48.2	95.3
LanguageChild=LanguageChild_English	61.0	83.7
MFSpeak_91=MFSpeak_91_English	38.5	93.0
LanguageComfy=LanguageComfy_English	30.8	95.3

Note: Labels are as follows (for example, row 1): SS= Saturday/Sunday, Speak_4B= Language child speaks from 4 pm to bedtime, selected as English; LanguageChild is the language the child uses and LanguageComfy is the language parents report the child is most comfortable with.

In Tables 1-4, the percentage of the sample column indicates the percentage of individuals meeting the row criterion in the sample who are in cluster 1. The percentage of the cluster column indicates the percentage of individuals in the cluster who meet the criterion. So for example, 94.6% of the children who speak Spanish from 9A - 1P on the weekends are in cluster 1. Of the children in cluster 1, 96% speak Spanish on the weekends from 9A - 1P.

Results from the three dimension tables suggest the language which the children speak on the weekends from 9AM until bedtime are strongly tied to cluster membership. In general, cluster 1 could be classified as Spanish-speaking, cluster 2 as Bilingual, and cluster 3 as English-speaking. However, given that our sample of English-speaking students is very small, we did not include this cluster in the CSR as a meaningful factor derived from the IGDIs-E family survey, called the Language Exposure Evaluation Report (LEER). These findings indicate that survey questions could be streamlined to ask questions that are specific to the child the language speaks on the weekend.

Once we were confident in our dimensions, we wanted to examine the descriptives of these two primary groups as well as the remainder of the sample for the Year 3 assessments on IGDIs-E. As such, we included scores of children who reported speaking all Spanish on the weekend in the Spanish group, scores of children who reported speaking both languages on the weekend in the both group, and all other students' scores in the Mixed Level Bilingual (MLB) group.

Tables 5-19 contain the descriptive results by season. General trends observed indicated that performance trends varied based on the measure and season of the year. Small sample sizes do result in a fair amount of instability in results, limiting generalization.

Table 5. *Picture Naming/Identificación de los Dibujos Fall*

	N	*M*	*SD*	*Min*	*Max*	*Range*	*Skew*	*Kurtosis*	*SE*
Both	88	-0.79	1.56	-4.01	2.89	6.90	-0.71	-0.09	0.17
Spanish	90	0.59	1.15	-2.75	4.18	6.90	-0.17	0.64	0.12
MLB	94	-0.39	1.72	-4.01	4.18	8.19	-0.37	0.08	0.18

Table 6. *Picture Naming/Identificación de los Dibujos Winter*

	N	*M*	*SD*	*Min*	*Max*	*Range*	*Skew*	*Kurtosis*	*SE*
Both	100	-0.38	1.51	-4.01	4.18	8.19	-0.16	0.43	0.15
Spanish	101	0.84	1.28	-4.01	2.89	6.90	-0.92	1.53	0.13
MLB	90	0.19	1.73	-4.01	4.18	8.19	-0.07	0.13	0.18

Table 7. *Picture Naming/Identificación de los Dibujos Spring*

	N	*M*	*SD*	*Min*	*Max*	*Range*	*Skew*	*Kurtosis*	*SE*
Both	97	-0.45	1.68	-4.01	2.89	6.90	-0.54	-0.23	0.17
Spanish	94	1.01	1.28	-4.01	4.18	8.19	-0.70	2.06	0.13
MLB	90	0.42	1.93	-4.01	4.18	8.18	-0.30	0.14	0.20

Table 8. *Expressive Verbs/Verbos (Expresivo) Fall*

	N	*M*	*SD*	*Min*	*Max*	*Range*	*Skew*	*Kurtosis*	*SE*
Both	85	-1.58	1.60	-4.77	1.04	5.81	-0.62	-0.59	0.17
Spanish	89	-0.04	1.35	-3.50	3.69	7.20	-0.11	0.49	0.14
MLB	90	-1.01	1.87	-4.77	3.69	8.46	-0.63	-0.02	0.20

Table 9. *Expressive Verbs/Verbos (Expresivo) Winter*

	N	*M*	*SD*	*Min*	*Max*	*Range*	*Skew*	*Kurtosis*	*SE*
Both	61	-1.04	1.37	-4.77	1.58	6.35	-0.57	-0.08	0.18
Spanish	45	0.19	1.61	-4.77	3.69	8.46	-0.56	1.26	0.24
MLB	61	-0.77	1.91	-4.77	2.41	7.18	-0.71	-0.35	0.24

Table 10. *Expressive Verbs/Verbos (Expresivo) Spring*

	N	*M*	*SD*	*Min*	*Max*	*Range*	*Skew*	*Kurtosis*	*SE*
Both	90	-1.20	1.74	-4.77	2.41	7.18	-0.74	-0.23	0.18
Spanish	96	0.36	1.28	-4.77	2.41	7.18	-1.29	2.63	0.13
MLB	88	-0.54	1.92	-4.77	3.69	8.46	-0.67	0.37	0.20

Table 11. *First Sounds/Primeros Sonidos Fall*

	N	*M*	*SD*	*Min*	*Max*	*Range*	*Skew*	*Kurtosis*	*SE*
Both	83	0.15	1.19	-3.81	4.01	7.82	0.46	2.22	0.13
Spanish	75	-0.11	1.21	-3.81	4.01	7.82	0.28	1.32	0.14
MLB	86	0.23	1.24	-2.59	4.01	6.60	0.99	1.05	0.13

Table 12. *First Sounds/Primeros Sonidos Winter*

	N	*M*	*SD*	*Min*	*Max*	*Range*	*Skew*	*Kurtosis*	*SE*
Both	58	0.70	1.70	-1.81	4.01	5.82	0.67	-0.73	0.22
Spanish	71	0.30	1.47	-1.81	4.01	5.82	0.75	-0.08	0.17
MLB	60	0.66	1.69	-1.81	4.01	5.82	0.65	-0.70	0.22

Table 13. *First Sounds/Primeros Sonidos Spring*

	N	*M*	*SD*	*Min*	*Max*	*Range*	*Skew*	*Kurtosis*	*SE*
Both	91	0.64	1.41	-2.59	4.01	6.60	0.45	-0.19	0.15
Spanish	89	0.57	1.40	-1.81	4.01	5.82	0.57	-0.54	0.15
MLB	87	0.73	1.49	-1.81	4.01	5.82	0.56	-0.40	0.16

Table 14. *Letter Naming/*Identificación de las Letras *Fall*

	N	*M*	*SD*	*Min*	*Max*	*Range*	*Skew*	*Kurtosis*	*SE*
Both	85	-0.08	1.05	-3.45	1.97	5.42	-1.07	2.05	0.11
Spanish	84	0.21	1.18	-3.45	3.27	6.72	-0.76	1.82	0.13
MLB	92	-0.20	1.11	-3.45	3.27	6.72	0.35	1.24	0.12

Table 15. *Letter Naming/*Identificación de las Letras *Winter*

	N	*M*	*SD*	*Min*	*Max*	*Range*	*Skew*	*Kurtosis*	*SE*
Both	96	0.37	1.36	-2.23	3.27	5.51	0.18	-0.44	0.14
Spanish	92	0.51	1.38	-3.45	3.27	6.72	-0.46	0.87	0.14
MLB	97	0.48	1.46	-3.45	3.27	6.72	0.14	-0.41	0.15

Table 16. *Letter Naming/*Identificación de las Letras *Spring*

	N	*M*	*SD*	*Min*	*Max*	*Range*	*Skew*	*Kurtosis*	*SE*
Both	96	0.59	1.21	-2.23	3.27	5.51	0.19	-0.12	0.12
Spanish	95	0.70	1.51	-3.45	3.27	6.72	-0.09	-0.03	0.16
MLB	88	0.61	1.43	-3.45	3.27	6.72	-0.24	0.25	0.15

Table 17. *Sound Identification/Identificación de los Sonidos Fall*

	N	*M*	*SD*	*Min*	*Max*	*Range*	*Skew*	*Kurtosis*	*SE*
Both	80	0.01	1.34	-3.33	3.13	6.45	-0.16	-0.21	0.15
Spanish	67	-0.28	1.38	-3.33	3.13	6.45	-0.22	0.00	0.17
MLB	89	-0.02	1.38	-3.33	4.35	7.67	0.21	1.12	0.15

Table 18. *Sound Identification/Identificación de los Sonidos Winter*

	N	*M*	*SD*	*Min*	*Max*	*Range*	*Skew*	*Kurtosis*	*SE*
Both	59	0.75	2.13	-3.33	4.35	7.67	0.21	-0.91	0.28
Spanish	71	-0.01	1.72	-3.33	4.35	7.67	0.14	0.12	0.20
MLB	59	0.64	1.98	-3.33	4.35	7.67	0.17	-0.33	0.26

Table 19. *Sound Identification/Identificación de los Sonidos Spring*

	N	*M*	*SD*	*Min*	*Max*	*Range*	*Skew*	*Kurtosis*	*SE*
Both	91	1.07	1.93	-3.33	4.35	7.67	0.26	-0.58	0.20
Spanish	93	0.57	1.71	-3.33	4.35	7.67	0.38	-0.11	0.18
MLB	88	1.06	1.78	-3.33	4.35	7.67	-0.03	-0.53	0.19

Language of instruction. The language in which a child is taught is another important factor in understanding how to assess and support early language and literacy development. Evidence suggests that many classroom level factors influence the degree to which SEB children can acquire and develop strong skills in English and Spanish. Specifically, language of instruction has been show to impact academic skills of SEBs based on the quantity and quality of that instruction in Spanish (Burchinal, Field, López, Howes & Pianta, 2012; Rolstad, Mahoney, & Glass, 2005). Quality of instruction is particularly important as teacher-child interactions have strong relationships to child level outcomes. (Hindman & Wasik, 2015; Justice, Mashburn, Hamre & Pianta, 2008; Mashburn et al., 2008; Hamre & Pianta, 2005). Further, when high quality language techniques, such as language modeling, are embedded in instruction, such approaches improve child-level outcomes in language and literacy for Latino and SEBs (Downer et al., 2012).

Claim 8:
IGDIs-E are appropriate for use with children who are native Spanish speakers, with exposure to Spanish ranging from Spanish dominant to English dominant.

During IGDIs-E development we explored the role of language of instruction by examining teacher reported language and scores on the Classroom Assessment and Scoring System (CLASS). We conducted a hierarchical linear model at three levels where the full model is specified as follows:

Claim 11:
IGDIs-E can be used within classrooms that use a variety of models of language of instruction (from English only, to some Spanish, to a balance of both).

Level 1: $$\text{Score}_{ijk} = \pi_{0jk} + \pi_{1jk}(Season)_{ijk} + e_{ijk}$$

Level 2: $$\pi_{0jk} = \beta_{00k} + r_{0jk}$$
$$\pi_{1jk} = \beta_{10k} + r_{1jk}$$

Level 3: $$\beta_{00k} = \gamma_{000} + \gamma_{001}(\text{Lang})_k + \gamma_{002}(\text{LMScore})_k + u_{00k}$$
$$\beta_{10k} = \gamma_{100} + \gamma_{101}(\text{Lang})_k + \gamma_{102}(\text{LMScore})_k + u_{10k}$$

γ_{000} represents the predicted initial status (Fall) for an English classroom (Lang = 0); and γ_{100} represents the seasonal learning rate for an English classroom. Results, depicted in Figures 3-6 indicate that there were initial differences between groups.

Language of instruction had a significant association with English Picture Naming's initial status, English and Spanish Sound identification's initial status and English Sound identification's growth rate. For the rest of the parameters, our data did not detect empirically significant differences. In total, when the language of instruction coefficients are significant, variance explained in level-3 variance components from unconditional to final model is large (35% for PN English).

Figure 3. Effects of Language of Instruction on Picture Naming English Growth

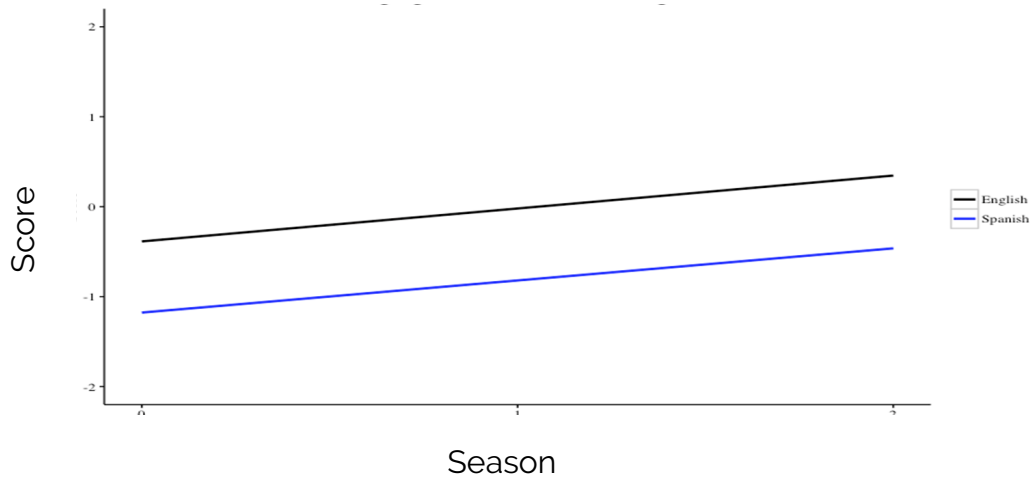


Figure 4. Effects of language Modeling on PN English Growth by language of instruction

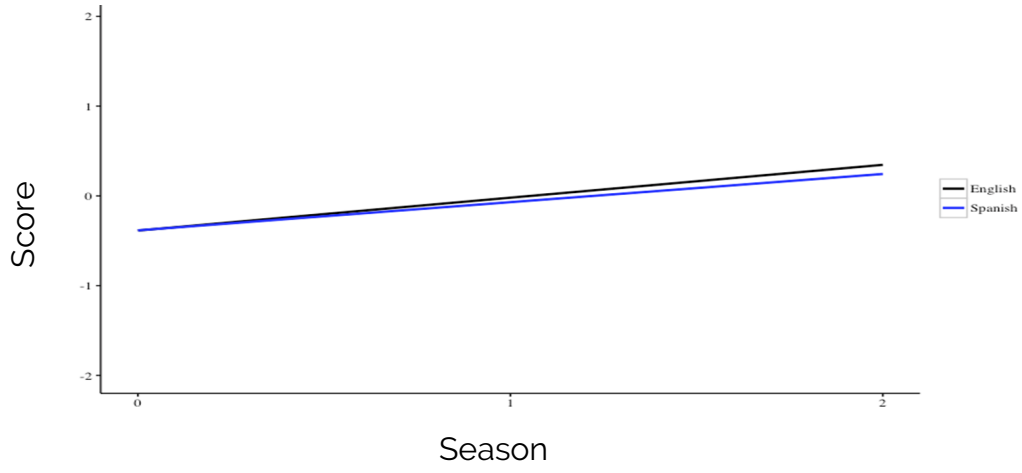


Figure 5. Effects of Language of Instruction on PN Spanish Growth

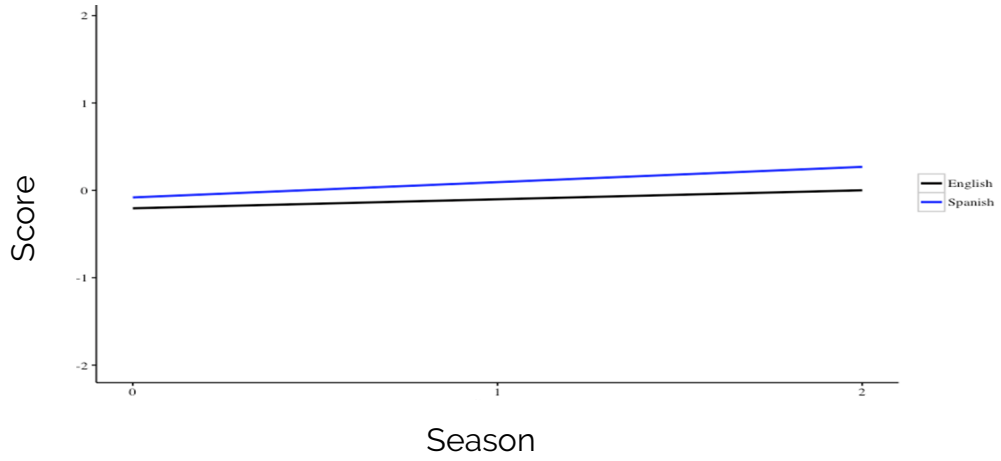
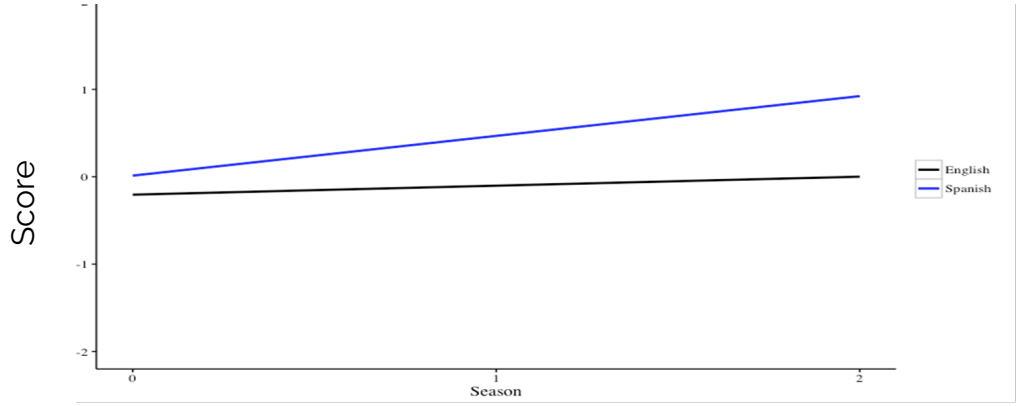


Figure 6. Effects of Language Modeling on PN Spanish Growth by Language of Instruction



Claim 11: *IGDIs-E can be used within classrooms that use a variety of models of language of instruction (from English only, to some Spanish, to a balance of both).*

Test Administration, Scoring, and Interpretation

Intended Audiences
The IGDIs-E are designed to be used by early childhood preschool programs serving 4-5-year-old Spanish-speaking children. The IGDIs-E can be administered to any child that can pass the sample items in each subtest including children with disabilities. Children can also have varying levels of Spanish proficiency as our development team carefully explored the performance of both simultaneous and sequential bilinguals and children who speak various dialects of Spanish.

Testing Environment
The test administrator needs a low table in a quiet area for testing each child individually. There should be minimal surrounding noise or activity to avoid distracting the child. Before starting the test, the test administrator may converse with the child in Spanish. When the actual testing begins, the test administrator must administer the entire test in Spanish only.

Average Testing Time
The average testing time for each of the IGDIs-E measures is 5 minutes. Time will vary depending on the conciseness of the child's responses.

Test Administration
Each of the five measures, Oral Language (2), Phonological Awareness (1), and Alphabet Knowledge (2) includes four sample items labeled Ejemplo A-D, followed by 15 test items numbered on the top right corner of each card. Items are always presented in the same order. During administration, each card is presented one at a time by holding the card up in front to show the pictures to the child. The administrator points to each picture as they label it given the standardized prompt printed in red on the back of the card.

Please see the Administrative Manual for specific guidelines for each measure.

Level of Spanish proficiency required for test administrators
Personnel in programs who are administering the IGDIs-E should be native Spanish speakers or have near-native proficiency. Although the administration prompts are clearly scripted, it is important that test administrators can produce Spanish that is clear, with sounds that are accurately articulated so as to be easily understood by preschoolers. Test administrators must also be able to understand child responses and score them accurately.

Claim 2: *IGDIs-E are designed to align with general outcome measure standards including: ease of use, scores related to meaningful long-term out comes, quick and efficient to deliver, meaningful score interpretation and inexpensive or easily accessible.*

Claim 14: *IGDIs-E scores can be used to inform instructional planning.*

SEBs know when to use each of their languages with different communicative partners, a skill called interlocutor sensitivity. Interlocutor sensitivity develops as early as 18 months (Pettito, Katerlos, Levy, Guana, Tetretault, & Ferraro, 2001). Given that even very young bilinguals are sensitive to the native language of the person with whom they are communicating, it is recommended that assessors should have native or near native fluency. For example, if the child assumes that the assessor speaks English, then he/she may conclude that it is appropriate to speak English with that assessor, even when an assessment is being conducted in Spanish. Although potentially difficult to accomplish in all early childhood programs, every effort should be taken to hire assessors with the appropriate Spanish language skills to elicit children's optimal performance. Further, it is important to test each language separately with different examiners and on different days if possible (i.e., a different assessor for each language on different days). This approach is most likely to elicit the child's best performance in each language.

Scoring Instructions
For each of the five measures, Oral Language (2), Phonological Awareness (1) and Alphabet Knowledge (2), the child can respond by pointing to the answer or by saying the answer. After the child responds to each item, the administrator circles the response given by the child on the score sheet. The administrator checks the box labeled *DK/NR* if child indicates they don't know or provides no response. Exact instructions for scoring each item are provided in the script protocol.

Please see the Administrative Manual for specific guidelines for scoring each measure.

Score Interpretation
Scores from each task indicate a child's ability level. A low score indicates that a child may be struggling with the domain represented by the specific task. IGDIs-E data can be used to determine if a child is on target in their Spanish early literacy development or if they are in need of more instructional support (potentially at a targeted-Tier 2 level, or at an intensive-Tier 3 level).

The IGDIs-E are designed to be used in complement with the English IGDIs, such that together each test provides valuable information to create an accurate picture of the child's total language profile. It is important to note that when testing SEBs, IGDIs-E should not be used without an English complementary measure if the long-term goal is for the child to be successful in English.

Claim 2:
IGDIs-E are designed to align with general outcome measure standards including: ease of use, scores related to meaningful long-term out comes, quick and efficient to deliver, meaningful score interpretation and inexpensive or easily accessible.

Claim 14:
IGDIs-E scores can be used to inform instructional planning.

Design Principles and Quality Indicators

Avoiding Bias through Intervention and Instructional Alignment

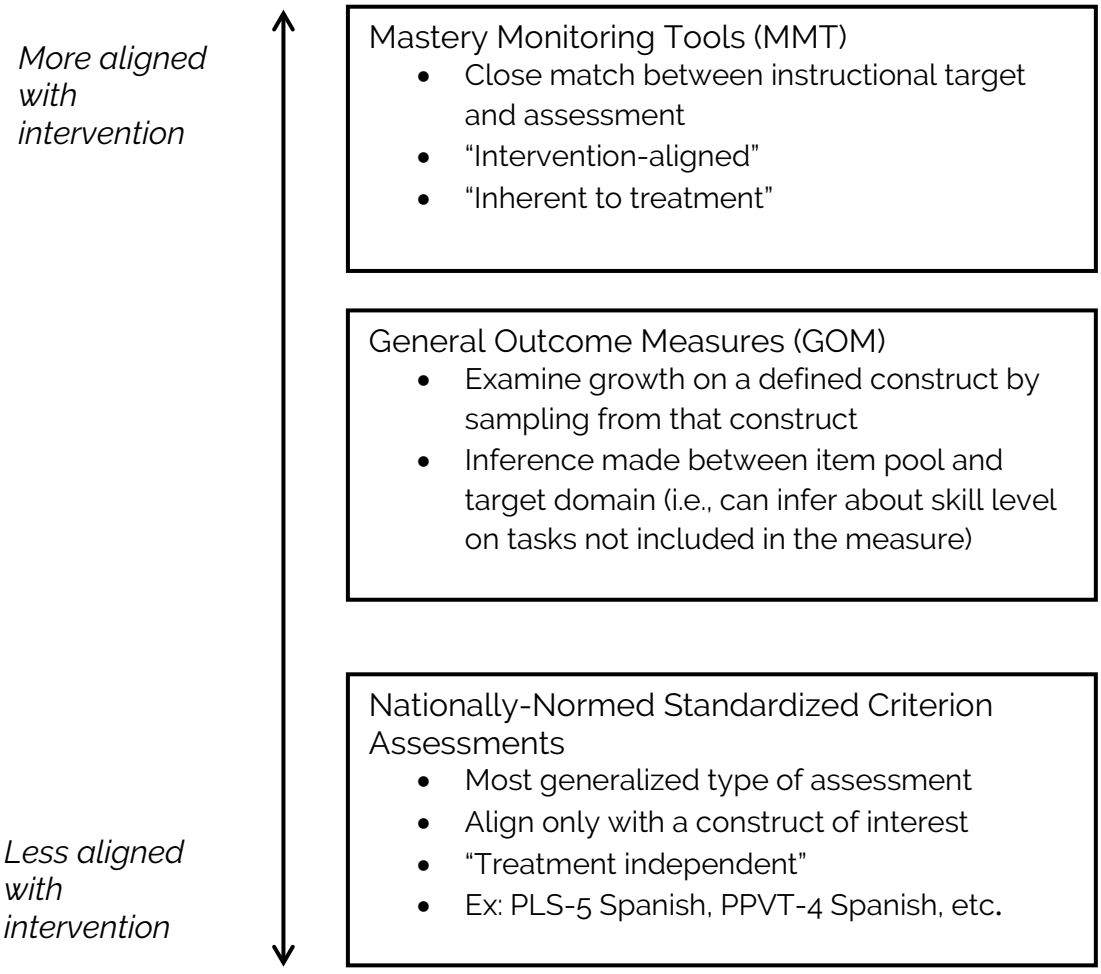
One common approach to measurement design is to ensure that measures are sensitive to intervention or instruction. The degree of alignment between the assessment and intervention can be conceptualized on a continuum where at one extreme the assessment includes items that are a direct match to the intervention targets, and at the other extreme there is only a generalized conceptual connection between the items of the assessment and the general domains represented in the intervention or instructional practice (Slavin & Madden, 2011). Conceptual arguments suggest that the weaker the alignment, the less sensitive to growth in the measurements (Polikoff, 2010).

One approach, referred to as Mastery Monitoring Tools (MMTs), represents an attempt to secure a close match between the instructional target and the assessment, illustrating a case example of the first extreme. In fact, there may be so much of an assessment-to-instruction match that one might argue that increases in performance on MMT measures are misleading because they may not generalize to the domain skill set. In this way, MMTs have been described as "intervention-aligned" or "inherent to treatment" measures (Slavin & Madden, 2011) such that students who receive intervention or instruction that matches the measure of interest have a distinct advantage in responding to the items.

In contrast, another approach, general outcome measurement, includes measures that examine growth and status on a defined construct by sampling from the construct. Kane (2013) argued that an important inference in a validity framework is the extrapolation from the universe of generalization (item pool) to the target domain. The extrapolation inference allows us to move from simple claims about test performance to claims about the full range of Spanish early language and literacy abilities on tasks not represented in the IGDIs-E.

The validity argument supporting the extrapolation inference includes evidence about item specifications (how items were developed), item scoring, item functioning, and item selection. Moving along to the center of the continuum, a measure that represents a more generalized approach to assessment that balances sensitivity to growth and generalization to the domain of interest may be beneficial (see Figure 7). At the other extreme of the continuum are the most generalized types of assessments, generally categorized as nationally-normed standardized criterion assessments (e.g. Preschool Language Scale-5 in Spanish; Peabody Picture Vocabulary Test-4 in Spanish). These types of measures assess Spanish early literacy and language skills as "treatment independent", demonstrating alignment only with the construct of interest.

Figure 7. The continuum of types of assessment tools based on their alignment to treatment or intervention.



Although each type of measure on the continuum may have utility in practice for specific purposes, we argue that measures that examine more generalized performance may be especially salient for SEB students because of the evidence for cross-linguistic transfer. Researchers suggest that a student's general early literacy and language knowledge (at the domain sub-skill level) in Spanish can support their performance and language acquisition in English (Cárdenas-Hagan et al., 2007; Gutierrez, Zepeda & Castro, 2010; Kuo & Anderson, 2010; Mancilla-Martinez & Lesaux, 2010; Oller & Eilers, 2002). Few if any researchers have examined the role of specific instructional targets in cross linguistic transfer. That is, evidence does not examine the degree to which specific words, phrases, units of grammar or syntax contribute to cross-linguistic transfer; instead, a more generalized approach is utilized (Cárdenas-Hagan et al., 2007; Durgunoğlu et al., 1993; Melby-Lervag & Lervag, 2011). Further, the gap between treatment dependent (MMTs) and treatment independent (standardized) measures is wide and apparent in the field of SEB research.

Claim 2:
IGDIs-E are designed to align with general outcome measure standards including: ease of use, scores related to meaningful long-term out comes, quick and efficient to deliver, meaningful score interpretation and inexpensive or easily accessible.

Challenges are present at both ends of the continuum, as many standardized measures are expensive, lengthy and cumbersome to deliver, and have a history of resulting in small effects when examining instructional practices or interventions (Barrueco et al., 2012). Although, MMTs limit the generalizability of skill acquisition and thus limit understanding of the role of cross-linguistic transfer for students in bilingual classrooms. Indeed, more generalizable measures are needed to adequately examine student performance within the context of an RTI model and to evaluate instruction and intervention.

Claim 13:
IGDIs-E were uniquely designed to attend to how Spanish language develops rather than by translating existing English measures.

Creating Measures to Fill the Gap: General Outcome Measures

Historically, measures of growth and development for children have been characterized across the continuum of alignment previously described, including MMTs or general outcome measures (Fuchs & Deno, 1991). For measures of development toward a long-term outcome, where change is indexed to a common scale or metric of growth (i.e., change over time), general outcome measures offer clear and distinctive advantages (Fuchs & Deno, 1991; McConnell & Greenwood, 2013).

General outcome measures are brief, easy-to-administer and psychometrically-robust measures of child achievement in a single developmental or academic domain. As a class, general outcome measures have been the subject of substantial research, development, evaluation, and adaptation (c.f., Deno, 1985, 1997, 2003; Fuchs & Deno, 1991; McMaster & Espin, 2007; Shinn, 1998; Stoner, Carey, Ikeda, & Shinn, 1994; Wayman, Wallace, Wiley, Tichá, & Espin, 2007). In short, contemporary researchers and practice standards suggest that GOMs have four defining features (Fuchs & Deno, 1991).

First, GOMs must be conceptually and empirically related to other, socially appropriate and meaningful (and typically future-oriented) outcomes. This essential characteristic of GOMs is central to their overall utility; like any developmental or academic measure, GOMs are not inherently valid, but rather we must validate each interpretation and use of a measure, including uses such as describing growth toward a desired outcome (Kane, 2013).

Second, GOMs must be cost-effective and easy to collect. In most applications, GOMs are used either for broad-scale screening of large groups of students and/or monitoring individual children's growth during periods of more intensive intervention. In both cases (and others), measures must be logistically feasible to implement to allow for broad-scale, frequent, and affordable administration.

Third, GOMs must be repeatable and sensitive to growth, ideally over relatively brief periods of time (with the potential to employ parallel or equivalent forms; Albano & Rodriguez, 2012). Because these measures are often used to assess growth at the individual and group level, they must detect relatively fine differences in performance. GOMs are used both to identify children who would benefit from additional or more intensive intervention and to assess effects of intervention in short enough time-frames that important adjustments can be made.

Fourth, GOMs must produce data that are direct and easy to interpret, and lead to clear actions on behalf of teachers and others. These are measures intended to describe and support improved intervention; as a result, the scores and data produced must support instructional monitoring and decision-making by teachers and others.

Claim 2:
IGDIs-E are designed to align with general outcome measure standards including: ease of use, scores related to meaningful long-term out comes, quick and efficient to deliver, meaningful score interpretation and inexpensive or easily accessible.

During assessment development, evidence contributed from measurement design and empirical approaches and contextual and applied approaches all influence our claims about validity, and as a result, directly impact the uses and applications of the measure. In best practice, early childhood measures for SEB students will attend to each contribution and continually evaluate how it impacts the use and interpretations of the tool.

The IGDIs-E are GOMs designed for the universal screening of the early language and literacy skills of Spanish-speaking preschoolers. Universal screening is conducted in general education settings three times a year to identify children who may be in need of more intensive instructional supports at the targeted (e.g. Tier 2) and intensive (e.g. Tier 3) levels. Universal screening is not diagnostic testing designed to identify language delay or impairment, but rather assessment that is directly tied to instructional decision-making. GOMs are the most common type of measure used within Multi-tiered Systems of Support (MTSS) because they are designed to efficiently measure instructionally relevant targets that are meaningful in predicting school performance (Fuchs & Fuchs, 2006).

Claim 2:
IGDIs-E are designed to align with general outcome measure standards including: ease of use, scores related to meaningful long-term out comes, quick and efficient to deliver, meaningful score interpretation and inexpensive or easily accessible.

Multi-Tiered Systems of Support

Multi-tiered Systems of Support (MTSS), also known as Response to Intervention models, provide instructional supports to students at three tiers: (a) universal core curriculum, (b) targeted intervention or (c) intensive intervention (Buysse & Peisner-Feinberg, 2013; Greenwood, Bradfield, Kaminski, Linas, Carta, & Nylander, 2011). Implementing MTSS in early childhood settings increases the need for universal screening tools that can be used with SEB preschool children. At the core of any MTSS model is the prevention of academic failure through the delivery of targeted instruction that is designed to meet the learning needs of every child. Central to this goal in a MTSS approach is the use of screening tools that provide robust and meaningful information regarding children's ability levels. As MTSS models are implemented more frequently in early childhood settings, the need for assessments that yield appropriate inferences and adequately trained professionals to administer those assessments will intensify. Thus, improving screening practices for SEB children is critical within this emerging MTSS framework in early childhood settings. Administering the IGDIs-E three times a year is an important and integral component of implementing high quality MTSS models with SEB preschool children.

Claim 14:
IGDIs-E scores can be used to inform instructional planning.

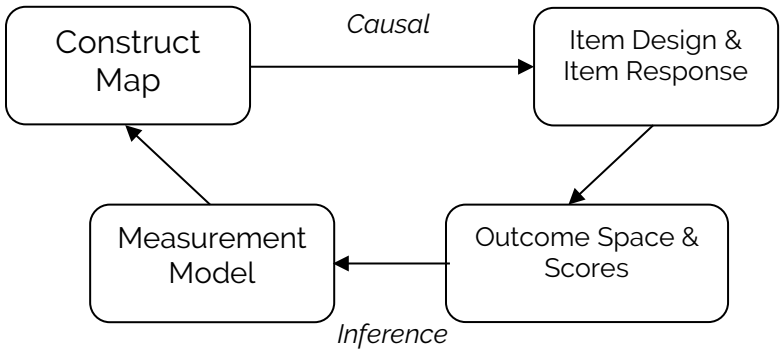
The Measurement Design Model

To develop IGDIs-E we utilized Wilson's (2005) constructing measures framework as a roadmap for assessment design. Wilson's framework allows for the conceptual foundation of a measure or task (in this case the interpretation and use argument) to be the driving force in task and item development. In this way, the assessment is built with the measurement model as an integral part of the process. Instead of building measures and then finding a measurement model to fit, researchers develop tasks with a measurement model in mind to allow for clear and consistent interpretations and uses.

Claim 1:
The IGDIs-E are psychometrically sound and theory-based, using Mark Wilson's constructing measures model and Rasch modeling for empirical item level statistics.

Wilson's model allows for a clear translation from conceptual and theoretical perspectives to an end product. Four building blocks incorporate relevant components of a validity argument at each step of the process: (a) define the construct of interest by creating a construct map; (b) use the map to support item design; (c) select and defining parameters for evaluating responses to the items, or outcome space; and (d) describe and interpret responses, or the measurement model. These building blocks allow inferences to be made about the constructs of interest (see Figure 8). As such, the model is intended to be cyclical by using the information defined in each building block to further refine the assessment.

Figure 8. The four building blocks of an item response model approach to measurement construction (adapted from Wilson, 2005).



Confirmatory evidence for validity suggests the measure is appropriately constructed; dis-confirmatory evidence suggests revisions and refinement must be made. Within Wilson's model validity is realized as a causal line from construct definition to item responses (a child's ability on the construct causes their responses to items) and the inferential line from outcome space definition through

the measurement model back to the construct (we score item responses, scale them through the measurement model, and make inferences about ability on the construct). This interpretive argument enables meaningful and useful inferences about a child's ability regarding Spanish early language and literacy based on item responses.

Defining the construct of interest. To establish the foundation of a measure, a conceptually strong presentation of the construct being measured is needed. Wilson's model recognizes this step as the construct map, presenting a conceptual representation of an underlying latent cognitive skill that exists along a continuum of no skill to advanced skill.

Specifically, defining and operationalizing the constructs of interest and developing the validity arguments that support intended interpretations and uses of the measure creates a conceptual anchor that influences each decision that is made within the assessment development process. Further, it is important to note that the construct map is entirely defined in regard to the interpretations presented within the conceptual argument. As such, the goal of the assessment, (whether it be screening as discussed here, or another goal such as diagnostic assessments), informs the nature and complexity of the construct of interest. This approach focuses on meaningful operationalization of the construct in support of the intended interpretations and uses of scores. See Figure 9 for the IGDIs-E phonological awareness construct map.

Phonological awareness. We define phonological awareness as the meta-linguistic ability to understand that spoken words are comprised of small sound units; to detect, discriminate between, and manipulate these structural components; and to perform these skills independent of word meaning (Anthony et al., 2011; Branum-Martin et al., 2006; Cardenas-Hagan, Carlson & Pollard-Durodola, 2007; Cisero & Royer, 1995; Durgunoglu, Nagy & Hancin-Bhatt, 1993; Gorman & Gillam, 2003; Kuo & Anderson, 2010).

Claim 1:
The IGDIs-E are psychometrically sound and theory-based, using Mark Wilson's constructing measures model and Rasch modeling for empirical item level statistics.

Claim 3:
IGDIs-E assess three different yet related domains of early literacy in Spanish- alphabet knowledge, oral language, and *phonological awareness.*

When defining constructs for Spanish-English bilinguals we argue that it is critically important to recognize the contribution of Spanish as a unique and complementary construct to English, rather than a translated equivalent to English. Others in the field have proposed methodologies for creating measures that have translated equivalents. We argue that Spanish has its own similar, but different trajectory than English, suggesting equating, and thus translating, is not an appropriate practice. Of those measures that are currently available for screening, none provide scales that can be compared in each language by examining the relative, but different, developmental trajectories in English and Spanish (Barrueco et al., 2012). As a result, we turn our attention to defining constructs that are not translated equivalents of English. Instead, with the conceptualization of Spanish language and literacy through operationally defined constructs, the measure development process proceeds to the creation of items as manifestations of these constructs.

Claim 7: *IGDIs-E are inclusive of a variety of Spanish dialects and socio-economic backgrounds that are representative of Spanish speaking populations.*

Item design. When designing items for SEB students it is important to attend to the features that allow for successful and meaningful interaction with the content. In practice, it is important to carefully select the images represented on items in each task with respect to fidelity to the construct, ensuring adequate construct representation and minimizing the presence of construct-irrelevant features. For SEB students, construct irrelevant features may be manifested or examined in at least three ways: cultural variability, target and distractor variables, and through differential item functioning.

Cultural variability. Cultural variability represents the differences in perceived cultural interpretations of a stimulus both within and across cultures. Specifically, some stimuli within items (e.g. images, illustrations, etc.) may be interpreted differently with the lens of differing cultures. For example, a picture of a slice of white bread in English might be interpreted as "pan" in Spanish. However, when the word "pan" is used in many native Mexican Spanish-speaking communities, it references a sweet round bread, not the standard sandwich variety. This inter-cultural variability must be addressed to prevent information in a given item from being misrepresented and resulting in inappropriate testing interactions. At the same time, we must also attend to stimuli within items that may be interpreted differently within a single cultural variable.

For example, although many families who engage in Spanish measures of early language and literacy may identify as Hispanic or Latino, they also may represent different dialectal regions. At least four major dialectal regions exist: Caribbean Spanish, Mexican Spanish, South American Spanish, and Central American Spanish. Across these dialectal regions, many items function appropriately, however, for others there are unique terms within region. For example, when shown an image of an orange, Mexican Spanish speaking students often respond with "naranja" or "mandarina"; however, Caribbean Spanish students respond with "china". To support claims that a measure of Spanish early language and literacy appropriately measures SEB student performance, we must attend to inter- and intra-culture variability.

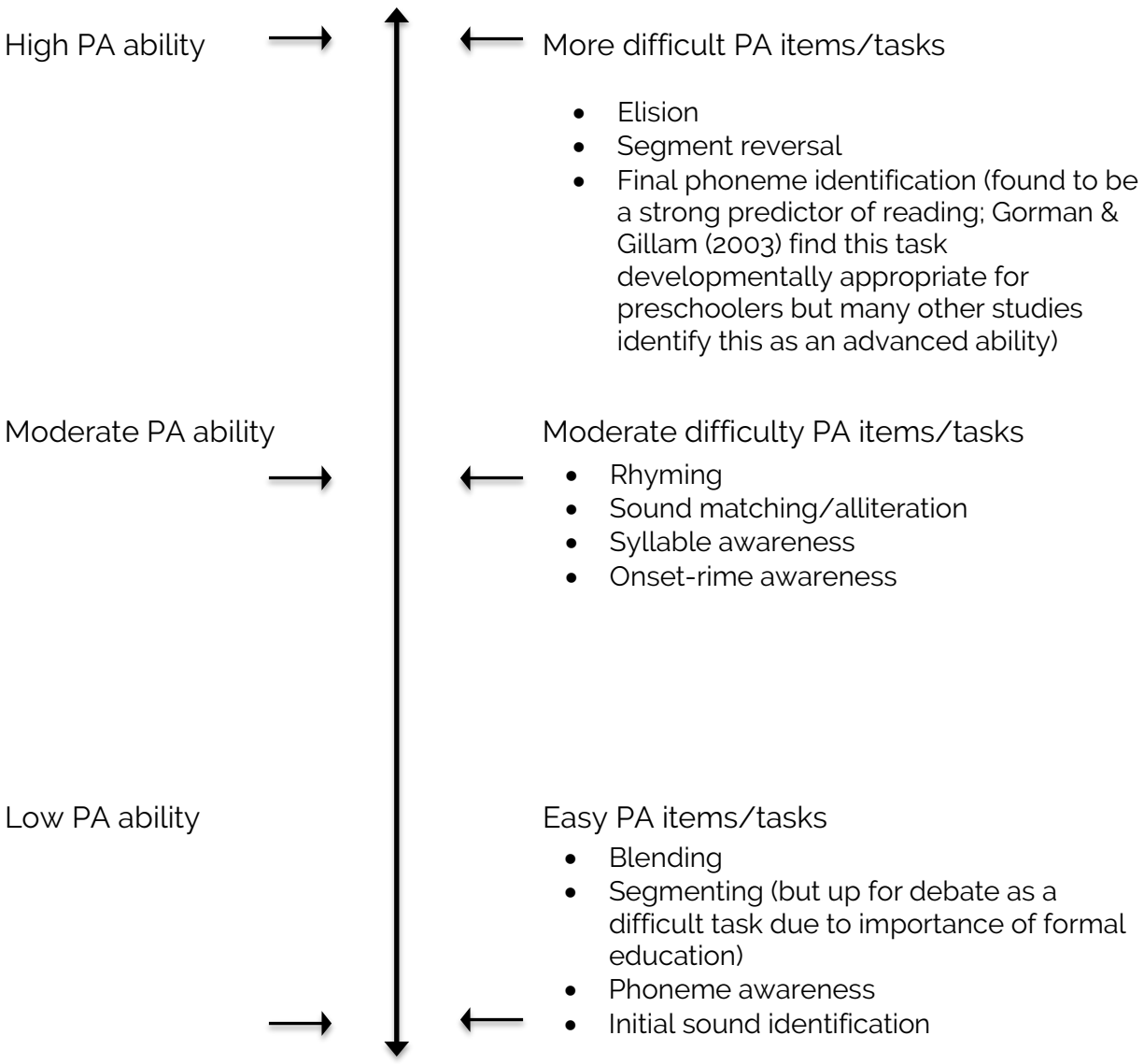
Target and distractor variables. When designing items, it is not only important to attend to the actual images and responses presented in the stimulus as the target, but also, when applicable (i.e. receptive tasks), to attend to the distractors that may be present within the item. For example, in a task that requires students to select the image that matches a target initial syllable provided, the child may also interact with one, two or three other distractors. The quality of these distractors must be addressed to ensure that the item contributes to the validity argument effectively. That is, if we do not control the distractors, then it is difficult to understand if the student selects the target or a distractor because of the information we intended them to use to make the decision, or possibly because of other information present in the distractor(s). Consider the following situation: a student is presented with an item that features a door (puerta), an airplane (avión) and hog (cerdo). If the administrator asks the student to find the word that begins with "puer", we would anticipate that the student would select the door ("puerta"). However, in some dialectal regions a hog is known as a "puerco". If the student draws on this experience he or she may experience some cognitive dissonance in this item, and could potentially choose the cerdo/puerco. By examining these factors in distractors we can limit construct-irrelevant features in items, and thus maximize our ability to contribute to the validity argument effectively.

Differential item functioning. Differential item functioning (DIF) can be used to examine how items perform at the group level for a given group (e.g. gender, socio-economic status, regional dialectal groups, etc.) to determine the degree to which characteristics of group membership interact with item-level functioning.

Claim 1:
The IGDIs-E are psychometrically sound and theory-based, using Mark Wilson's constructing measures model and Rasch modeling for empirical item level statistics.

Construct Map

Figure 9. Phonological awareness construct map example.



In this way we can determine if group-level factors are influencing item performance which may indicate inadvertent influences in test design, introducing variability or bias that may otherwise be unaccounted for.

As with any meaningful assessment, a test should function equally well for all students who interact with the items. When all other item level characteristics are common, a test with robust evidence for validity demonstrates limited DIF in identified group membership. Specifically, DIF analysis examines the degree to which items function differently across a priori group membership, given overall ability level. In our measurement model, described below, each item is an indicator of the construct domain and performance on the item is a simple function of ability in the domain. When individuals with the same ability but from different groups perform differently on any given item, this suggests that group membership is affecting item response, not just ability in the domain. This violates the assumption that each item is an indicator of the same construct, since now the DIF item is also functioning as an indicator of group membership, differentially performing. This suggests that the item is potentially biased, capturing construct-irrelevant factors (group membership), perhaps privileging one ethnic group, language group, or gender, based on group-specific knowledge or experience.

We do not want test items to measure group-specific knowledge or experience, but to be exchangeable indicators of the common domain. By exploring DIF for relevant groups, tests can be designed to limit bias and reduce instances of construct-irrelevant variance, which directly contributes to robust evidence to support the validity arguments – that the test is a measure of the intended domain and not student characteristics.

Outcome space. In the typical interpretation and use argument, we find an inference from the observable behavior (child responses) of interacting with the item, to the score a child receives. The conversion from this observable behavior to a score that recognizes information regarding the construct is the outcome space. This scoring inference assumes the appropriateness of scoring criteria or scoring rules – that we give credit to responses that are true indicators of the child's trait level on the construct.

Claim 1:
The IGDIs-E are psychometrically sound and theory-based, using Mark Wilson's constructing measures model and Rasch modeling for empirical item level statistics.

This demands that we recognize the factors presented in item design (cultural variability, target and distractor analysis and DIF) so that we do not attribute high scores to irrelevant or biased responses. By identifying construct-aligned responses as high value responses, we can best support the inference to the broader domain and meaningfully contribute to evidence for the validity arguments.

Measurement model. As the final contribution in measure design, the measurement model must be specified. Here we discuss the Rasch model as an ideal candidate for use in assessments that feature single latent-trait constructs.

Rasch modeling (Rasch, 1980) provides a methodology for scaling and scoring a test under a latent-trait model, assuming that the item responses are due to a single underlying ability – Spanish language and early language and literacy in three domains.

The Rasch model provides a guiding framework for measure development in that it allows the construction of a measure that is aligned with the construct of interest. In doing so, we can evaluate items and item features (characteristics of items such as number of elements presented and complexity of the task) vis-à-vis the item difficulty, its location on the continuum of the underlying trait present in a given construct. This allows us to defend the construct, as the items theorized to be easier actually turn out to be easier and those items that were theorized to be more difficult are empirically more difficult. This mapping to the construct provides important content and construct-related validity evidence, supporting the intended interpretation of the score scale and the underlying ability regarding Spanish early language and literacy.

There are several statistics provided with Rasch analysis that allow for the evaluation of item functioning and of the measure as a whole. Indicators of item fit are provided, answering the question: How does this item contribute to the construct? An item fits the Rasch model if children with lower ability tend to respond incorrectly and children with higher ability tend to respond correctly as evaluated by in-fit and out-fit statistics. If item responses are random or are not consistent with the underlying ability, then item fit indices will be large, suggesting the item is problematic.

Claim 1:
The IGDIs-E are psychometrically sound and theory-based, using Mark Wilson's constructing measures model and Rasch modeling for empirical item level statistics.

Similarly, person fit statistics are provided, answering the question: Does the child response behavior fit the Rasch model? A child fits the Rasch model if the child correctly responds to items that are easier, given their estimated ability, and responds to more difficult items incorrectly – that their ability predicts the correctness of their responses to each item. If a child's item responses are random (i.e. due to guessing) or are not consistent with their underlying ability, the person fit index will be large, suggesting the child's responses do not fit the model and potentially, a problem with the items.

With the Rasch model, analyses are available that allow for the evaluation of mean ability statistics for each item response, to answer the question: to what degree does a given item aid in identifying high and low achieving students? The Rasch model assumes that items are uniform in their discrimination. This does not mean that the discrimination of all items is exactly the same, but that the variation in discrimination does not distort our interpretation of scores across the score scale. The measurement construction process is designed to identify items that do demonstrate differences in ability levels based on each score value so they can be eliminated or revised.

Beneficial features of the Rasch model. There are several benefits for using the Rasch model in the item design process. First, Rasch has the benefit of providing a metric where students and items are on a common scale. Generally speaking, person ability and item difficulty are measured on a logistic metric, where the average difficulty of items is centered at 0 and items range from requiring low levels of ability (-4) to high levels of ability (+4). Because of the difficult-to-interpret nature of the scale score, person scores are typically transformed into a metric that has more general appeal and can support other intended uses of scores. However, for the purposes of this manual, we report Rasch model statistics in logit units.

Second, if the model fits the data, the estimation of child ability is invariant across items, such that the child's score does not depend on which items were administered, as all items are scored on a common scale (Andrich, 2004; Sick, 2010). A child's ability is determined as it relates to the ability required to get the items correct. Item difficulty is then defined as the ability required to respond correctly to the item – not the proportion of children that respond correctly as in classical test theory.

Claim 1:
The IGDIs-E are psychometrically sound and theory-based, using Mark Wilson's constructing measures model and Rasch modeling for empirical item level statistics.

Thus, children's scores are based on their modeled ability rather than on normative performance of the items sets they received. When the model fits the data, item difficulty is invariant across children.

Third, when the Rasch model is selected a priori, measures can be constructed to support the model by attending to model assumptions during item design. This approach promotes a model for constructing measures, rather than a model of post-hoc data analysis. Although it is true that using 2PL or 3PL models will account for more variance, it is important to note that the focus of test design is not to explain maximum variance; it is to maximize accurate and objective measurement. Researchers remind us that Rasch is a parsimonious model and Rasch person parameter estimates consistently correlate with 2PL models above .90 (de Ayala, 2013), as we report below.

Finally, the Rasch model estimates measurement error more appropriately than traditional reliability statistics by providing the conditional standard error of measurement (CSEM) for each ability level. Thus, the reliability of producing a child ability at any one level differs from the reliability at any other level. Therefore, reporting CSEM allows the user to consider the error present in the student's ability at many levels, rather than as a traditional summary statistic.

Taken together, the information provided in the Rasch model contributes empirical evidence for the validity arguments, supporting appropriate interpretations and uses of early childhood language and literacy assessments for SEB students.

Claim 1: The IGDIs-E measures are psychometrically sound and theory-based, using Mark Wilson's measure design model and Rasch modeling for empirical item level statistics.

Item Development

Item Content

Our goal in each domain was to include culturally and linguistically relevant content that aligns with the construct maps previously presented. The following section describes how we selected content for each domain measured by the IGDIs-E.

Alphabet knowledge. In 2010 the Royal Spanish Academy voted to drop the "Ch" and "Ll" from the Spanish alphabet and now the Spanish alphabet officially includes 27 letters. (http://www.latintimes.com/spanish-royal-academy-eliminates-characters-y-ch-and-ll-alphabet-and-changes-names-others-164688).

We therefore included all letters that were officially part of the Spanish alphabet, but we also included the "Ll" as many bilingual programs and Spanish language teaching materials may still include this letter in Spanish early literacy instruction. We did not include the "ch" letter as initial interactions with programs illustrated that children were not engaging this letter in the classroom. With respect to letter sound content, most Spanish letters or phonemes have only one sound, given the language's shallow orthography. We excluded the letter sound for "w" as it is not native to Spanish and has the same pronunciation as the English "W". Similarly, we removed the letter "K" from our content pool because of its non-native inclusion in Spanish (Raynolds & Uhry, 2010). We also included the Spanish "ñ" sound based on its commonality.

Oral language. Our first step in developing the oral language items was to select the corpus of words in which we drew vocabulary from to design items. We cataloged vocabulary words in three ways. First, we reviewed the corpus of words on the MacArthur Inventarios del Desarrollo de Habilidades Comunicativas (Jackson-Maldonado, Thal, Marchman, Newton, Fenson, & Conboy, 2003). The Spanish version of the MacArthur was developed including 1,872 Spanish-English bilingual children between the ages of 30 months and 8 years in the US. The corpus of words on the test reflects high frequency words in the vocabulary of young Spanish-speaking children in the US. We therefore used this lexicon as our initial source of words. Second, we reviewed existing early childhood curricula in Spanish such as the Creative Curriculum (Dodge, Colker, Heroam, 2002), the Mexican Federal preschool curriculum (Secretaria de educación publica, 2010), and Estrellita (Myers, 2016). Finally, we cataloged all words in over 100 Spanish picture books by frequency and included high frequency words in our corpus of vocabulary. As a result of these three efforts we created a database of 1236 Spanish vocabulary words.

Phonological awareness. To develop items for phonological awareness IGDIs-E measures, we used the vocabulary database noted in oral language item development.

Item Design
With content to draw on, including target images, sound and letters, we turned our attention to the process of item design. Each domain featured unique approaches to designing, piloting and revising items.

Alphabet knowledge. The development of AK measures began with both expressive and receptive tasks on letter names and sounds. Initial attempts were made with expressive and receptive approaches for letter and sound tasks. The expressive letter naming approach was excluded because of ceiling and floor effects due to discrepancies between students who appeared to know many letters in Spanish and students who responded primarily in English. During the item design process, we constructed alphabet knowledge items by manipulating factors within image/letter presentation and within the phonetic and production similarities of the letters and sounds. For the letter naming tasks we manipulated item difficulty by considering the variables present in each distractor, as well as the number of distractors available (2 or 3). For example, to design item that included maximally different distractors we excluded distractors that had common orthographic features (e.g. letters may have "tails" that hang below the line, letters with closed loop structures, such as p, q, etc.). To design items we anticipated being more difficult that were minimally different from their distractors, we included distractors that had common orthographic features. For example, easier distractors for the target letter “t” might include the letters “o” and “m,” whereas more difficult distractors might include the letters “j” and “f.” Similarly, we explored letter difficulty in parallel research to note which letters demonstrate more difficult concepts. For example, vowels are more likely than consonants to have multiple sounds (although still fairly shallow in Spanish), as such, items with a target vowel and distractors that included two other vowels were considered.

Many of the features manipulated in letter sound items overlapped with our approach for writing phonological awareness first sound items. For letter sound tasks distractors were manipulated based on sound complexity and similarity. For example, easier distractors for the letter sound /m/ might include the letter sounds /k/ and /t/, because /m/ is a bilabial and nasal sound and these distractors are not. Furthermore, /k/ and /t/ are both plosives, whereas /m/ is not. A more difficult set of distractors would be /n/ and /b/.

Claim 1:
The IGDIs-E are psychometrically sound and theory-based, using Mark Wilson's constructing measures model and Rasch modeling for empirical item level statistics.

Claim 7:
IGDIs-E are inclusive of a variety of Spanish dialects and socio-economic backgrounds that are representative of Spanish speaking populations.

The letter sound /n/ shares the nasal quality of the /m/ sound, and is also not a plosive. Although the /b/ sound is plosive, it is similar to the /m/ sound in terms of pronunciation, both requiring bilabial articulation. We describe the process in additional detail within the Phonological Awareness section. Sound identification items were developed by sound complexity, bilabials, stops sounds, and diagraphs.

Oral language. To design OL items, we started by using our corpus of words to find appropriate images. Graphic design techniques were used to removing construct-irrelevant features, or any additional details in the image that could distract a child from attending to the main idea of the image (e.g., all backgrounds were made plain white when appropriate). Before our pilot study and before each year's field study, the process of selecting images took several iterations. Initial selections were based on the clarity of the image and expert content review of multiple image depictions of the same vocabulary word. During each year's study, we examined item-level statistics and identified items that may have functioned poorly due to ambiguous imaging. In these cases, we consulted the graphic designer to revise existing images or to find entirely new images. For example, in the *Verbos-Expresivo/Expressive Verbs* item "estar asustado" (to be scared), the image showed a boy with an open-mouthed, surprised look on his face. To improve item functioning, we had our designer add a spider dangling from the ceiling to the image for context. We made every attempt to include images that were actual photographs of words to reduce the inferences being made between the vocabulary word and the depiction.

Acceptable correct responses. Before our initial pilot study, our Spanish-speaking team tried to identify dialectal variations of words and all possible correct responses to describe each image. Beyond the initial item drafting process, however, selection of acceptable correct responses was driven by child responses and empirical criteria. During the pilot studies and during each year of field testing, we cataloged all child responses for each item to compute item-level response frequencies. These responses included those that were correct, incorrect, and every other response given, even if they were entirely unaligned with the image. We computed frequency counts for all responses and if a particular response was given by more than 5% of children, did not already appear on the back of the card, and was deemed appropriate by our Spanish-speaking team, we added it to the list of acceptable correct responses for that item and rescored the data as necessary. For example, for an image of a sailboat, we originally accepted "barco" (boat) as correct. Upon examining response frequencies, we added "velero" (sailing ship) and "bote" (boat) as other acceptable correct responses. It is important to note that for those responses that occurred more than 20% of the time (1 in 5 students), but were clearly incorrect, for example, saying "gato" for an image of a dog, we flagged the item as having a problematic image and revised it as previously noted. In this way correct answers during item design were driven by empirical data and expert content review.

For many items in *Verbos–Expresivo/Expressive Verbs*, it was common for children to give a more generic response (e.g., "hacer burbujas" [to make bubbles] for "soplar" [to blow bubbles]). In most cases, these generic responses technically described what was happening in the picture, so when they appeared in 5% or more of children's responses for a given item, then we determined whether the generic response could be accepted as correct. Sometimes the use of a more generic verb signaled less sophisticated vocabulary skills (e.g., "hacer comida" [to make food] instead of "cocinar" [to cook]), but in other cases, the generic verb was actually correct based on colloquial use or dialect (e.g., "[h]echar agua" [to throw or pour water] for "regar" [to water]). To evaluate responses we used a feature, function, class framework. If the verbs described a feature of the image, a function of the image, or a class of the image, we generally did not accept the answer as correct. For example, for the verb "manducir"/to drive we noted that a feature of the image such as "he's turning the wheels", a function of the image "he's going from place to place" or a class of image "he's doing transportation" were all incorrect.

For a few *Verbos – Expresivo/Expressive Verbs* items, additional acceptable correct responses were added because multiple actions were occurring in the image. For example, for an image of a hand passing a dollar bill to another hand, we originally considered "pagar/dar dinero" (to pay or to give money) as the correct response, but based on child response frequencies, we decided to add "recibir" (to receive) in case the child attended to the hand that was receiving the money.

Phonological awareness. To develop phonological awareness items we carefully considered the structure of Spanish words. For many words in Spanish it is difficult to isolate the initial phoneme for onset rhyme tasks (Escamilla, 2000). In this task we therefore included items that targeted the initial syllable (ga-to) and the initial phoneme (a-beja). Once item level targets (e.g., Spanish sounds) were identified, we isolated and removed construct irrelevant features, including background context in photographs and distractors with common features that were not part of the target skill (e.g., in a first sounds task, we excluded distractors that rhymed). We also carefully selected the images from the image bank to reflect culturally relevant and developmentally appropriate targets.

Claim 1:
The IGDIs-E are psychometrically sound and theory-based, using Mark Wilson's constructing measures model and Rasch modeling for empirical item level statistics.

Claim 7:
IGDIs-E are inclusive of a variety of Spanish dialects and socio-economic backgrounds that are representative of Spanish speaking populations.

When designing *Primeros Sonidos/First Sounds* items, we manipulated distractors in ways similar to the other domains. We considered the factors previously noted, maximally different distractor, (differing across multiple categories of inference) or minimally different (differing across one or few categories of inference). For example, maximally different *Primeros Sonidos/First Sounds* items (which theoretically represent an item that requires less ability) featured targets that contrasted distractors by syllables that were clearly different within each sound (e.g. target */que/*, foils: */pa/*, */fo/*), where minimally different *Primeros Sonidos/First Sounds* items featured targets that contrasted distractors that only manipulated one phoneme (e.g. target */man/*, foils */mun/*, */pan/*). This strategy was used for items with phonemes at the target level and manipulation varied as a function of the tongue position or articulation used to say the words. For example, in item where the target sound is */pa/* and the item features pato and two distractors, balon and mano, we have intentionally included three images where we controlled the number of syllables (all have two), the similarity in sound structure, all include the */o/* sound, and the similarity in initial phoneme articulation, where */pa/* and */ba/* and */ma/* are all bilabials.

Qualitative item testing
Once items were drafted we used small pilot groups to test feasibility and usability. This process allowed us to exclude measures that show limited promise in the measurement design framework. To complete the pilot testing, we recruited 10 students for each of the 16 IGDIs-E tasks, with a total sample of 33 students across sites. All students were four or five years old, spoke Spanish as a native or dominant language, attended a preschool or school readiness program, and would enter kindergarten the following academic year.

Data collection and fidelity standards. Data collectors included three fluent Spanish-speaking graduate students funded by the IGDIs-E project as Graduate Research Assistants (GRA). Prior to pilot data collection, each GRA was observed using a fidelity checklist and attained 100% fidelity on each of the IGDIs-E pilot tasks. All student interactions with data collectors were video-recorded for later coding. The pilot testing occurred for 12 of the 16 IGDIs-E pilot tasks between October and December 2012. The remaining four measures required further development and were piloted between April and August 2013.

Qualitative analysis. Three utility standards were created to limit the burden on student and practitioner use and interpretation. First, the task could not continue on to piloting if it included tedious or overwhelming materials or test kits. Many early childhood measures come with large supply kits, manipulatives, and multiple manuals. These measures often take a significant amount of time to deliver, are difficult to maintain if materials are missing or damaged, and can be confusing for practitioners with many separate parts and protocols. Second, tests must

demonstrate a cost-benefit balance such that producing the task would not outweigh the benefit of the scores achieved. As such, any task that was particularly expensive or time-consuming to produce was eliminated from the pool of potential measures for piloting. Finally, any task that provided redundant information already available in a psychometrically sound format in the field was eliminated.

Of 24 original tasks, 16 were produced for initial pilot testing, with 7 measures removed from the pool of tasks based on utility standards, limited variability in performance or poor construct alignment, including:

1. *Defined Language Interactions/Interacciones Definidas Lingüísticas,*
2. *Analogies/Analogías,*
3. *Story Comprehension- Recall and Prediction/Comprensión de la historia: el retiro y la predicción,*
4. *Detection/Detetcion,*
5. *Definitional Vocabulary (Receptive)/Vocabulario de Definiciones (Receptivo),*
6. *Letter Naming (Expressive)/Identificacion de las Letras (Expresivo)* and
7. *Definitional Vocabulary (Expressive)/Vocabulario de Definiciones (Expresivo))*

Functionality rubrics. A qualitative rubric evaluated each task's functioning in the field and addressed each measure's adherence to design and development standards. The qualitative rubric included the following criteria: (a) active engagement of child, (b) valid response patterns obtained from child, (c) ease of use by administrator, and (d) timeliness of measure administration and scoring. GRAs rated each of these four components using a 0-3 scale, where 0 represented an unsatisfactory and unresolvable measure that did not achieve its desired outcome, and where 3 represented a superior measure that achieved its desired outcome to the highest standard and was considered for further testing without reservations. Two independent coders reviewed each measure via video-recorded child interactions and used the qualitative rubric to arrive at an overall score for each measure. Overall scores ranged from 0-12.

Active engagement was rated based on coder-observed child attention to the task, the extent to which the child seemed to enjoy the task, and whether the child responded to the administrator when asked a question. Valid response patterns were determined by the degree to which meaningful data could be obtained from children. Data were considered meaningful by the coders when children responded thoughtfully and demonstrated understanding of what was being asked of them. A measure produced unreliable or invalid data when children guessed or consistently chose item distractors on the left, center, or right of the card. Ease of administrator use was determined by the success with which an administrator could give the task to a child. This rating was completed with end users in mind: would the procedures allow for successful administration of the task by someone with minimal academic training (i.e., paraprofessionals)? To evaluate the timeliness with which a measure

could be delivered and scored, video coders timed the child's interaction with each measure. If administration of a measure took more than 5 minutes or if scoring took more than 1 minute, a lower rating was given, as according to General Outcome Measures (GOM) standards, IGDIs-E measures must be quick and easy to administer (see pages 35-37 for more information about General Outcome Measure qualities).

Qualitative results. Results for the qualitative rubric are provided in Table 20. We removed five tasks from the candidacy pool for field testing based on low qualitative scores.

Table 20. *Qualitative Rubric Results*

Task	Qualitative Criteria				
	Active Engagement	Valid Response Patterns	Easy to Use	Timely to Deliver	Grand Total Score
Phonological Awareness:					
Rimar/Rhyming	3	2	2	3	10
Primeros Sonidos/First Sounds	3	2	2	3	10
Mexclar/Blending	3	2	3	3	11
Elision	3	2	2	2	9
Oral Language					
Identificación de los Dibujos/Picture Naming	3	3	3	2	11
~~*WODB*~~	~~2~~	~~1~~	~~2~~	~~2~~	~~7~~
~~*Categories*~~	~~2~~	~~1~~	~~2~~	~~2~~	~~7~~
Functions/Funciones	3	3	2	2	10
Receptive Verbs	3	3	2	3	11
Verbos – Expresivo/Expressive Verbs	3	2	2	2	9
Alphabet Knowledge					
~~*Letter Detection*~~	~~3~~	~~2~~	~~2~~	~~3~~	~~10~~
~~*Exp Letter Naming*~~	~~2~~	~~1~~	~~3~~	~~3~~	~~9~~
Identificación de las Letras/ Letter Naming	3	2	3	3	11
Identificación de l0s Sonidos /Sound Identification	3	2	2	3	10
Contextualized Oral Language					
Storybook	2	3	2	2	9
~~*Vamos a Hablar*~~	~~1~~	~~1~~	~~2~~	~~2~~	~~6~~

Note. Measures that were removed due to low quality results are crossed out.

Item Review and Revisions

Student trials. During the first year of our design process we tried out multiple measure formats and refined our pool to the existing measures presented in this manual. For the core 5 measures (*Identificación de los Sonidos /Sound Identification, Identificación de las Letras/ Letter Naming, Identificación de los Dibujos/Picture Naming, Verbos – Expresivo/Expressive Verbs, Primeros Sonidos/First Sounds*), we engaged students in trials to determine which items were functioning and which were problematic across two additional years of item trials.

National samples including four dialectal groups were recruited for these item trials including Mexican Spanish, Carribean Spanish, Central and South American Spanish, and mixed group membership Spanish. Using Rasch model statistics, we noted items with poor point-biserial correlations, poor in-fit or out-fit, and very low or very high p-values. These items were discarded or revised and recalibrated based on student response patterns and expert reviews. A summary of items that were discarded or revised, including the revision process, is provided in Appendix A.

Differential item functioning analysis. As previously noted, DIF is an important test to explore to ensure bias is not present within an assessment. We explored DIF for three variables: (a) sex, (b) Mexican/non-Mexican dialect, and (c) level of language exposure. First, we calibrated all items and compared calibration results for each group. Upon calibration, we set empirical and substantive criteria. These criteria guided our decisions to remove items that displayed statistically meaningful bias for clear reasons. Our empirical criteria were (a) C-level DIF contrast statistic according to Educational Testing Services (ETS) guidelines (Zwick, 2012), (b) statistically significant Rasch-Welch probability, and (c) adequate group sample size.

The DIF contrast variable represents the difference in item location (difficulty parameter estimates) between the two groups. When the contrast is negative, it favors, or requires less ability for, the reference group; when the contrast is positive, it favors, or requires less ability for, the focal group. ETS recommends that items with C-level DIF (Rasch contrasts greater than .64) should be further evaluated for potential bias (Linacre, 2016; Zwick, 2012). When the DIF contrast is large enough to represent an empirically meaningful difference in

Claim 1:
The IGDIs-E are psychometrically sound and theory-based, using Mark Wilson's constructing measures model and Rasch modeling for empirical item level statistics.

Claim 7:
IGDIs-E are inclusive of a variety of Spanish dialects and socio-economic backgrounds that are representative of Spanish speaking populations.

item difficulty between the two groups, the Rasch-Welch test shows statistical significance. In other words, a statistically significant Rasch-Welch probability (< .05) suggests we can reject the null hypothesis because the two estimates are empirically different. Finally, to complete item calibrations that are statistically robust, we require a sample of at least 100 members per group. As such, items with group sizes smaller than 100 were not considered for removal because the DIF results were too unstable to make a reliable decision, even if the DIF contrast and Rasch-Welch empirical criteria were met.

Items that met all three empirical criteria for at least one comparison variable (i.e., sex, dialect, or level of language exposure) moved to the substantive evaluation stage. Our substantive evaluation consisted of team review of the item's content. If there was an identifiable reason why the item favored one group over another, then we removed the item; if a reason could not be identified, then the item remained in our viable pool. In all, two scenarios led to an item being removed from our pool: (a) the item met all three empirical criteria and the substantive criteria for one comparison variable, or (b) the item did not meet the substantive criteria but met all three empirical criteria across two or more DIF comparison variables.

Sex. When considering the role of sex on item functioning, we wanted to ensure no bias could be attributed to being a boy or a girl when taking the IGDIs-E. As such, DIF was completed comparing boys and girls on each IGDIs-E measure. Table 21 lists all items that met one of the two scenarios described above (i.e., met empirical and substantive criteria for the sex variable, or met empirical criteria across sex and one other DIF comparison variable). The table provides information about item content, the difficulty for each group (represented in logits), the DIF contrast (also represented in logits), and the Rasch-Welch probability. For the sex contrast (boys v. girls), 13 items met our criteria for removal across the five measures. These items included 6 that favored boys and 3 that favored girls for Picture Naming; 2 that favored boys and 0 that favored girls for Expressive Verbs; 1 that favored boys and 0 that favored girls for Letter Naming; 1 that favored boys and 0 that favored girls for Sound Identification; and no items for First Sounds.

Claim 1:
The IGDIs-E are psychometrically sound and theory-based, using Mark Wilson's constructing measures model and Rasch modeling for empirical item level statistics.

Claim 7:
IGDIs-E are inclusive of a variety of Spanish dialects and socio-economic backgrounds that are representative of Spanish speaking populations.

Table 21. *Differential Item Functioning Results for Sex (Boys v. Girls)*

Item ID	Item Content	Difficulty for Girls	Difficulty for Boys	DIF Contrast	Rasch-Welch Prob.	Favors
Identificación de los Dibujos/Picture Naming						
110019	Tren/ferrocarril	0.38	-0.33	0.70	.002	Boys
110028	Maleta/equipaje/bulto	1.20	1.97	-0.77	<.001	Girls
110044	Pelota/bola/bolita/balón	0.40	-0.36	0.75	.001	Boys
110048	Martillo	1.92	1.07	0.85	<.001	Boys
110049	Tigre	2.13	1.38	0.75	<.001	Boys
110054	Vestido/traje	0.68	2.37	-1.70	<.001	Girls
110057	Bate de beisbol/bate	2.96	1.47	1.49	<.001	Boys
110076	Camión/troca	1.97	1.33	0.65	.004	Boys
110079	Mariposa	-0.31	0.42	-0.73	.001	Girls
Verbos (Expresivo)/Expressive Verbs						
140010	Jugar fútbol/patear	-0.68	-1.39	0.70	.001	Boys
140048	Deslizar/resbalar	1.55	0.86	0.69	.003	Boys
Identificación de las Letras/Letter Naming						
150059	s/z/O (z)	1.64	0.90	0.74	.017	Boys
Identificación de los Sonidos/Sound Identification						
160033	Ll/w/l (l)	1.90	1.22	0.68	.021	Boys
Primeros Sonidos/First Sounds						
None						

Note: Girls formed the reference group and boys formed the focal group.

Dialectal groups. We repeated DIF analyses to explore potential item bias based on dialectal representation. Due to the demographic constraints of our sample, we were only able to separate children who spoke a Mexican dialect of Spanish from children who spoke any other dialect (e.g., Caribbean, dialects from Central and South American countries). Our dialectal groups from other regions did not include over 100 children in each group, and therefore we could not produce 100 valid responses per item. Thus, the two groups for DIF analysis were Mexican dialect and Non-Mexican dialect. Across the five measures, 10 items met our criteria for removal (i.e., met empirical and substantive criteria for the dialect variable, or met empirical criteria for dialect and one other DIF comparison variable). These items included 2 that favored Mexican dialects and 4 that favored non-Mexican dialects for Picture Naming; 1 that favored Mexican dialects and 0 that favored non-Mexican dialects for Expressive Verbs; 2 that favored Mexican dialects and 0 that favored non-Mexican dialects for Letter Naming; 1 that favored Mexican dialects and 0 that favored non-Mexican dialects for Sound Identification; and no items for First Sounds. These items are listed in Table 22.

Table 22. *Differential Item Functioning Results for Dialect (Mexican v. Non-Mexican)*

Item ID	Item Content	Difficulty for Non-Mex	Difficulty for Mex	DIF Contrast	Rasch-Welch Prob.	Favors
Identificación de los Dibujos/Picture Naming						
110036	Arroz	-0.19	1.63	-1.82	<.001	Non-Mex
110028	Maleta/equipaje/bulto	0.82	1.96	-1.13	<.001	Non-Mex
110044	Pelota/bola/bolita/balón	0.99	-0.65	1.64	<.001	Mex
110049	Tigre	1.33	2.11	-0.78	.002	Non-Mex
110057	Bate de beisbol/bate	1.53	2.40	-0.87	.015	Non-Mex
110084	Chile	1.27	-1.10	2.37	<.001	Mex
Verbos (Expresivo)/Expressive Verbs						
140048	Deslizar/resbalar	2.10	0.77	1.37	<.001	Mex
Identificación de las Letras/Letter Naming						
150013	j/h/a (j)	0.93	0.05	0.89	.004	Mex
150059	s/z/O (z)	0.56	-0.12	0.67	.005	Mex
Identificación de los Sonidos/Sound Identification						
160033	Ll/w/l (l)	2.12	1.35	0.77	.032	Mex
Primeros Sonidos/First Sounds						
None						

Note. Non-Mexican dialects formed the reference group; Mexican dialects formed the focal group.

Language exposure. Finally, we explored the role of language exposure on item functioning. To form groups, we scored language exposure items from the family survey completed by parents/guardians, with a maximum exposure score of 16 for each language (Spanish, Both, English). Scores closer to 16 indicated higher levels of exposure for the given category. We created categorization using the 75th percentile such that group membership required scores indicative of the greater than or equal to the 75th percentile, which in raw score equivalents was a score greater than 11 out of 16. Children who scored 12-16 for Spanish exposure comprised the Spanish-dominant group, and children who scored 12-16 for exposure to both Spanish and English (by parents noting "Both") comprised the Both group. The sample of children who scored 12-16 for English exposure was too small for DIF analysis; further, English-dominant bilinguals are not the target population for the IGDIs-E. Table 23 displays the language exposure results. When contrasting Spanish-dominant with Both, 3 items met our criteria for removal (i.e., met empirical and substantive criteria for the level of language exposure variable, or met empirical criteria across language exposure and one other DIF comparison variable). These items included 1 that favored Spanish-dominant and 1 that favored Both for Picture Naming; 1 that

favored Spanish-dominant and 0 that favored Both for Letter Naming; and no items for Expressive Verbs, Sound Identification, or First Sounds.

Table 23. *Differential Item Functioning Results for Language Exposure (Spanish v. Both)*

Item ID	Item Content	Difficulty for Both	Difficulty for Spanish	DIF Contrast	Rasch-Welch Prob.	Favors
Identificación de los Dibujos/Picture Naming						
110036	Arroz	-0.25	1.34	-1.60	<.001	Both
110079	Mariposa	0.21	-0.43	0.64	.030	Spanish
Verbos (Expresivo)/Expressive Verbs						
None						
Identificación de las Letras/Letter Naming						
150013	j/h/a (j)	1.29	0.37	0.92	.010	Spanish
Identificación de los Sonidos/Sound Identification						
None						
Primeros Sonidos/First Sounds						
None						

Note. Children exposed to Both formed the reference group, and children who were Spanish-dominant formed the focal group.

Expert review. To further explore item development, a group of five national experts from bilingual assessment, Spanish language development, and early childhood research reviewed the expanded item sets of selected tasks. All experts had native or near-native proficiency in Spanish as well as knowledge of dialect differences and linguistic differences for which they were evaluating content. They provided a detailed review and qualitative analysis of the oral language and phonological awareness tasks in a one-on-one interview. Overall, responses were positive. Reviewers provided recommendations about the usage of alternative dialectal responses to some of our items. For example, a child of Puerto Rican descent may be more likely to respond with *rueda* instead of *llanta* to a picture of a tire, or with *china* instead of *naranja* to a picture of an orange. In addition to feedback on dialectal considerations for individual items, reviewers also provided feedback on the use of some measures. For example, some reviewers suggested that rhyming in Spanish was not as salient as it is in English and thus may not contribute as much to Spanish early literacy development. Furthermore, in relation to the measure first sounds, reviewers suggested that item targets should include whole syllables rather than just the initial phoneme because of the syllabic nature of Spanish.

Outcome Space

Scoring Responses

As previously noted, the outcome space identifies how we interpret the child response as a score. The IGDIs-E measures employ a dichotomous item scoring approach. In alignment with each construct map, responses were scored as correct (1) or incorrect (0). During the development period partial credit models were tested to examine the degree to which a meaningful hierarchy of knowledge was present in each task, however, partial credit models did not improve the measurement model (test information function) and were therefore not used.

The IGDIs-E model employed dichotomous scoring for three reasons. First, the dichotomous model provided empirically robust scores based on item fit indices. The receptive tasks that include dichotomous scoring in a multiple choice format provided response choices that demonstrated no clear hierarchy. That is, of the distractors, we did not systematically control for or attempt to create distractors with difficulty that were relatively easier or more difficult than the target or other distractors and as such could not assign partial credit to receptive items as there was no way to identify credit quantities for each response choice outside of the target. For expressive tasks, including *Identificación de los Dibujos/Picture Naming* and *Verbos – Expresivo/Expressive Verbs* , we did consider partial credit scoring, but found that a hierarchy of scores added little to no value to the measurement model. As such, partial credit models were not used in these tasks.

Second, to maximize ease of administration and standardize procedures we determined the measures should be dichotomously scored (correct or incorrect) for each measurement approach (expressive or receptive) because of the dramatic variability present in the level of expertise of administrators. Early childhood practitioners administer assessments with a range of educational backgrounds.

Claim 1:
The IGDIs-E are psychometrically sound and theory-based, using Mark Wilson's constructing measures model and Rasch modeling for empirical item level statistics.

Claim 4:
IGDIs-E are developmentally appropriate for SEB 4-5-year-old children.

Classrooms with Spanish speaking students may have support for those students in their L1 only at a paraprofessional or interpreter level, which may have had no experience with assessment training. Therefore, to ensure standardization we wanted to make the scoring as simple and straightforward as possible.

Third, early childhood research frequently presents items in a dichotomous approach given preschool performance is highly variable (Schweinhart, DeBruin-Parecki, & Robin, 2004). At the preschool level, distractibility and variable responses are common. Preschool children frequently offer responses that align with the construct being measured and are easy to score, but in other circumstances preschoolers wander off-task to engage in conversation that is not construct related. Although the universe of scores that may potentially be correct are limited by dichotomous scoring, it also offers the benefit of seeking particular responses from children rather than evaluating all possible responses children may provide to discern if they offered a response that meets the criteria for various scores.

Claim 4:
IGDIs-E are developmentally appropriate for SEB 4-5-year-old children.

After DIF items were removed, a homogeneous set of items for each measure were produced and therefore responses were defended as unbiased or valid representations of the construct of interest. However, even with the benefits of dichotomous scoring in an expressive and multiple-choice response model there are some challenges that must be reviewed.

Claim 1:
The IGDIs-E are theory-based.

First, multiple choice items introduce the concept of guessing as there is the potential that a child could guess the correct response rather than relying on their knowledge to construct an answer. For all measures we examined item level statistics to evaluate the contribution of guessing on item level responses. All receptive measures include two, three or four choices within each item for response selection. If chance is playing a major role in calibration it will produce poor fit indices and poorly characterize performance on the construct. To examine how guessing contributed to item level information we examined in-fit and out-fit statistics. Item-total correlations are attenuated when random responses (guessing) have a significant presence; thus item discrimination will be low. If guessing is present, in-fit and out-fit will be inappropriately high.

The finalized item sets for all measures showed no impact of guessing. In addition, researchers have documented longstanding controversies regarding random responses (guessing) in a 3PL model and the value of its application (Chiu & Camilli, 2013). We posit the 3rd parameter is not a necessary inclusion in the model given the item review process we specified. In sum, these findings provide empirical evidence regarding our claim that the IGDIs-E are meaningful representations of the Spanish early language and literacy constructs.

Second, dichotomous scoring in expressive and multiple choice formats can potentially prevent administrators from scoring a response as correct if it was not identified in the expressive response key. For example, if a child provides a response outside of the listed correct responses such as "davenport" for "sofa" that is indeed correct for an expressive item, but not included in the key, the response will be scored as incorrect and the child will not receive credit for an item with content that they know. Similarly, if the child provides a response on a receptive task that is not provided within the choices but is correct, he or she will again not receive credit, when the underlying skill maybe be mastered (for example, if an item offers the target cat and response choices of bat, dog and chair, but the child says "cat, rat, the rhyme" and chooses not to select or name a response on the item, he or she would get the item incorrect).

Third, dichotomous scoring may increase standardization, but is not entirely protective of administration errors. During validation studies for IGDIs-E a review of data collector identifiers and frequency of discontinuation was analyzed using a chi square analysis. Results indicated that some assessors inappropriately discontinued children on two IGDIs-E measures significantly more than is predicted using a chi square likelihood ratio. A detailed analysis illustrated that three data collectors over-discontinued students. These results suggest differences in sample card presentation were present among data collectors and even with standardization in procedures, the assessment protocols were not uniform across data collectors. We present this information to note that although dichotomous scoring can be advantageous, it will not prevent all errors in scoring as a result of administration inconsistencies.

Claim 1:
The IGDIs-E are psychometrically sound and theory-based, using Mark Wilson's constructing measures model and Rasch modeling for empirical item level statistics.

Inter-Rater Reliability and Fidelity of Implementation

Scoring rules for administration rely on the training of the assessor in accurate delivery of each measure. As such, each measure requires the assessor achieve 90% accuracy before beginning testing with a student. To achieve 90% accuracy the assessor must review the measure content and all administration materials and then engage a trained assessor with fidelity in a testing session. During the session the trainer uses the fidelity of implementation checklist for the relevant measure to assess fidelity. At the same time the assessor must score the items he or she delivers. At the same time the trainer scores the same items and items are checked for inter-rater reliability. Assessors receive immediate feedback from the trainer if they do not achieve 90% the first time. Assessor are allowed three back to back trials to achieve 90% fidelity and inter-rater reliability (via Kappa). If assessors do not achieve 90% on fidelity of implementation and on inter-rater reliability they must practice using the measure with three adults or children and then attempt another round of fidelity of implementation and inter-rater reliability with the trainer.

Claim 1:
The IGDIs-E are psychometrically sound and theory-based, using Mark Wilson's constructing measures model and Rasch modeling for empirical item level statistics.

The Measurement Model (Rasch)

Rasch Argument and Fit

The Rasch (1980) model provides a guiding framework for measure development to directly align with the construct of interest. In doing so, we can evaluate items and item features (characteristics of items such as number of elements presented and complexity of the task) vis-à-vis the item difficulty, which is the item's location on the continuum of the underlying trait present in a given construct. This allows us to defend the construct, as the items theorized to be easier actually turn out to be easier and those items that were theorized to be more difficult are actually more difficult empirically. This mapping to the construct provides important content-related validity evidence, supporting the intended interpretation of the score scale and the underlying ability regarding Spanish early language and literacy. The Rasch model is explained in further detail on pages 23-25.

Claim 13:
IGDIs-E were uniquely designed to attend to how Spanish language develops rather than by translating existing English measures.

In the IGDIs-E model we used Rasch to calibrate all items and produce item level statistics that were evaluated to determine which items were functioning appropriately in the model (as well as which were not). Two assumptions are important to evaluate empirically for the Rasch model. Because of the complexity of the data collected at this point and because of the relatively modest sample sizes, multiple model analyses were conducted as an evaluation of the assumptions.

First we tested the empirical model fit of the model; as with all IRT models, we assume the mathematical model fits. Second, we conducted a series of confirmatory factor analyses. These analyses were conducted with forms from all five domains.

To satisfy the assumption of mathematical model fit, items and persons were calibrated with the Rasch (discrimination set to equal 1 for all items, item location/difficulty allowed to vary across items), 1PL (1-parameter logistic model, estimating the average discrimination and fixing it across items; allowing item location to vary), 2PL (similar to 1PL, allowing item discrimination to vary), and 3PL (similar to 2PL, allowing lower asymptote to vary) models using Bilog Version 3 (du Toit, 2003; Zimowski, Muraki, Mislevy & Bock, 2003). To pragmatically simplify the comparison of these multiple models employing the same data, we examined correlations among person locations (person ability scores), as obtaining person scores is the core reason for employing the measurement model, placing items and persons on the same scale as a representation of the construct. A correlation table is provided for one form from each domain, to illustrate the consistency of model agreement (Table 24). In part, these different IRT scoring models result is essentially the same ordering of persons (near perfect correlations) because the skills are relatively focused and consistently defined.

All unidimensional IRT models, not surprisingly, assume the underlying trait is a unidimensional construct; a single latent trait is being measured by the items. The extent to which a unidimensional model fit each form was evaluated through confirmatory factor analyses (CFA). CFA was conducted with Mplus Version 7 (Muthén & Muthén, 2012a). Three measures of model fit provide different aspects of fit, including the root mean-squared error of approximation (RMSEA), the extent to which the model fits reasonably well in the population; comparative fit index (CFI), the relative fit to a more restricted baseline model; and the Tucker-Lewis index (TLI), which compensates for the effect of model complexity. Multiple indicators of fit should be examined (Muthén & Muthén, 2012b).

The general criteria for model-data fit are as follows (Brown, 2015):

RMSEA < .05 is good fit, RMSEA < .08 is adequate fit;
CFI > .95 is good fit, CFI > .90 is adequate fit;
TLI > .95 is good fit, TLI > .90 is adequate fit.

Based on these criteria, each operational form of measure resulted in good to adequate fit to a unidimensional model (mostly in the good range), meeting the unidimensionality assumption of the Rasch model. These results are presented in Tables 25 to 29.

Table 24. *Correlations among IRT Model Person Scores by Measure*

	Rasch	1PL	2PL	3PL
First sounds				
Rasch	.89			
1PL	.989	.90		
2PL	.988	.998	.90	
3PL	.989	.992	.996	.88
Picture naming				
Rasch	.91			
1PL	.984	.91		
2PL	.984	.991	.91	
3PL	.985	.984	.996	.90
Expressive verbs				
Rasch	.90			
1PL	.983	.91		
2PL	.987	.990	.91	
3PL	.988	.985	.998	.90
Letter naming				
Rasch	.89			
1PL	.988	.91		
2PL	.985	.994	.91	
3PL	.985	.983	.991	.89
Sound identification				
Rasch	.89			
1PL	.989	.90		
2PL	.979	.983	.90	
3PL	.976	.968	.989	.88

Note. IRT model reliabilities are on the diagonal. Correlations are based on one form in each domain, each containing 25 items, with sample sizes: FS (367), PN (369), EV (364), LN (338), SI (365).

Table 25. *Confirmatory Factor Analysis Results for* ***Primeros Sonidos***

Form	*# of items*	*# of children*	*RMSEA*	*CFI*	*TLI*
100214001	25	333	.029	.954	.950
100214002	25	289	.035	.927	.920
100214003	25	304	.034	.973	.971
100214004	25	275	.036	.950	.945
100315001	15	1051	.042	.963	.957
100717001	25	240	.029	.988	.987
100717002	25	238	.035	.983	.981

Table 26. *Confirmatory Factor Analysis Results for* ***Identificación de los Dibujos***

Form	*# of items*	*# of children*	*RMSEA*	*CFI*	*TLI*
110214001	25	321	.026	.990	.989
110214002	25	290	.034	.980	.978
110214003	25	305	.018	.995	.994
110214004	25	271	.046	.967	.964
110315001	15	1171	.051	.961	.955

Table 27. *Confirmatory Factor Analysis Results for* ***Verbos Expresivos***

Form	*# of items*	*# of children*	*RMSEA*	*CFI*	*TLI*
140214001	25	317	.021	.990	.989
140214002	25	288	.028	.963	.959
140214003	25	302	.021	.984	.983
140214004	25	276	.032	.948	.942
140315001	15	1126	.038	.991	.989
140717001	25	162	.024	.991	.990
140717002	25	115	.012	.999	.999
140717003	25	156	.023	.992	.992

Table 28. *Confirmatory Factor Analysis Results for* ***Identificación de las Letras***

Form	*# of items*	*# of children*	*RMSEA*	*CFI*	*TLI*
150214001	25	338	.035	.941	.936
150214002	25	292	.039	.928	.921
150214003	25	310	.026	.981	.979
150214004	25	277	.031	.964	.960
150315001	15	1120	.059	.926	.914
150717001	25	160	.036	.982	.980
150717002	25	165	.038	.980	.978

Table 29. *Confirmatory Factor Analysis Results for* ***Identificación Sonidos***

Form	*# of items*	*# of children*	*RMSEA*	*CFI*	*TLI*
160214001	25	328	.035	.970	.967
160214002	25	293	.037	.975	.972
160214003	25	301	.029	.970	.967
160214004	25	280	.043	.951	.946
160315001	15	1073	.059	.953	.946

Note. RMSEA is the root mean-squared error of approximation. CFI is the comparative fit index. TLI is the Tucker-Lewis index.

Concurrent Calibration
Data were collected during two academic years, 2013-2014 (calibration study) and 2014-2015 (pilot study). At total of 970 children completed the IGDIs-E measures across the states of CA, FL, IL, KS, MN and UT. Sampling occurred using two different designs across the years. Each design is described here.

Calibration study sampling design (2013-2014). Each measure consisted of 75 items and items that were divided across four different forms. Each form consisted of 25 items assembled in blocks and adjacent forms had 8 or 9 items in common (see Table 30). All forms were used across the three seasons. The study was designed so that each child would see all measures per season and for each measure a child would see a different form each season. This approach allowed for strategic item testing without fatiguing preschool age children.

Table 30. *Item Sampling Design for the Calibration Study 2013-2014*

Measures	Form 101	Form 102	Form 103	Form 104
Identificación de los Dibujos/Picture Naming	Block A (1-8) Block B (9-16) Block C (17-25)	Block D (26-33) Block E (34-41) Block C (17-25)	Block F (42-50) Block E (34-41) Block G (51-58)	Block H (59-66) Block I (67-75) Block G (51-58)
Verbos (Expresivo)/ Expressive Verbs	Block A Block B Block C	Block D Block E Block C	Block F Block E Block G	Block H Block I Block G
Primeros Sonidos/First Sounds	Block A Block B Block C	Block D Block E Block C	Block F Block E Block G	Block H Block I Block G
Identificación de las Letras/ Letter Naming	Block A Block B Block C	Block D Block E Block C	Block F Block E Block G	Block H Block I Block G
Identificación de los Sonidos /Sound Identification	Block A Block B Block C	Block D Block E Block C	Block F Block E Block G	Block H Block I Block G

Note. (#-#) indicates item positions.

Pilot study sampling design (2014-2015). The pilot study consisted of testing new items. The number of new items by measure varied and were as follows: *Identificación de los Dibujos/Picture Naming* = 42, *Verbos (Expresivo)/Expressive Verbs* = 36, *Primeros Sonidos/First Sounds* = 29, *Identificación de las Letras/ Letter Naming* = 35 and *Identificación de los Sonidos /Sound Identification* = 43. New items were grouped into blocks (Block A, B, C, etc.) of 7 or 8 items. Each block was only administered in a given season. For instance, Block A was only administered in the fall. Each child only saw a given block once. Pilot items were administered along with 20 items from the calibration study. For each measure, the administration

consisted 15 screening items (similar in difficulty around a cut score), 5 anchor items (spread across the scale) and 7-8 pilot items (of unknown difficulty) to always appear in this order. When not all 27/28 items could be administered in one sitting, pilot items were administered at a later time (typically the next day). For each measure, screening and anchor item sets remained constant for all three seasons and for all children (see Table 31). In addition, children who saw the screening items but did not see the anchor or pilot items were also included in the analyses.

Table 31. *Item Sample Design for Pilot Study*

Form	Season	Measure Set	Calibration items	New piloted items
Form A	Fall	Screening (15)	Anchor (5)	Pilot –Block A (7)
Form B	Fall	Screening (15)	Anchor (5)	Pilot –Block B (7)
Form C	Fall	Screening (15)	Anchor (5)	Pilot –Block C (7)
Form D	Winter	Screening (15)	Anchor (5)	Pilot –Block D (7)
Form E	Winter	Screening (15)	Anchor (5)	Pilot –Block E (8)
Form F	Winter	Screening (15)	Anchor (5)	Pilot –Block F (8)
Form G	Spring	Screening (15)	Anchor (5)	Pilot –Block G (8)*

Note. (#) indicates number of items. *This block was created from items in Blocks A-F with fewer responses. Not all forms were used for all measures.

Children for whom the measure was discontinued (i.e. did not pass the sample cards) were excluded from the analyses. In the case of the pilot study, children with responses for pilot items but no responses for the anchor or screening items were excluded. Similarly, children who did not respond to at least 40% of the items were not included in the analysis because we hypothesized their responses patterns were significantly impacted by error variance. For the remaining cases, items with missing responses were considered as not-administered for the item calibration. Both, the calibration study and pilot study were originally designed so a child would be tested once every season. Rasch item calibrations were estimated using Winsteps 3.72 (Linacre, 2011).

Initial calibrations in the studies described here were used to examine new items and create an initial item bank. However, the most robust calibration of items after all revisions previously noted must occur as a concurrent calibration. In a concurrent calibration the items are allowed to independently vary and scale with no items serving as anchors to an existing scale. As such, after two years of initial item bank construction we moved to a finalized concurrent calibration that produced the finalized IGDIs-E item statistics presented in Tables 33-37.

Participant sample. Concurrent calibration included students who responded to items in Years 2 and 3 (2013-2015; *n*=970). We collected demographic information for the entire sample, however some families did not return the demographic form, or did not respond to some of the questions on the survey. As such, the sample that contributed to each demographic differs. Table 32 provides the percentage of students represented in each demographic group.

Table 32. *Student Level Demographics for IGDIs-E Concurrent Calibration Sample*

Characteristic	*Percentage (%)*
Female	50.1
Special education services	6.2
Ethnicity/Race	
Latino (general)	54.4
Mexican	16.6
Puerto Rican	12.3
Caribbean	2.1
Central American	4.1
South American	1.0
Multiple races/ethnicities	8.8
Other	0.7
Regional Representation	
Midwest (MN, IL, KS)	30.1
FL	25.4
CA	26.3
UT	18.2
Languages spoken to the child from ages 0 to 1	
Spanish	71.1
Both Spanish and English	24.8
Language the child uses when talking at home	
Spanish only	48.8
Both	32.5
English	10.3
Other	8.4
Household weekly income	
Less than $500	65.0
$501 – 700	24.8
$701 – 900	4.2
More than $901	6.1

Table 32. *Student Level Demographics for IGDIs-E Concurrent Calibration (cont.)*

Characteristic	*Percentage (%)*
Mothers highest level of education	
6th grade or less	16.4
Less than 12th grade	20.1
GED	9.9
High school diploma	16.1
Some education after high school / vocational program	17.3
Associate degree (AA)	5.3
College degree (BA/BS)	9.5
Graduate/ professional degree	5.3

Resulting item level calibrations. The total sample include 975 students across the two years. Students received item sets three times a year (fall winter and spring) and scores were treated as independent cases across season. As such, some items have a *Count* as large as approximately three times the sample size. The calibration results are provided in Tables 33-37 for all 5 measures. In each table we provide the item identifier, the Winsteps output statistics, noted as Measure (item location or difficulty), Count (the number of students who responded to the item), Score (number of students who got the item correct), Standard Error of Measurement (SEM), Mean-Square Infit, Mean-Square Outfit and point-biserial correlation. Descriptions of each variable's criteria for evaluation are provided in the item analysis section on page 48-51.

Item analysis. For each measure we analyzed item level statistics to determine if the item was fitting appropriately. We used the Point-Biserial correlation to examine the degree to which performance on a given item correlates with total score. We used in-fit and out-fit mean square statistics to determine how the item fits the Rasch model, discarding items with fit below an absolute value of 0.5 and above an absolute value of 1.5. We used the *p*-value, or ratio of SCORE to COUNT variables (proportion correct) to evaluate how useful the item was in the item pool. We eliminated items that were very difficult or very easy (above a *p*-value of .8, or below a *p*-value of .2). Items were removed sequentially and recalibrated to examine continuous fit statistics. All items discarded are noted in Appendix B.

Table 33. *Identificación de las Letras/ Letter Naming Item Calibrations*

Item	Measure	Count	Score	SEM	In.Msq	Out.Msq	Pbs
150005	-0.43	954	635	0.08	1.11	1.07	0.28
150006	-0.36	316	221	0.14	0.89	0.78	0.47
150007	-1.63	317	276	0.18	0.84	0.72	0.41
150008	0.49	1418	718	0.06	1.03	1.00	0.28
150009	0.32	316	183	0.13	1.13	1.27	0.31
150010	0.27	1421	780	0.06	0.92	0.84	0.37
150011	-1.19	318	261	0.16	0.90	0.73	0.39
150012	0.19	1423	799	0.06	0.91	0.85	0.38
150013	0.55	313	167	0.13	1.20	1.40	0.26
150080	-0.88	318	248	0.15	0.97	0.88	0.35
150015	0.04	1427	844	0.06	0.93	0.85	0.36
150016	-0.05	1431	868	0.06	0.95	0.89	0.32
150017	-0.47	955	641	0.08	0.88	0.81	0.46
150018	0.34	318	181	0.13	0.92	0.89	0.48
150019	0.75	937	408	0.08	1.13	1.16	0.30
150020	-0.88	321	250	0.15	1.00	0.92	0.34
150021	0.36	597	351	0.10	1.08	1.09	0.35
150081	0.64	1708	842	0.06	0.95	0.91	0.38
150023	-2.16	604	556	0.16	0.90	0.74	0.32
150082	-0.94	602	480	0.11	1.08	1.17	0.25
150083	-1.01	600	485	0.11	0.98	0.88	0.35
150026	-0.64	600	454	0.11	0.93	0.82	0.43
150027	-0.11	602	405	0.10	0.89	0.78	0.48
150029	-0.21	587	404	0.10	1.10	1.12	0.29
150030	-2.16	285	264	0.24	0.92	1.25	0.31
150032	0.83	285	151	0.14	1.03	1.03	0.41
150033	0.54	282	164	0.14	1.31	1.43	0.21
150084	-0.15	920	572	0.08	1.16	1.14	0.25
150035	-1.43	286	248	0.19	0.91	0.66	0.36
150036	1.18	285	133	0.14	1.33	1.38	0.25
150037	-0.47	282	211	0.15	1.06	0.93	0.32
150038	0.31	585	358	0.10	0.93	0.85	0.49
150039	-0.87	588	468	0.11	1.05	1.11	0.31
150040	0.19	587	371	0.10	0.90	0.83	0.50
150041	-1.87	590	532	0.15	0.89	0.82	0.36
150042	1.01	584	285	0.10	1.00	1.00	0.45
150043	1.24	583	263	0.10	1.18	1.19	0.35
150044	-0.05	585	395	0.10	0.85	0.76	0.53

Item	Measure	Count	Score	SEM	In.Msq	Out.Msq	Pbs
150045	-0.42	587	431	0.11	0.86	0.72	0.50
150046	-0.22	298	207	0.14	0.96	1.03	0.43
150047	0.61	1394	681	0.06	1.04	1.02	0.30
150048	-0.57	300	224	0.15	1.05	1.14	0.35
150049	-0.05	304	203	0.14	1.09	1.17	0.36
150050	-0.31	299	212	0.15	0.93	0.90	0.46
150051	-0.61	300	226	0.15	1.07	0.93	0.34
150052	-0.48	300	220	0.15	1.02	0.91	0.40
150053	-0.17	299	205	0.14	0.90	0.79	0.51
150054	0.38	302	179	0.14	0.92	0.82	0.50
150055	1.52	552	217	0.11	1.23	1.28	0.31
150056	0.41	561	330	0.10	0.90	0.83	0.52
150057	-0.93	562	449	0.12	0.98	0.86	0.37
150058	1.91	1669	474	0.07	1.17	1.18	0.42
150059	0.05	560	365	0.10	1.23	1.51	0.25
150060	0.51	1659	857	0.06	1.13	1.15	0.26
150061	0.82	1668	770	0.06	0.96	0.91	0.39
150062	-1.24	564	473	0.13	0.90	0.67	0.42
150063	-0.40	895	593	0.08	0.97	0.93	0.38
150064	-0.05	260	174	0.15	0.93	0.90	0.44
150065	0.07	259	168	0.15	1.01	0.94	0.40
150066	-0.40	262	191	0.16	1.06	1.09	0.31
150067	-0.47	264	195	0.16	0.85	0.79	0.49
150068	0.69	257	138	0.15	1.28	1.40	0.20
150069	-0.44	262	192	0.16	0.89	0.80	0.46
150070	-0.17	1368	864	0.06	0.99	0.99	0.30
150071	0.38	1370	728	0.06	1.04	1.05	0.31
150072	0.27	1367	756	0.06	0.90	0.86	0.39
150073	-0.43	262	192	0.16	0.97	0.89	0.39
150074	0.67	257	139	0.15	1.16	1.18	0.30
150075	1.30	1352	494	0.07	1.04	1.05	0.36
150076	1.13	263	121	0.15	1.15	1.18	0.32
150077	0.34	1364	737	0.06	0.95	0.93	0.35
150078	0.42	258	151	0.15	1.14	1.21	0.31
150079	-0.68	261	201	0.16	0.90	0.82	0.44
150086	0.52	99	38	0.23	1.07	1.10	0.23
150088	-0.67	101	64	0.23	0.97	0.90	0.32
150090	-0.57	101	62	0.22	0.90	0.85	0.39
150091	0.49	98	38	0.23	1.01	1.05	0.26

Item	Measure	Count	Score	SEM	In.Msq	Out.Msq	Pbs
150093	0.31	112	51	0.21	1.05	1.07	0.26
150094	0.27	112	52	0.21	0.94	0.89	0.37
150095	0.38	111	49	0.21	0.90	0.87	0.41
150096	-0.18	111	62	0.21	0.88	0.84	0.43
150098	-0.02	112	59	0.21	0.93	0.89	0.37
150099	-0.26	111	64	0.21	1.01	0.99	0.27
150100	0.65	157	66	0.18	1.16	1.11	0.26
150101	-1.55	160	131	0.22	1.01	0.96	0.27
150103	0.55	159	70	0.18	1.01	0.99	0.39
150104	0.32	159	77	0.18	0.97	0.93	0.41
150106	0.39	157	74	0.18	1.06	1.05	0.31
150107	-1.33	112	90	0.26	1.05	0.83	0.27
150108	0.07	109	61	0.22	0.72	0.60	0.65
150109	1.04	109	42	0.23	1.08	1.08	0.41
150110	0.69	108	48	0.23	1.16	1.15	0.32
150111	0.68	108	48	0.23	1.15	1.16	0.33
150112	-0.13	111	67	0.22	0.91	0.84	0.48
150113	0.37	108	55	0.22	0.86	0.73	0.56
150114	0.66	155	77	0.19	0.92	0.87	0.49
150115	0.05	155	95	0.19	0.90	0.87	0.48
150116	0.52	155	81	0.18	0.92	0.92	0.48
150117	0.45	155	83	0.18	0.85	0.78	0.54
150118	0.59	154	78	0.19	1.13	1.17	0.31
150119	0.73	155	75	0.19	1.01	0.99	0.42
150120	0.22	155	90	0.18	0.88	0.98	0.50

Table 34. *Identificación de los Sonidos /Sound Identification Item Calibrations*

Item	Measure	Count	Score	SEM	In.Msq	Out.Msq	Pbs
160005	-0.69	314	240	0.15	0.91	0.95	0.45
160006	0.39	1021	514	0.07	0.94	0.89	0.48
160007	-0.59	309	231	0.15	1.10	1.08	0.32
160008	0.35	307	183	0.14	1.32	1.54	0.23
160009	-0.20	311	215	0.14	0.93	0.85	0.49
160020	0.04	311	203	0.14	1.26	1.46	0.28
160080	0.67	308	168	0.14	0.87	0.85	0.57
160023	-0.09	1016	598	0.07	1.09	1.09	0.34
160024	0.32	311	188	0.14	0.83	0.72	0.58
160081	-1.01	311	251	0.16	1.07	0.93	0.31
160011	-1.26	314	263	0.17	0.87	0.61	0.43
160012	-0.12	1220	743	0.07	0.96	0.88	0.38
160013	0.25	313	193	0.14	0.77	0.67	0.63
160014	0.08	313	202	0.14	0.94	0.89	0.50
160025	0.05	308	199	0.14	0.93	0.93	0.50
160026	0.41	602	364	0.10	0.87	0.77	0.56
160027	0.79	1502	704	0.06	0.95	0.93	0.48
160028	1.13	1660	682	0.06	1.14	1.19	0.38
160029	-1.07	389	316	0.14	1.12	1.83	0.22
160015	0.55	1645	832	0.06	1.34	1.51	0.21
160016	0.29	1511	842	0.06	1.01	0.98	0.39
160017	-0.13	596	411	0.10	0.89	0.76	0.53
160018	0.51	594	348	0.10	0.88	0.81	0.56
160082	-0.35	595	432	0.11	1.22	1.41	0.28
160031	0.45	287	176	0.15	0.84	0.71	0.61
160032	0.48	285	174	0.15	0.91	0.83	0.56
160033	1.88	286	111	0.15	1.56	1.80	0.23
160034	0.33	1202	644	0.07	0.88	0.80	0.49
160035	0.08	1004	571	0.07	1.03	1.05	0.42
160036	0.66	286	165	0.15	0.85	0.74	0.60
160037	-1.15	288	242	0.18	0.89	0.74	0.43
160038	-0.10	580	405	0.11	1.03	1.07	0.39
160039	0.41	571	351	0.10	0.99	0.92	0.48
160040	0.26	572	366	0.10	1.28	1.36	0.26
160041	0.07	577	387	0.10	0.89	0.84	0.52
160042	1.18	1292	512	0.07	1.19	1.26	0.30
160043	0.09	581	388	0.10	0.85	0.77	0.54
160044	-1.15	578	486	0.13	0.98	0.80	0.34

Item	Measure	Count	Score	SEM	In.Msq	Out.Msq	Pbs
160045	0.19	575	374	0.10	0.92	0.83	0.51
160046	-0.68	291	227	0.16	0.86	0.64	0.48
160047	0.02	287	193	0.15	1.26	1.41	0.22
160048	0.49	1352	679	0.07	0.90	0.84	0.47
160050	-1.14	293	246	0.17	1.02	0.94	0.32
160051	0.49	292	174	0.14	1.00	1.00	0.45
160054	0.24	1354	743	0.06	0.92	0.86	0.48
160055	-0.09	558	383	0.11	0.87	0.77	0.51
160056	0.10	359	226	0.13	1.22	1.34	0.25
160058	-0.09	555	381	0.11	0.99	1.02	0.42
160059	-0.88	366	287	0.14	1.08	1.21	0.26
160060	0.31	553	342	0.10	0.99	1.04	0.46
160061	0.44	1618	850	0.06	0.92	0.87	0.47
160062	0.65	1613	786	0.06	0.82	0.76	0.53
160063	-2.91	267	254	0.32	0.97	0.88	0.26
160064	0.17	262	165	0.15	1.24	1.47	0.29
160065	-0.86	267	210	0.17	0.88	0.74	0.48
160067	0.09	264	170	0.15	1.05	0.98	0.42
160068	-0.10	268	181	0.15	0.87	0.88	0.54
160069	-1.02	268	216	0.18	1.23	1.38	0.20
160070	0.19	1328	735	0.07	0.82	0.72	0.51
160071	0.51	1324	656	0.07	0.82	0.76	0.51
160072	-0.11	266	180	0.16	0.90	0.92	0.52
160073	1.22	1314	492	0.07	1.25	1.25	0.29
160074	-0.67	265	202	0.17	1.02	0.98	0.40
160075	-0.87	978	708	0.08	1.18	1.09	0.29
160076	0.22	1321	724	0.07	0.90	0.84	0.46
160077	-0.32	267	190	0.16	0.99	0.94	0.43
160078	-0.84	269	211	0.17	1.00	0.93	0.40
160079	0.61	1304	619	0.07	1.08	1.06	0.36
160083	0.37	106	45	0.22	1.07	1.02	0.32
160087	0.47	105	42	0.22	1.10	1.17	0.28
160088	-0.06	106	54	0.22	0.86	0.79	0.50
160089	-0.02	106	53	0.22	1.08	1.05	0.31
160091	-0.71	105	68	0.22	0.80	0.73	0.49
160092	0.19	105	49	0.22	1.01	0.98	0.29
160093	0.55	104	41	0.22	1.01	0.95	0.33
160094	0.76	105	37	0.22	0.92	0.86	0.42
160098	-0.94	148	97	0.19	1.02	1.02	0.29

Item	Measure	Count	Score	SEM	In.Msq	Out.Msq	Pbs
160100	0.93	148	44	0.20	1.15	1.37	0.23
160101	0.26	148	62	0.19	0.94	0.87	0.41
160103	0.13	147	65	0.19	0.98	0.91	0.38
160104	-0.55	106	69	0.24	1.02	0.89	0.47
160105	0.59	104	47	0.24	1.09	1.13	0.44
160106	-0.15	106	62	0.24	0.75	0.61	0.66
160109	0.20	105	55	0.24	1.06	1.04	0.48
160110	-1.37	106	82	0.27	0.87	0.62	0.49
160111	0.76	146	62	0.19	0.97	0.93	0.41
160112	-0.01	144	82	0.19	1.14	1.14	0.24
160113	0.70	148	65	0.19	0.90	0.83	0.48
160115	-0.32	147	93	0.19	0.96	1.01	0.37
160116	-0.86	147	107	0.20	1.09	0.98	0.26
160117	-0.50	147	98	0.19	1.02	1.05	0.32
160118	0.05	119	74	0.21	0.87	0.75	0.52
160119	-0.69	122	92	0.23	1.04	1.19	0.30
160120	0.79	121	59	0.21	1.27	1.29	0.25
160121	0.43	122	68	0.21	0.89	0.78	0.52
160122	-0.34	120	83	0.22	1.15	1.09	0.25
160123	-0.69	121	91	0.23	0.86	0.63	0.48
160124	-0.06	122	79	0.21	1.09	0.95	0.33
160125	0.92	122	57	0.21	0.94	0.99	0.50

Table 35. *Identificación de los Dibujos/Picture Naming Item Calibrations*

Item	Measure	Count	Score	SEM	In.Msq	Out.Msq	Pbse
110005	2.87	293	70	0.16	1.06	5.20	0.21
110006	-0.47	266	200	0.18	1.01	0.89	0.51
110007	-0.57	1298	932	0.08	0.93	0.97	0.46
110035	1.26	293	143	0.14	1.35	1.54	0.23
110009	-0.16	910	592	0.09	0.75	0.62	0.64
110036	0.58	274	164	0.15	1.37	2.10	0.23
110037	-0.99	280	225	0.19	0.93	0.92	0.60
110013	-1.77	304	263	0.22	1.14	1.08	0.49
110014	1.32	1416	577	0.06	1.23	1.61	0.23
110015	-0.65	288	221	0.18	0.87	0.70	0.60
110016	-0.27	286	207	0.17	0.76	0.58	0.68
110017	-0.81	281	224	0.19	1.22	1.65	0.37
110018	-1.53	297	253	0.22	0.90	0.65	0.60
110019	-0.42	276	204	0.17	1.17	1.16	0.47
110020	-0.45	277	208	0.17	1.11	1.51	0.43
110038	-2.56	577	521	0.19	0.78	0.39	0.61
110022	0.08	507	331	0.12	0.84	0.75	0.60
110023	-0.72	1578	1169	0.07	0.87	0.75	0.49
110024	-1.00	497	394	0.14	0.90	0.73	0.60
110025	-2.28	581	516	0.18	0.89	1.13	0.55
110026	-1.12	557	446	0.14	0.84	0.65	0.63
110027	-1.45	558	461	0.15	0.74	0.51	0.67
110028	0.96	436	232	0.12	1.14	1.22	0.36
110029	-0.46	466	343	0.13	0.97	0.88	0.57
110030	-1.11	258	205	0.19	0.97	0.77	0.54
110032	1.44	203	86	0.17	1.03	1.19	0.40
110033	0.58	1296	687	0.07	0.96	0.89	0.45
110040	-0.46	262	187	0.17	0.84	0.76	0.60
110041	-2.68	265	242	0.27	0.91	0.60	0.49
110042	-3.22	278	258	0.30	1.02	1.43	0.43
110043	-0.20	1206	802	0.08	0.87	0.81	0.47
110044	-0.07	1308	824	0.07	1.07	1.22	0.43
110045	-2.12	556	482	0.16	0.92	1.01	0.55
110046	-0.39	1675	1147	0.07	1.26	1.38	0.29
110047	-1.18	492	393	0.14	1.15	1.08	0.47
110048	0.55	1184	644	0.07	0.86	0.80	0.47
110049	1.57	493	187	0.11	1.15	1.34	0.31
110050	-0.65	527	384	0.13	1.04	1.09	0.54

Item	Measure	Count	Score	SEM	In.Msq	Out.Msq	Pbse
110051	-1.43	547	442	0.14	0.70	0.42	0.70
110052	-0.05	225	151	0.17	0.87	0.77	0.59
110053	-0.99	1338	1021	0.08	1.12	1.21	0.37
110054	1.46	267	105	0.15	1.06	0.99	0.37
110055	2.62	197	42	0.20	0.96	1.67	0.32
110056	-0.85	270	202	0.18	0.91	0.72	0.61
110057	1.75	220	76	0.16	1.17	3.25	0.21
110058	-2.45	281	248	0.25	0.97	0.53	0.55
110059	1.82	262	90	0.15	1.17	4.61	0.18
110060	-0.36	1045	699	0.08	1.07	1.06	0.45
110061	2.93	357	65	0.15	0.91	0.65	0.33
110062	0.88	1603	761	0.06	1.24	1.49	0.20
110063	-1.77	521	437	0.16	0.87	0.80	0.59
110064	-2.04	539	460	0.16	0.77	0.54	0.63
110065	-1.64	528	436	0.15	0.84	0.58	0.62
110066	-0.28	435	298	0.13	0.97	0.98	0.55
110068	-2.24	537	465	0.17	0.70	0.52	0.62
110069	-1.64	238	198	0.22	1.11	0.88	0.46
110070	-0.39	219	151	0.19	0.86	0.76	0.61
110071	0.01	154	101	0.20	1.17	1.10	0.34
110072	2.07	205	60	0.18	1.02	1.11	0.35
110073	-0.79	230	172	0.19	0.86	0.72	0.62
110074	-0.18	999	639	0.08	1.31	1.56	0.27
110075	-0.11	1178	746	0.08	0.84	0.76	0.52
110076	1.66	1253	414	0.07	1.04	1.44	0.31
110077	-1.26	237	189	0.21	0.89	0.73	0.60
110079	-0.33	1275	848	0.07	0.76	0.65	0.56
110080	-0.72	1264	917	0.08	0.88	0.80	0.51
110081	-3.27	252	233	0.32	0.75	0.28	0.54
110082	-0.36	1285	866	0.07	0.89	0.85	0.48
110083	1.88	903	258	0.08	1.08	1.23	0.31
110084	-0.87	199	154	0.21	1.27	1.44	0.33
110085	2.45	216	49	0.19	1.05	0.95	0.31
110088	-0.53	96	58	0.25	0.98	1.04	0.47
110090	-0.34	113	66	0.24	0.71	0.60	0.70
110091	-0.01	114	60	0.23	0.89	0.82	0.60
110093	-0.68	104	68	0.25	1.10	0.95	0.42
110094	-0.21	100	61	0.26	0.65	0.52	0.75
110095	0.51	113	55	0.22	0.97	0.94	0.53

Item	Measure	Count	Score	SEM	In.Msq	Out.Msq	Pbse
110096	-0.20	104	66	0.25	1.20	1.38	0.38
110098	1.09	100	40	0.24	0.95	0.87	0.51
110099	-0.34	97	60	0.25	0.86	0.81	0.54
110101	0.07	135	87	0.20	0.92	0.85	0.43
110102	1.15	135	57	0.20	0.85	0.93	0.51
110104	2.41	132	28	0.24	0.91	1.14	0.41
110106	2.61	100	19	0.29	1.15	1.24	0.20
110108	0.90	136	63	0.20	0.96	1.17	0.45
110111	1.47	100	34	0.25	0.97	0.86	0.45
110113	0.81	109	46	0.23	0.81	0.64	0.60
110114	-0.85	114	81	0.25	0.70	0.51	0.66
110115	0.49	120	60	0.22	1.37	1.79	0.23
110116	1.48	109	34	0.24	0.88	0.72	0.52
110118	2.13	103	28	0.25	1.05	1.07	0.23
110120	-0.04	134	88	0.22	0.77	0.66	0.63
110122	-0.08	114	78	0.24	0.86	0.91	0.60
110124	-1.11	137	109	0.26	0.87	0.64	0.56
110125	0.68	94	49	0.24	1.21	1.78	0.23
110126	1.52	111	43	0.22	1.02	0.96	0.28
110127	-0.76	98	77	0.29	0.86	0.65	0.48
110129	0.67	107	58	0.22	0.96	0.88	0.41

Table 36. *Primeros Sonidos/First Sounds Item Calibrations*

Item	Measure	Count	Score	SEM	In.Msq	Out.Msq	Pbse
100080	-0.58	315	233	0.14	1.02	1.10	0.30
100007	-0.31	315	218	0.14	1.07	1.13	0.29
100081	-0.36	317	222	0.14	0.90	0.82	0.44
100009	-0.85	320	249	0.15	0.91	0.75	0.39
100020	0.36	312	177	0.13	0.99	1.03	0.41
100021	0.91	316	148	0.13	0.97	0.97	0.44
100022	1.15	308	130	0.14	1.00	0.99	0.45
100023	0.49	317	173	0.13	0.99	1.01	0.40
100082	0.45	1352	685	0.06	0.93	0.89	0.39
100083	-1.25	318	265	0.16	1.00	0.90	0.27
100084	-0.46	758	505	0.09	1.00	0.97	0.36
100012	-1.80	318	283	0.19	0.99	0.93	0.24
100085	-0.50	317	229	0.14	0.96	0.90	0.37
100014	-0.59	755	520	0.09	0.99	0.96	0.35
100025	0.41	1360	700	0.06	0.97	0.94	0.36
100026	-0.10	599	381	0.10	1.04	1.00	0.32
100027	-0.07	597	376	0.10	0.96	0.94	0.41
100086	0.93	591	261	0.10	1.13	1.17	0.33
100029	0.28	590	333	0.10	1.00	1.03	0.39
100015	-1.03	598	474	0.11	0.94	0.79	0.34
100016	-0.82	598	455	0.10	1.07	1.18	0.23
100017	-0.54	600	429	0.10	1.01	0.95	0.32
100018	-0.27	594	396	0.10	0.96	0.94	0.38
100030	0.29	276	149	0.14	1.08	1.06	0.32
100031	0.22	279	155	0.14	1.08	1.07	0.31
100087	-0.01	1323	769	0.06	1.00	1.04	0.32
100088	0.58	279	135	0.14	1.03	1.11	0.38
100034	-0.50	281	194	0.14	1.06	1.04	0.26
100035	0.03	280	165	0.14	1.04	1.08	0.33
100036	-0.57	283	199	0.14	0.96	0.98	0.35
100037	-0.46	719	467	0.09	1.01	1.02	0.34
100038	0.67	576	269	0.10	1.05	1.06	0.40
100039	0.30	1616	850	0.06	1.04	1.03	0.34
100040	-0.74	1620	1153	0.06	0.94	0.95	0.30
100041	-0.53	577	394	0.10	0.99	1.01	0.39
100042	0.19	1608	875	0.06	1.03	1.02	0.33
100043	0.80	573	255	0.10	1.00	0.97	0.46
100044	0.24	1617	869	0.06	0.88	0.86	0.43

Item	Measure	Count	Score	SEM	In.Msq	Out.Msq	Pbse
100045	0.39	574	298	0.10	1.00	1.06	0.44
100046	0.94	290	121	0.14	1.27	1.36	0.32
100047	-1.23	291	228	0.16	1.00	0.98	0.36
100048	-0.41	296	193	0.14	0.92	0.98	0.49
100049	0.29	1339	703	0.06	1.10	1.13	0.29
100050	0.17	295	163	0.14	0.90	0.91	0.55
100051	0.74	295	133	0.14	1.18	1.25	0.36
100052	-1.58	295	243	0.17	0.93	1.10	0.41
100053	0.53	294	144	0.14	0.78	0.68	0.65
100054	0.33	1332	688	0.06	0.89	0.88	0.41
100055	0.16	1591	890	0.06	1.15	1.19	0.26
100056	-1.48	558	466	0.13	0.95	0.76	0.38
100057	-0.15	557	354	0.10	0.88	0.77	0.52
100058	0.92	558	251	0.10	0.85	0.82	0.57
100059	0.39	558	302	0.10	1.01	1.00	0.46
100060	-0.93	554	422	0.11	1.02	1.02	0.38
100061	0.69	554	272	0.10	1.18	1.21	0.34
100062	0.21	553	319	0.10	1.01	0.97	0.43
100063	-0.52	705	475	0.09	0.93	0.84	0.43
100064	-0.14	259	174	0.15	1.01	0.97	0.37
100065	0.86	257	125	0.15	1.20	1.32	0.30
100066	-1.06	262	212	0.17	0.84	0.69	0.48
100067	-0.58	265	197	0.16	0.87	0.73	0.47
100068	0.26	698	368	0.09	1.10	1.11	0.34
100069	0.73	261	133	0.15	1.24	1.26	0.27
100070	-0.40	1305	864	0.07	0.95	0.92	0.31
100071	-0.27	261	180	0.15	1.05	1.06	0.33
100072	0.27	1306	707	0.07	0.93	0.93	0.40
100073	0.22	1297	714	0.07	0.88	0.84	0.40
100074	-0.54	1294	886	0.07	1.10	1.23	0.20
100075	-1.32	264	223	0.18	0.95	0.86	0.34
100076	0.47	1309	659	0.07	1.14	1.20	0.28
100077	0.86	257	125	0.15	0.91	0.86	0.50
100078	-0.23	261	180	0.15	0.99	0.99	0.38
100079	-0.83	261	205	0.17	1.06	1.35	0.25
100089	0.86	99	34	0.24	0.98	1.05	0.38
100090	0.26	84	38	0.24	1.12	1.10	0.23
100091	0.54	100	40	0.23	0.92	0.90	0.44
100092	0.81	100	35	0.24	1.16	1.25	0.25

Item	Measure	Count	Score	SEM	In.Msq	Out.Msq	Pbse
100095	0.32	99	44	0.23	1.01	1.01	0.36
100097	0.50	104	41	0.22	0.99	0.99	0.35
100099	0.17	104	48	0.22	0.88	0.87	0.45
100100	0.26	104	46	0.22	0.95	0.92	0.38
100101	0.24	103	46	0.22	0.95	1.00	0.39
100109	0.00	1	0	0.00	1.00	1.00	0.00
100110	-0.21	152	91	0.18	1.02	0.98	0.34
100112	-0.30	154	95	0.18	0.89	0.88	0.47
100113	-0.22	152	92	0.19	0.86	0.78	0.49
100114	-0.40	154	98	0.19	0.87	0.80	0.47
100115	0.66	149	65	0.19	1.08	1.06	0.34
100116	1.01	101	40	0.25	1.19	1.20	0.40
100117	-0.41	103	67	0.23	1.01	0.89	0.39
100118	0.34	103	53	0.23	0.96	1.05	0.50
100119	-0.09	102	60	0.23	0.84	0.73	0.55
100120	0.73	103	46	0.24	1.03	1.01	0.49
100121	0.43	102	51	0.24	0.96	0.89	0.51
100122	-0.57	103	70	0.24	1.00	1.37	0.38
100123	0.88	101	42	0.25	1.21	1.23	0.39

Table 37. *Verbos (Expresivo)/Expressive Verbs Item Calibrations*

Item	Measure	Count	Score	SEM	In.Msq	Out.Msq	Pbse
140106	0.92	250	95	0.15	1.08	1.08	0.31
140006	-2.34	280	237	0.22	0.99	1.01	0.56
140020	0.76	216	88	0.16	0.95	0.98	0.38
140008	0.51	265	121	0.14	1.01	0.96	0.40
140009	-2.79	287	251	0.25	0.84	0.93	0.56
140010	-0.62	1727	1006	0.06	1.11	1.26	0.35
140011	-2.22	292	243	0.21	0.98	0.86	0.52
140012	-2.25	275	231	0.22	0.83	0.61	0.63
140103	-0.10	1563	780	0.06	1.02	1.01	0.38
140014	0.97	1371	428	0.07	1.06	1.27	0.29
140015	0.46	788	320	0.08	0.98	0.92	0.40
140016	2.53	208	30	0.22	0.99	1.39	0.23
140017	-2.33	296	247	0.21	1.36	1.74	0.37
140018	-1.16	1657	1111	0.06	1.16	1.31	0.33
140104	0.74	273	109	0.14	1.09	1.06	0.32
140099	2.31	268	43	0.18	1.02	0.96	0.25
140021	1.69	517	118	0.12	1.13	1.26	0.22
140022	0.19	1595	704	0.06	0.89	0.80	0.45
140105	1.33	349	99	0.13	1.00	1.15	0.29
140024	-3.07	562	500	0.19	0.80	0.61	0.55
140025	-1.78	1972	1483	0.07	0.82	0.72	0.53
140026	-0.50	1961	1115	0.05	0.99	0.93	0.40
140027	-0.89	1962	1243	0.06	0.94	0.93	0.44
140028	-1.86	1995	1519	0.07	0.95	0.88	0.47
140029	2.66	395	45	0.17	0.97	0.82	0.25
140031	-2.33	810	660	0.11	0.95	0.91	0.49
140032	-3.25	261	236	0.28	1.08	2.90	0.29
140107	-0.36	252	146	0.15	1.16	1.23	0.29
140034	-1.26	235	169	0.18	1.03	1.26	0.43
140101	2.15	215	35	0.20	0.96	1.04	0.25
140044	1.70	238	51	0.17	0.96	0.80	0.31
140045	-1.53	256	194	0.18	1.13	1.43	0.35
140047	-0.73	250	161	0.16	1.24	1.42	0.22
140048	1.50	488	113	0.12	0.93	1.01	0.33
140049	0.81	1954	643	0.06	1.08	1.05	0.33
140050	-1.44	319	230	0.16	1.11	1.27	0.40
140051	0.70	1911	660	0.06	1.03	1.22	0.32
140052	-3.64	528	473	0.21	0.92	0.82	0.50

Item	Measure	Count	Score	SEM	In.Msq	Out.Msq	Pbse
140053	1.43	978	227	0.09	0.97	0.97	0.33
140054	1.27	299	81	0.15	1.11	1.32	0.22
140055	-0.72	480	298	0.11	0.95	0.91	0.46
140057	2.29	226	29	0.21	0.93	0.90	0.30
140058	-0.68	274	161	0.15	0.98	1.03	0.51
140059	0.53	218	85	0.17	0.98	0.94	0.45
140061	-1.12	1677	1088	0.06	0.86	0.77	0.52
140063	0.66	252	92	0.15	1.02	1.02	0.39
140065	1.03	263	79	0.15	0.96	1.06	0.40
140067	-0.58	1667	937	0.06	0.98	0.97	0.43
140068	0.88	448	141	0.12	0.96	0.92	0.38
140069	0.74	409	134	0.12	1.18	1.26	0.24
140071	-1.19	1975	1300	0.06	1.01	1.05	0.40
140073	0.87	364	117	0.13	0.93	0.90	0.42
140074	-2.00	503	380	0.14	1.08	1.34	0.40
140075	-1.40	1929	1334	0.06	0.88	0.84	0.50
140076	0.46	487	184	0.11	0.92	0.89	0.45
140077	0.83	202	61	0.18	0.91	0.86	0.44
140079	1.29	244	53	0.18	0.99	0.96	0.35
140080	4.73	220	4	0.60	1.04	0.34	0.28
140081	4.49	231	5	0.52	1.05	0.60	0.27
140082	3.92	235	8	0.40	0.90	3.18	0.21
140084	0.84	226	66	0.17	1.07	1.12	0.36
140085	3.83	203	7	0.43	1.06	0.85	0.24
140086	1.88	223	33	0.21	0.98	0.79	0.32
140087	0.91	221	61	0.17	1.09	1.00	0.31
140092	0.80	539	167	0.11	0.99	1.09	0.34
140093	0.35	238	89	0.16	1.07	1.02	0.31
140096	0.35	216	81	0.16	0.91	0.89	0.47
140097	0.18	236	96	0.15	1.06	1.02	0.34
140098	-1.92	248	183	0.18	1.15	1.42	0.28
140144	-1.40	95	62	0.27	0.98	0.86	0.54
140108	0.90	108	29	0.25	1.22	1.32	0.27
140109	-0.35	103	48	0.24	1.15	1.32	0.37
140110	-2.27	97	74	0.31	0.85	0.56	0.62
140116	-0.75	110	65	0.23	0.92	0.85	0.49
140117	0.27	99	39	0.24	1.06	1.25	0.36
140118	-0.10	110	51	0.22	1.08	1.06	0.39
140119	-1.48	100	70	0.27	0.97	0.96	0.52

Item	Measure	Count	Score	SEM	In.Msq	Out.Msq	Pbse
140123	2.30	134	17	0.29	0.88	2.45	0.26
140124	0.40	164	63	0.18	1.08	1.26	0.25
140125	-2.50	164	140	0.26	0.92	0.73	0.45
140127	1.02	130	36	0.22	0.95	1.23	0.31
140139	0.42	99	44	0.23	1.13	1.30	0.27
140141	-1.70	93	72	0.31	0.90	1.01	0.53

STANDARD SETTING

Background

MTSS are designed to provide a structure for evaluating student performance and for differentiating instructional interventions (allocated at three tier levels: Tier 1: universal, Tier 2: targeted, and Tier 3: intensive) based on children's demonstrated need (Greenwood, Carta, McConnell, Goldstein, & Kaminski, 2009; Vaughn & Fuchs, 2003; Coleman, Roth, & West, 2009; Jackson, Pretti-Frontczak, Harjusola-Webb, Grisham-Brown & Romani, 2009). MTSS require measures that efficiently identify children in need of more intensive levels of intervention (Fuchs & Fuchs, 2006; McConnell & Missall, 2008; Vaughn & Fuchs, 2003) and often encompass Response to Intervention models.

IGDI-E was designed specifically for use in an MTSS, providing quick and reliable information about students' Spanish early literacy and language performance to meaningfully inform intervention candidacy of young SE bilinguals. By using the measures to define candidacy for tiered intervention, we aim to improve the likelihood that appropriate instruction and interventions will be provided, in turn improving the possibility for the student to develop English reading skills (e.g., Al Otaiba, et al., 2011).

In many existing MTSS frameworks, school districts use a normative approach to identify candidates for Tier 2 and Tier 3 intervention. Early research on MTSS suggested 80% of students in a typical classroom are candidates for Tier 1 or success in the universal curriculum, 15% are candidates for Tier 2 or targeted intervention, and 5% are candidates for Tier 3 or intensive intervention (Greenwood, Bradfield, Kaminski, Linas, Carta, & Nylander, 2011). These results have been used to examine classroom-level performance by applying the normative distribution to student scores, such that student scores are rank ordered and the top 80% of the performance group are considered Tier 1 candidates, whereas 15% are Tier 2 candidates and the lowest 5% are Tier 3 candidates. Rank ordering allows educators to map resources and efforts to students easily, with the results always producing manageable groups for intervention while ensuring the students who need the most support receive it.

Claim 10: *Performance standards at each tier level were set based on empirical data, expert review panels, and information from parents and teachers, and revised based on longitudinal analysis of children's Kindergarten performance.*

Claim 15: *Item difficulties are stable across seasons.*

However, norm-referenced methods for determining candidacy within MTSS also come with challenges. Using the normative method, students are evaluated through the lens of the observed distribution levels without regard to which students have the necessary early literacy and language skills to be proficient readers in later grades. That is, using the normative method, 80% of students will be identified as Tier 1 candidates without regard for the degree to which these students are on target in the early literacy domains. Normative methods for using MTSS focus on the performance of students relative to other students, rather than on student performance relative to the ability of interest. With a normative approach, high-quality teachers who apply effective instruction are able to bring all students in their classroom to a level consistent with later reading success, however they will still identify some Tier 2 and Tier 3 candidates. Likewise, a teacher who has limited skill may also identify most students at the Tier 1 level, when in fact, they may not actually be Tier 1 candidates. This is particularly true for bilingual students given the heterogeneity in the population, the lack of teacher knowledge in how to deliver effective instruction to SEB students, and the documented need for language support in both L1 and L2 (Hoff, 2013).

A criterion-referenced method for identifying candidates more accurately examines student performance relative to expectations for performance (i.e., knowledge, skills, and abilities). In a criterion-referenced approach, a standard is set for each Tier level that corresponds with benchmarks or identified skill sets known to predict student performance at a level consistent with later reading success. One example of existing criterion-referenced methods for MTSS is the use of English IGDIs 2.0 cut scores to identify candidates for Tier 1 and Tier 2/3 candidacy, developed through the Center for Response to Intervention in Early Childhood (Greenwood, et al., 2008).

Standard Setting: A Brief Introduction, Rationale and Methods
Standard setting is a class of methods employed to set cut scores on tests that are associated with performance levels. Although there are many methods available (see Cizek & Bunch, 2007), the method chosen should be consistent with the interpretative argument of the test; that is, standard setting procedures should support the intended interpretations and uses of the test.

All standard setting methods involve human judgment at some stage of the process. However, methods differ in a number of important ways. Procedures generally focus on test takers (person-centered) or test items and tasks (test-centered). Procedures also generally employ performance level descriptors (PLDs). Standard setting results in identifying one or more cut-scores that define the performance level associated with test scores, such as certified, licensed, proficient, accomplished, passed, etc. Each performance level must be described adequately to support score interpretation and use: What does a proficient student know or what can a proficient student do? Finally, there is a process that translates PLDs, a

review of test takers or test items, to identification of one or more scores on the test that separate performance levels.

The *Testing Standards* (AERA, APA, & NCME, 2014) provide guidance regarding the selection, use, and documentation of standard setting procedures: The methods used to establish performance cut scores should be documented (Standard 7.4); the participants in the standard setting process and their relevant expertise should be documented (Standard 7.5); and their training and engagement in the process should be documented (Standard 1.9). In the *Testing Standards*, Chapter 5 devotes cluster 4 standards to setting cut scores. Among these standards, documentation is required describing the rationale and procedures used (Standard 5.21), allowing judges to employ their knowledge and experience in a reasonable way (Standard 5.22), and when feasible, providing sound empirical evidence that categories defined by cut scores are associated with relevant criteria (Standard 5.23). Many of these standards can be achieved through the use of empirically grounded standard-setting methods with sufficient validity evidence.

The use of standard setting in early childhood assessment is relatively new. An important functional step in MTSS is the identification of children who are likely to benefit from Tier 2 or 3 interventions. This identification process is improved through the collection and use of relevant information. That is, information relevant to the skills seen as important for successful language and literacy development through early childhood and early elementary school is important to assess and monitor to ensure successful development and progress. IGDIs-E were developed to contribute criterion-referenced information to support teaching and learning in SEB students.

In this study, we explored the viability of setting performance standards to support the use of IGDIs-E in MTSS by addressing the following questions: *What methods are appropriate for early childhood assessments with bilingual children? Are performance standards in this context sensitive to differences in standard setting methods? Do judges and expert informants understand their roles, understand the procedures sufficiently to complete the assigned tasks, and feel confident in the results of their work? What are the preliminary classification rates of a multi-state sample of SEB children among the three tiers of supports?*

MTSS are predicated on the availability of assessment results with sufficient evidence of reliability and validity to defend the identification of individual children who will likely benefit from different tiers of support. A review of current standard setting practices suggested the contrasting groups design is most appropriate for developing scores for selection and placement decisions (Cizek & Bunch, 2007). The most common method of setting performance standards in K-12 achievement tests is an item-mapping method referred to as the Bookmark procedure. By employing

multiple methods, we are able to evaluate the sensitivity of performance standards due to choice of methods, and to gain a deeper understanding of performance expectations for Spanish language and literacy development among preschool aged children. Moreover, these processes and their results provide additional vehicles for the evaluation of validity evidence for the IGDIs-E.

Level of performance and performance categories. The contrasting group method requires that experts identify the performance level of participating children based on operationally defined domains of early literacy. In the PLD surveys (available upon request), bilingual teachers reviewed a carefully constructed operational definition of the domain, reflected on their students' current level of skills particular to that domain, and identified those students who are successful in typical instruction (Tier 1) or who need moderate (Tier 2) or significant (Tier 3) support. Parents also received a PLD survey in which they reflected on their child's ability in each domain and provided a rating associated with each tier level descriptor. It was important to include parents in a standard setting process because they have been found to be more accurate reporters of their child's Spanish language ability than preschool teachers who are often monolingual English speakers (Limbos & Geva, 2001; Bedore, Peña, Joyner, & Macken, 2011; Mancilla-Martinez, Gamez, Vagh, & Lesaux, 2016). Parents and teachers completed PLDs for each season: fall, winter and spring.

To provide a stronger basis for standard-setting with IGDIs-E given the limited use of standard-setting methodologies in early childhood assessment, an item-mapping procedure based on the method developed by CTB/McGraw-Hill, the Bookmark procedure (Mitzel, Lewis, Patz, & Green, 2001), was also used. The Bookmark procedure involves an expert panel that reviews an ordered-item booklet (a booklet of test items in increasing order of difficulty) and identifies the place (i.e., item) in the booklet that defines the point that differentiates student performance at each tier level. Once students' performance levels were recommended by all three methods, teacher, parent and expert judgments were compared. Cut scores were then averaged across sources to provide a more stable standard of performance.

Standard setting sample. Once standard setting procedures were concluded we produced cut scores using the contrasting groups design and applied these cut scores to child level responses on the IGDIs-E items. Families and children participating in the standard setting study are described in Table 38.

Table 38. *Participating Child and Parent Characteristics (n=396)*

Characteristic	Percent %
Female	47.5
Special education services	4.0
Ethnicity/Race	
Latino (general)	52.1
Mexican	15.4
Puerto Rican	14.3
Caribbean	2.5
Central American	2.8
South American	1.5
Multiple races/ethnicities	11.5
Languages spoken to the child from ages 0 to 1	
Spanish	71.0
Both	24.3
Language the child uses when talking at home	
Spanish only	48.5
Both	24.0
English	12.0
Other	15.5
Household weekly income	
Less than $500	63.1
$501 – 700	22.7
$701 – 900	7.7
More than $901	6.6
Mothers highest level of education	
6th grade or less	13.9
Less than 12th grade	18.0
GED	12.1
High school diploma	11.3
Some education after high school / vocational program	16.6
Associate degree (AA)	9.7
College degree (BA/BS)	13.1
Graduate/ professional degree	5.4

ROC analysis and classification accuracy. Using all three methods we produced the finalized cut scores. Receiver operator curves are depicted for parent and teacher performance level descriptors, including the related Area Under the Curve (AUC) statistics to examine the classification accuracy of each approach (Table 39).

Table 39. *Final Tier 2/3 Cut Scores by Method*

Measure	Bookmarking	Teacher PLD	Parent PLD	Average
Identificación de los Dibujos (PN)	1.02	0.29	0.51	0.61
Verbos Expresivo (EV)*	-0.58	-0.64	-0.57	-0.60
Primeros Sonidos (FS)	-0.43	0.48	0.61	0.22
Identificación de las Letras (LN)	-0.22	0.66	0.80	0.41
Identificación de los Sonidos (SI)	-0.12	0.95	1.02	0.62

* *Verbos Expresivo* (EV) results exclude the Minnesota Bookmarking result.

When examining the ROC and AUC statistics for each measure we set the target sensitivity statistic at .70 and identified the pint that maximized specificity. We argue that in practice, teachers are more likely to over-identify students for intervention and then remove them based on initial performance than they are likely to independently and reliability identify students who are in need of intervention, but were not identified by the measure using the identified cut score. As such, we established a sensitivity of .70 as a minimum. As depicted in Table 40, for some measures teacher and parent confidence in their scores and ratings were highly variable, potentially resulting in low specificity statistics.

Table 40. *Sensitivity, Specificity and AUC for all Measures*

	Teacher PLD			Parent PLD		
	Sens	Spec	AUC	Sens	Spec	AUC
Identificación de los Dibujos (PN)	.70	.60	.69	.70	.50	.66
Verbos Expresivo (EV)*	.71	.59	.73	.71	.57	.72
Primeros Sonidos (FS)	.71	.46	.61	.72	.47	.57
Identificación de las Letras (LN)	.70	.52	.66	.71	.60	.72
Identificación de los Sonidos (SI)	.70	.38	.60	.70	.61	.70

The contrasting groups design yielded cut scores that yielded sufficient sensitivity to be used during the IGDIs-E design process. However, a long-term goal of the IGDIs-E project was to secure improved cuts based on criterion-related predictive validity evidence in Kindergarten. In this way, scores on IGDIs would establish cut scores for achieving an identified status (typically 50 to 60% probability) on a standardized measure of English or Spanish literacy skills at the end of Kindergarten. For clarity, the cut scores from the predictive validation study are provided in Table 41. However, the full study design of the predictive validation study is described later in this manual on pages 112-116.

Table 41. *Sensitivity and Specificity for (Predictive) Validity Coefficients (LiMER)*

	Probability of prediction	Cut anchor scale	IGDIs-E theta cut	Sens	Spec
Identificación de los Dibujos (PN)	.60	CELF Exp. Language	0.69	.74	.64
Verbos Expresivo (EV)*	.55	CELF Exp. Language	-0.27	.77	.57
Primeros Sonidos (FS)	.50	CELF Phono. Awareness	0.67	.65	.65
Identificación de las Letras (LN)	.50	DIBELS	0.61	.69	.58
Identificación de los Sonidos (SI)	.50	TERA	0.89	.67	.67

Sample-specific frequencies of tier level candidacy. Once cut scores were identified, we examined the frequencies at which each tier level group (Tier 1 and Tier 2/3) would be present in our field test population to provide data that examines the distribution of young SEB students in each tier. Following the standard setting process, a new group of students was recruited in the 2014-2015 academic year. This sample included the same states and sites as reported in the participant section. Students received the same five IGDIs-E measures but items were selected as close to each cut score as possible to create the testing forms. We determined 15 items were required to produce a reliable student ability score for each measure (by examining changes in SEM due to number of items). Thus, each measure included 15 items with item difficulties as close to the cut score as possible. This sample included 471 children, and of this sample, 330 provided demographic information (see Table 42).

Table 42. *Participating Family Demographics for Children Completing IGDIs-E Screening Sets (n=330)*

Characteristic	%
Female	53.2
Special education services	7.6
Latino (general)	57.1
Mexican	18.0
Puerto Rican	9.6
Caribbean	1.6
Central American	5.6
South American	0.3
Multiple races/ethnicities	5.6
Other	2.2
Languages spoken to the child from ages 0 to 1	
Spanish	71.2
Both	25.5
Language the child uses when talking at home	
Spanish only	49.1
Both	42.7
English	8.2
Household weekly income	
Less than $500	67.2
$501 – 700	27.4
$701 – 900	0.0
More than $901	5.5
Mothers highest level of education	
6th grade or less	19.5
Less than 12th grade	22.6
GED	7.3
High school diploma	22.0
Some education after high school / vocational program	18.1
Associate degree (AA)	0.0
College degree (BA/BS)	5.2
Graduate/ professional degree	5.2

Note: The multiple ethnicities group includes those who selected two or more ethnicities. Twenty percent of parents did not report household income (n=56) and 15% did not include mother's highest level of education (*n*=43).

Results presented in Table 43 illustrate the sample size and the percentage of the total sample in each performance level (Tier 1 or Tier 2/3) for each season. Samples vary across seasons and measures due to discontinuing the measure or child absences. In these results we include a SEM group, labeled Tier M, where scores fall within ±1 SEM of the cut score. Because of measurement error, as is the case in all tests, our confidence regarding true candidacy recommendation within this 1 SEM range is limited and more information is needed to make a meaningful decision. Tier M includes children in Tiers 2/3, plus those falling within 1 SEM of the cut score. With this constraint, three groups are reported: Tier 1, Tier M, and Tier 2/3. Trends indicate that the Tier 1 candidacy groups generally grow from fall, to winter, to spring for all IGDIs-E measures with the largest percentage of candidates for Tier 1 observed in spring for the *Primeros Sonidos/First Sounds* measure (46%). In reciprocal fashion, the Tier 2/3 group decreases across seasons for all measure except *Primeros Sonidos/First Sounds*, where we observe a spike in Tier 2/3 candidacy (44%) during winter.

Table 43. *Tier Level Frequencies by IGDIs-E Measure and Season*

Measure	Season	Total sample	Tier 1	Tier M	Tier 2/3
Identificación de los Dibujos/ Picture Naming	Fall	399	26%	38%	36%
	Winter	409	36%	37%	27%
	Spring	370	42%	32%	26%
Primeros Sonidos/ First Sounds	Fall	352	26%	38%	36%
	Winter	198	38%	28%	44%
	Spring	342	46%	30%	24%
Identificación de las Letras/Letter Naming	Fall	378	23%	31%	46%
	Winter	394	43%	25%	32%
	Spring	355	45%	27%	28%
Identificación de los Sonidos/ Sound Identification	Fall	350	13%	29%	58%
	Winter	197	31%	24%	45%
	Spring	350	34%	33%	33%
Verbos Expresivo/ Expressive Verbs	Fall	383	31%	30%	39%
	Winter	280	40%	28%	32%
	Spring	313	41%	25%	34%

VALIDITY EVIDENCE

Expanding on Kane's Model

At its core, the IGDIs-E approach to constructing measures rests on validation through an argument-based presentation of assumptions, inferences, and intended uses regarding students' performance (Kane, 2013). The interpretation and use argument (IUA) provides clear guidance to structure our approach to assessment design by requiring substantive evidence for each argument or claim presented. As such, the goal of the validation process is to defend the measure's specified arguments and claims (Kane, 2013). This approach transitions away from perceptions of validity as a checklist and gathering validity evidence for the sake of validity or exploring the degree to which a tool can be used for identified purposes by confirming the tool is measuring what it was intended to measure. Instead, this view of validity builds on initial models presented by Messick (1980) and others (Kane, 2013) and demonstrates that although test design may specify the intended use of a measure, demonstrating evidence to support such claims is a separate process. There are two components to this process: the IUA describing inferences, assumptions, claims and proposed uses of the measure; and the validity argument detailing the evidence gathered to defend the IUA (Kane, 2013). Validation is the dynamic process of continually informing these arguments, such that validity is not achieved; but rather it is fluid and continually evolving based on the current IUA. Given that validity is a process of strategically collecting evidence, we focus on an approach that emphasizes conceptual rigor in design and empirical evidence to define and support the IUA at the item level.

Psychometric Evidence

All of the item-level statistics reported above provide evidence regarding the psychometric quality of items and total scores. To illustrate psychometric evidence for the IGDIs-E measures we have completed two primary analyses that depict the trends and expected trajectories of performance. Here we include the descriptive statistics and growth trajectories for each measure.

Descriptive performance. To examine the distribution of student performance across the sample used in the IGDIs-E validation process we examined descriptive statistics including mean, range, minimum, maximum, standard deviation, and percentage of zero scores for each measure by each season. Results are provided in Tables 44-46.

Claim 4:
IGDIs-E are developmentally appropriate for SEB 4-5-year-old children.

Claim 5:
IGDIs-E are a set of screening measures that can accurately identify student skill level in Spanish within the context of differentiated instruction or within a multi-tiered system of support.

Claim 7:
IGDIs-E are inclusive of a variety of Spanish dialects and socio-economic backgrounds that are representative of Spanish speaking populations.

Claim 10:
Performance standards at each tier level were set based on empirical data, expert review panels, and information from parents and teachers, and revised based on longitudinal analysis of children's Kindergarten performance.

Claim 12:
IGDIs-E have the potential to be used in a variety of early childhood programs.

Table 44. *IGDIs-E and English IGDIs Descriptive Statistics for Fall 2014-2015*

	*N**	*M*	*SD*	*Range*	*Skew*	*SE*	Percentage of discontinues/(n)
Spanish							
PN	393	0.24	1.52	8.07	-0.57	0.08	7% / 33
EV	383	-0.87	1.73	6.92	-0.73	0.08	11% / 48
FS	340	0.24	1.20	7.79	0.72	0.06	20% / 83
LN	371	0.22	1.27	7.81	0.11	0.07	12% / 52
SID	342	0.03	1.43	7.84	0.01	0.08	19% / 81
English							
PN	373	-1.15	1.57	5.77	0.01	0.08	9% / 39
FS	262	1.04	1.24	7.79	0.75	0.08	9% / 29
SID	380	-0.41	1.73	7.79	-0.16	0.09	7% / 31
RH	279	-0.77	2.00	7.79	-0.13	0.12	32% / 133

Table 45. *IGDIs-E and English IGDIs Descriptive Statistics for Winter 2014-2015*

	N*	M	SD	Range	Skew	SE	Percentage of discontinues/(n)
Spanish							
PN	409	0.66	1.41	8.07	-0.64	0.07	6% / 26
EV	393	-1.82	2.34	8.46	-o.13	0.11	11% / 40
FS	432	0.47	2.06	7.79	-0.155	0.10	14% / 62
LN	434	0.44	1.45	7.81	0.35	0.09	9% / 39
SID	432	0.27	2.05	7.84	0.01	0.10	11% / 48
English							
PN	428	-1.27	1.41	8.06	0.35	0.07	5% / 23
FS	424	1.57	1.9	7.83	-0.68	0.09	7% / 30
SID	427	0.74	1.9	7.87	0.41	0.09	9% / 40
RH	427	-0.33	2.25	7.84	-0.13	0.11	23% / 111

Table 46. *IGDIs-E and English IGDIs Descriptive Statistics for Spring 2014-2015*

	N*	M	SD	Range	Skew	SE	Percentage of discontinues/(n)
Spanish							
PN	387	0.65	1.72	8.07	-0.81	0.09	3% / 13
EV	357	-1.05	2.16	8.46	-0.59	0.11	8% / 28
FS	386	0.28	1.98	7.79	-0.34	0.10	11% / 44
LN	386	0.56	1.77	7.81	-0.37	0.09	7% / 28
SID	386	0.62	2.10	7.84	-0.04	0.10	9% / 35
English							
PN	349	0.01	1.45	6.58	-0.39	0.08	4% / 15
FS	349	1.22	1.50	7.96	-0.80	0.08	4% / 17
SID	349	1.07	1.88	7.90	0.11	0.10	4% / 17
RH	348	0.02	2.12	7.85	0.24	0.11	18% / 65

Estimating growth. Growth over time is an important statistic used to examine and evaluate progress, or how student performance changes across the academic year. IGDIs-E are used to produce student ability scores at three seasonal time points and evaluate that ability against a benchmark to determine tier level candidacy. Because IGDIs-E are screening tools we posit that limited information should be made about seasonal growth because the tools are not designed to measure progress in this way. Instead, our IUA notes that the tools are designed to identify candidates for tiered intervention. Essentially, the IGDIs-E make a yes/no decision about performance that answers the question: Is this student on track in the current early literacy and language instructional environment, or do they need more intensive intervention? Examining growth across three seasons is counterintuitive because growth and status (yes/no) are not compatible. As such, to measure growth over time, we needed to use an expanded set of items that were designed to produce ability estimates at more frequent intervals.

This type of data help us to examine student abilities over time (fall, winter and spring) so that we can understand if they are growing adequately- that is, what is the expected level of growth on Spanish early language and literacy skills over one academic year as measured by Spanish IGDIs? To explore this question, we examined growth curves for 75 students who were a subsample of the Year 2 IGDIs-E study.

Sample participants. As noted, a sample of 75 children (33 Females) was included in the growth analyses. We were interested in modeling growth for students in "business as usual" classrooms across the entire academic year where no identified interventions were in place to examine models for typical growth trajectories. During Year 2, we recruited a strategic sub-sample for monthly assessment. Our recruitment pool included 100 students selected from the four regional dialects (Mexican, Caribbean, South American, and mixed regional representation) stratified by age in months with at least 3 students in each monthly age bracket between 4 (48 months) and 5 years (60 months) of age. We specifically selected students who represented all of the ages in months present in the pre-Kindergarten year for each regional dialect to create a most complete picture of potential growth. HLM was used to estimate the intercept and growth curves for each child. The original design of the study called for a child to be tested seven times in nearly equally spaced time point during the school year. However, in practice, testing dates varied across children and not every child was tested at every time point (on average, a child was tested about four times during the study). Pairwise deletion was used to exclude cases where the measure was discontinued or a child did not respond to any item for a given measure in a given time point (EV=4.6%, PN=7%, FS=5.2%, LN=4.5%, SID=7.5%, and SB=1.7%). The statistical software HLM (version 6) was used to complete the analyses.

Analysis and growth estimate results. The HLM model was a two-level growth model across five measurement waves during the 2013–14 academic year. The time variable, *Months*, was defined as the amount of time in months that had elapsed from the first administration (10-15-2013) of the IGDIs-E tasks (where 1 month = 30 days). The scores were in Rasch logits. Last recorded date of IGDIs-E administration was 05-22-2014.

Claim 6: *IGDIs-E are designed to measures change over a school year.*

The intercept and slope can be interpreted as follows in the unconditional model:

- INTERCEPT: The intercept can be interpreted as the *IGDIs-E* score at the first data point (10-15-2013) (initial status).
- SLOPE: The interpretation of the slope is the change in *IGDIs-E* score per month (growth rate).

To investigate the intercept and slope parameters, a random-effects (unconditional) model was estimated. This was based on the following specification, for child *i* at time *t*. The Level-1 model estimates the intercept (π_{0i}) and slope (π_{1i}) for each child *i*. The Level-2 model estimates the grand-mean for both the intercept (β_{00}) and slope (β_{10}) across children.

Level-1 Model: $$Score_{ti} = \pi_{0i} + \pi_{1i}(Months)_{ti} + e_{ti}$$

Level-2 Model: $$\pi_{0i} = \beta_{00} + r_{0i}$$
$$\pi_{1i} = \beta_{10} + r_{1i}$$

Results are presented in Table 47, and indicate significant growth for all IGDIs-E tasks with the exception of *Expressive Verbs/Verbos Expressivo*. Growth estimates for the *Identificación de los Sonidos/ Sound Identification* and *Identificación de los Dibujos/Picture Naming* tasks illustrate approximately 0.10 logits of growth per month, suggesting in a business as usual pre-Kindergarten classroom, performance should increase approximately 1.0 Rasch logits over the course of an academic year. *Identificación de las Letras/Letter Naming* produced similar growth estimates; however the growth was unreliable, indicating substantial instability in the estimates and limiting our ability meaningfully interpret the slope for this task.

For *Primeros Sonidos/First Sounds*, growth estimates suggest .18 logits of *growth* per month, indicating that in a business as usual pre-Kindergarten classroom, performance should increase approximately 1.8 Rasch logits over the course of an academic year. For the *Verbos Expresivo/Expressive Verbs* task, growth was negative, illustrating a decline in performance across the academic year. This may be because the *Expressive Verbs* task has no equivalent capacity in English, given the differential salience of verbs in Spanish as compared to nouns in English. The negative growth present for this task may illustrate an instance of language loss, as noted in the research literature by Anderson and others (2004).

Table 47. *HLM Growth Estimates for IGDIs-E Items*

Measure	Fixed effects	*Coefficient*	*SE*(β)	*p*-value	*Reliability* of slope
Verbos Expresivo/ Expressive	Intercept, β_{00}	-0.68	0.17	.00	
	Slope, β_{10}	-0.04	0.02	.18	.37
Identificación de los Dibujos/Picture Naming	Intercept, β_{00}	0.6	0.20	.00	
	Slope, β_{10}	0.09	0.02	.00	.48
Primeros Sonidos/ First Sounds	Intercept, β_{00}	0.47	0.15	.00	
	Slope, β_{10}	0.18	0.02	.00	.31
Identificación de las Letras/Letter Naming	Intercept, β_{00}	0.74	0.17	.00	
	Slope, β_{10}	0.10	0.02	.00	.04
Identificación de los Sonidos/ Sound Identification	Intercept, β_{00}	0.83	0.19	.00	
	Slope, β_{10}	0.09	0.03	.00	.29

It is important to note that this study of growth only included 75 students, and as a result, some of the growth estimates are particularly unreliable (e.g. *Letter Naming*). Although K-12 measures suggest reliability of the slope should approach .8 for robust measures, there are few (if any) standards for appropriate estimates in early childhood when performance is highly variable due to child-level factors that may contribute a level of error that is significantly larger than that of K-12 student performance.

One way to improve the analysis of growth over time is to increase the sample size and number of time points for each measure. As IGDIs-E research continues, we anticipate updating technical reports of growth with forms designed for this purpose: progress monitoring measures. Please consult the IGDIs-E progress monitoring technical manual for updated growth estimates.

Construct-Related Validity Evidence

Another important factor in establishing a validity argument is the degree to which a measure illustrates the construct of interest. IGDIs-E provide construct-related validity evidence through four approaches: item maps, principal component analyses, construct representation analyses, and correlations between IGDIs-E measures.

Claim 4: *IGDIs-E are developmentally appropriate for SEB 4-5-year-old children.*

Item maps. Wright item maps are visual depictions of how items are distributed relative to child level abilities. A measure is adequately representing a construct when item distributions mirror the expected distribution and rank-order of item difficulties. This ordering should also map onto student abilities, based on student characteristics. For example, if preschool age SEB children have only very limited knowledge of nouns, their abilities, as a group, would be very low. Therefore adequate construct coverage would include items that are very easy and match the identified abilities of the students. Figures 10-14 depict the IGDIs-E item maps. In these maps, items are depicted on the right-hand side and child abilities are depicted on the left-hand side. A measure is meaningfully capturing child level abilities when the items are distributed in ways that mirror the child ability distributions. Figures 10-14, resulting item maps, illustrate that all IGDIs-E measures have adequate construct coverage and meaningfully align to the majority of student abilities. Those students whose abilities that are very high or very low have fewer items that match their capacity, suggesting potential floor and ceiling effects.

Figure 10. *Identificación de los Sonidos /Sound Identification Item Map*

REPORTED: 2237 PERSON 99 ITEM

```
          PERSON - MAP - ITEM
              <more>|<rare>
    4 .############  +
                     |
                   . |
                .### |
                   . |
                  .# T
              .###### |
    3              .#  +
              .####  |
                  #  |
                .##  |
              .#####
              .#####
                 .#
    2          .#### S+
              ######    160033
              .#####
              .####
               #####   |T
              .####    | 160073
            .#######   | 160028  160042
    1          .####  +  160100
            .######    | 160027  160120  160125
             .#####    |S 160036  160062  160080  160094  160111  160113
          .######## M  | 160015  160018  160071  160079  160093  160105
         .#########   | 160006  160026  160031  160032  160039  160048  160051
                      | 160061  160083  160087  160121
              .#####  | 160008  160013  160016  160024  160034  160040  160054
                      | 160060  160076  160101
      .############  | 160014  160035  160043  160045  160056  160064  160067
                      | 160070  160092  160103  160109
    0     .#########  +M 160020  160025  160041  160047  160088  160089  160112
                      | 160118  160124
       .############  | 160009  160012  160017  160023  160038  160055  160058
                      | 160068  160072  160106
        .###########  | 160077  160082  160115  160122
           .########  | 160117
         .#########    160007  160104
           ######## S |S 160005  160046  160074  160091  160119  160123
             .#####    | 160059  160065  160075  160078  160116
   -1       .######  +  160029  160069  160081  160098
               ###    | 160037  160044  160050
              .##     | 160011
             .###    T 160110
                .
               .#
                .
   -2            .  +
              .## T
                .  |
                   |
                   |
                .  | 160063
   -3           .  +
                .  |
                   |
                   |
                   |
                   |
                   |
   -4           .  +
             <less>|<frequ>
EACH "#" IS 10. EACH "." IS 1 TO 9
```

Figure 11. *Identificación de los Dibujos/Picture Naming Item Map*

REPORTED: 2283 PERSON 106 ITEM

```
          PERSON - MAP - ITEM
              <more>|<rare>
    5          .####  +
                      |
                      |
                      |
                      |
                      |
                      |
                    .
    4               .  +
                    .  |
                    .  |
                    . T|
                  .#   |  110109  110123
                  .#   |
                  .#   |
    3           .###   +  110061
                ####   T| 110005
                 ###   |  110055  110106
              .######   |  110085
                .####   |  110104
              .###### S |  110118
    2 .############   +  110072
              .######   |  110059  110083  110128
       .############   |  110057  110076
      .#############   S| 110032  110049  110054  110111  110116  110126
       .############   |  110014  110035
       .############   |  110098  110102
    1  .############   +  110028  110130
        .##########    |  110062  110108  110113
        .########## M  |  110119  110125  110129
          .########   |  110033  110036  110048  110095  110115
          .#######    |  110100
          .#######    |
    0     .#######   +M| 110022  110044  110052  110071  110091  110097  110101
                       |  110120  110122
            .#####     |  110009  110043  110074  110075  110094  110096
            .#####     |  110016  110046  110060  110066  110070  110079  110082
                       |  110090  110099
              ####     |  110006  110007  110019  110020  110029  110040  110088
             .####     |  110015  110023  110050  110080  110086  110093
             .#### S   |  110017  110056  110073  110084  110114  110127
   -1          .###   +  110024  110037  110053
               .###    |  110026  110030  110047  110124
                .##    |  110077
                .##    |S 110018  110027  110051
                .##    |  110065  110069
                .##    |  110013  110063
   -2           .##   +  110064
                 .#    |  110045  110068
                 .# T  |  110025
                 .#    |  110038  110058
                 .#    |  110041
                  .    |T
   -3             .   +
                 .#    |  110042
                  .    |  110081
                  .    |
                       |
                  .    |
   -4             .   +
                 .#    |
                  .    |
                  .    |
                       |
   -5         ######  +
              <less>|<frequ>
EACH "#" IS 10. EACH "." IS 1 TO 9
```

Figure 12. *Primeros Sonidos/First Sounds Item Map*

REPORTED: 2218 PERSON 98 ITEM

```
PERSON - MAP - ITEM
          <more>|<rare>
4  .###########  +
                 |
                 |
                 |
                 |
             .#  |
             ##  |
             .#  |
3             .  +
           .###  |
             .# T|
             .#  |
            .##  |
             .   |
             .#  |
2       #######  +
           .###  |
            .##  |
          .##### S|
           .###  |
           .###  T|
        .########|  100022
1          .###  +  100046  100086  100116
        ######## |  100021  100043  100058  100065  100077  100089  100092
                 |  100123
          .#####  |S 100038  100051  100061  100069  100115  100120
        ######### |  100053  100088  100091  100093  100097
              .## M|  100020  100023  100025  100045  100059  100076  100082
                 |  100121
       .########  |  100029  100030  100031  100039  100044  100049  100054
                 |  100068  100072  100073  100090  100095  100096  100100
                 |  100101  100118
       .########  |  100042  100050  100055  100062  100099
 0   .###########  +M 100027  100035  100087
       .########  |  100026  100057  100064  100110  100119
       .########  |  100007  100018  100071  100078  100081  100108  100112
                 |  100113
      .#########  |  100034  100037  100048  100070  100084  100085  100114
                 |  100117
   #############  |  100014  100017  100036  100041  100063  100067  100074
                 |  100080  100122
          .##### S|S 100040
           .####  |  100009  100016  100060  100079
-1       .#######  +  100015  100066
             .##  |
           .####  |T 100047  100075  100083
                .  |  100056
              .#   |  100052
                .  |
              .#   |  100012
-2             . T+
                 |
                 |
                 |
                .|
                 |
                 |
-3               +
                 |
                .|
                 |
                 |
                 |
             <less>|<frequ>
EACH "#" IS 12. EACH "." IS 1 TO 11
```

Figure 13. *Verbos (Expresivo)/Expressive Verbs Item Map*

REPORTED: 2618 PERSON 93 ITEM 186

```
MEASURE    PERSON - MAP - ITEM
              <more>|<rare>
     5          .#  +
                    |  140080
                    |  140081
                    |
                    |
     4            .  +
                  .    140082  140085
                    |
                    T
                  .   |
                  .   |
     3            .  +
                  .   |
                  . T|  140029
                  .   |  140016
                .##   |  140057  140099  140123
                .#    |  140101
     2          .#  +
                .#   S| 140086
              .####   |  140021  140044
              .###    |  140048  140053  140138
             .#####   |  140054  140079  140105  140121  140132
             .##### S |
     1   .##########  +  140014  140065  140106  140127  140133
           .######   |  140020  140049  140068  140073  140077  140084  140087
                     |  140092  140108
         .##########  |  140051  140063  140069  140104  140137
           .########  |  140008  140015  140059  140076
          .#########  |  140093  140096  140117  140124  140139
       .############  |  140022  140097
     0       .#######  +M 140140
        .###########  M|  140103  140118
            .#######   |  140107  140109
            .#######   |  140026  140067
           .########  |  140010  140047  140055  140058  140116
               .####  |  140027
    -1        .######  +
                .###   |  140018  140061  140071
                ####   |  140034  140075  140144
                .## S  |  140045  140050  140119
                .##    |  140136  140141
                ###   S| 140025  140028  140131
    -2          .##   +  140074  140098
                .##    |  140011  140012
                .##    |  140006  140017  140031  140110
                 .#    |  140125
                 .#    |
                  . T|  140009
    -3            .   +  140024
                 .#    |  140032
                  .    |  140134
                 .#   T|
                  .    |  140052
                  .    |
    -4            .   +
                  .    |  140135
                  .    |
                  .    |
                  .    |
    -5    ##########   +
              <less>|<freq>
  EACH "#" IS 14: EACH "." IS 1 TO 13
```

Figure 14. *Identificación de las Letras/ Letter Naming Item Map*

REPORTED: 2295 PERSON 102 ITEM

```
          PERSON - MAP - ITEM
             <more>|<rare>
 4    .#########  +
                  |
                .
               .#
               .#
                .
            .#####
                . T
 3           .##  +
               .#
              .##
                .
             .###
            .####
            .####
               .#
 2       .#######  +
             .### S|  150041
              .##  |
           .#######
           .#######  |T 150038
              .##   |
         #########  |  150058  150084
             .###   |  150027  150059
 1       .#######  +  150033  150078
         .########    150023  150044
        .#########  |S 150019  150051  150061
         .######## M|  150008  150015  150038  150057  150060  150069  150079
                       150081
        .#########  |  150008  150013  150043  150047  150056  150058  150060
                       150072  150082
         .########  |  150009  150012  150018  150021  150029  150039  150045
                       150054  150060  150061  150062  150064  150073  150075
      .###########  |  150010  150012  150031  150055  150063  150070
     .############  |  150048  150077
 0         .####  +M 150015  150016  150016  150035  150040  150042  150047
                       150067
            ######  |  150010  150027  150044  150053  150065  150083
    .#############  |  150029  150037  150068
         .########  |  150005  150006  150036  150041  150046  150049  150056
         .########  |  150017  150028  150043  150050  150052
     .############ S|  150026  150039  150042  150058  150059  150062
             .####  |S
              ####  |  150020  150030  150040  150080  150082
-1       .#######  +  150083
               .#
              .##    |  150011  150045
                .    |  150026  150076
             .###    |T 150070
               .     |  150007
                .
                . T  |  150032
-2              .  +
                     |  150021  150023
               .#
                .
                .
-3              .#  +
             <less>|<frequ>
EACH "#" IS 10. EACH "." IS 1 TO 9
```

Principal component analyses. Principal Components Analysis (PCA) of Rasch residuals (Wright, 1996) is a method used to investigate unidimensionality. PCA of Rasch residuals looks at patterns within the data that are not explained by the Rasch measures (i.e. item difficulties and person abilities). We explored PCAs for all IGDIs-E measures using all items from the current calibration. PCA reports illustrate the observed variance explained in each measure. Resulting statistics offer a breakdown of the raw variance explained in the measure by the children (persons) and the items. This variance is represented in a factor that is identified with an eigenvalue; results are preferable when the Eigenvalue is above 1.8 to 2.0. When the measure has ideal construct coverage and dimensionality the contrast between the Eigenvalues observed by the Rasch model and the next factor, or first, contrast should be large. PCA results are presented in Tables 48-52 for each measure.

Table 48. *PCA results for Verbos (Expresivo)/Expressive Verbs (n=93)*

	Eigenvalue	Observed	Expected
Total raw variance in observations	151.68	100%	100%
Raw variance explained by measures	58.68	38.7%	38.5%
Raw variance explained by persons	27.31	18.0%	17.9%
Raw Variance explained by items	31.37	20.7%	20.6%
Raw unexplained variance (total)	93.00	61.3%	61.5%
Unexplained variance in 1st contrast	1.68	1.1%	1.8%

Results in Table 49 indicate that the ratio of Rasch variance explained by items (20.7%) with the unexplained variance by the first contrast (1.1%) is very large, such that the Rasch dimension is almost 19 times the secondary dimension. The Eigenvalue of the second dimension (1.7) indicates a strength of about two items. These results support the unidimensionality assumption.

Table 49. *PCA results for Identificación de los Dibujos/Picture Naming (n=106)*

	Eigenvalue	Observed	Expected
Total raw variance in observations	173.30	100%	100%
Raw variance explained by measures	67.30	38.8%	38.8%
Raw variance explained by persons	38.55	22.2%	22.2%
Raw Variance explained by items	28.75	16.6%	16.6%
Raw unexplained variance (total)	106.00	61.2%	61.2%
Unexplained variance in 1st contrast	1.85	1.1%	1.7%

Results in Table 50 indicate that the ratio of Rasch variance explained by items (16.6%) with the unexplained variance by the first contrast (1.1%) is very large, such that the Rasch dimension is almost 15 times the secondary dimension. The Eigenvalue of the second dimension (1.85) indicates a strength of about two items. These results support the unidimensionality of the measure.

Table 50. *PCA results for Primeros Sonidos/First Sounds (n=98)*

	Eigenvalue	Observed	Expected
Total raw variance in observations	129.05	100%	100%
Raw variance explained by measures	31.05	24.1%	24.1%
Raw variance explained by persons	23.31	18.1%	18.1%
Raw Variance explained by items	7.74	6.0%	6.0%
Raw unexplained variance (total)	98.00	75.9%	75.9%
Unexplained variance in 1st contrast	1.93	1.5%	2.0%

Results in Table 51 indicate that the ratio of Rasch variance explained by items (6.0%) with the unexplained variance by the first contrast (1.5%) is large, such that the Rasch dimension is approximately 4 times the secondary dimension. The Eigenvalue of the second dimension (1.9) indicates a strength of about two items. These results suggest the second dimension can be neglected and the measure has a unidimensional structure.

Table 51. *PCA results for Identificación de las Letras/ Letter Naming (n=102)*

	Eigenvalue	Observed	Expected
Total raw variance in observations	140.00	100%	100%
Raw variance explained by measures	38.00	27.1%	27.5%
Raw variance explained by persons	25.62	18.3%	18.6%
Raw Variance explained by items	12.37	8.8%	9.0%
Raw unexplained variance (total)	102.00	72.9%	72.5%
Unexplained variance in 1st contrast	1.93	1.4%	1.9 %

Results in Table 52 indicate that the ratio of Rasch variance explained by items (8.8%) with the unexplained variance by the first contrast (1.4%) is large, such that the Rasch dimension is approximately 6times the secondary dimension. The Eigenvalue of the second dimension (1.9) indicates a strength of about two items. These results suggest the second dimension can be neglected and the measure has a unidimensional structure.

Table 52. *PCA results for Identificación de los Sonidos/ Letter Sounds (n=99)*

	Eigenvalue	Observed	Expected
Total raw variance in observations	139.23	100%	100%
Raw variance explained by measures	40.23	28.9%	29.1%
Raw variance explained by persons	31.32	22.5%	22.6%
Raw Variance explained by items	8.91	6.4%	6.4%
Raw unexplained variance (total)	102.00	71.1%	70.9%
Unexplained variance in 1st contrast	1.90	1.4%	1.9 %

Results in Table 52 indicate that the ratio of Rasch variance explained by items (6.4%) with the unexplained variance by the first contrast (1.4%) is large, such that the Rasch dimension is approximately 5 times the secondary dimension. The Eigenvalue of the second dimension (1.9) indicates a strength of about two items. These results suggest the second dimension can be neglected and the measure has a unidimensional structure. See Raîche (2005) for more background.

Construct Representation Analyses

A third approach to evaluating the construct is to examine the degree to which construct irrelevant features contribute to item fit. Assessment of young children can be especially subject to construct-irrelevant components where the interpretations of scores may be affected by the inappropriateness of the assessment in representing a construct.

Construct underrepresentation (failure to fully capture the intended construct) and construct contamination (influence by construct-irrelevant factors) can present threats to score interpretations and use (AERA/APA/NCME, 2014). One way to examine the degree to which a construct is adequately represented by items is through expressive or construct responses to the items. The benefits of constructed-response (CR) item formats include closer resemblance of actual learning tasks, permitting the testing of complex cognitive behavior and allowing for multiple solution strategies; however, they may introduce construct-irrelevance arising from unexpected interpretations (Haladyna & Rodriguez, 2013). The use of CR items must support the validity argument for the construct being measured and how the construct is manifested through the use of CR formats (Haladyna & Rodriguez, 2013; Wilson, 2005).

To examine construct representation in IGDIs-E we examined how one of the CR measures, *Verbos Expresivos/Expressive Verbs*, illustrated the construct of interest based on student responses. Three potential sources of construct underrepresentation and contamination were investigated: *the prompt* to instigate a verbal response, *the stimulus* (or image used to solicit a response), and the *influence of English language on the Spanish construct*. Finally, the relation between of responses and ability level were evaluated.

Claim 1:
The IGDIs-E are psychometrically sound and theory-based, using Mark Wilson's constructing measures model and Rasch modeling for empirical item level statistics.

Claim 3:
The IGDIs-E measures three different yet related domains of early literacy in Spanish alphabet knowledge, oral language and *phonological awareness.*

To evaluate each variable we classified responses on 75 of the *Verbos Expresivos /Expressive Verbs* items based on response features (verbs, functions, descriptions, no response or other) and language (Spanish, English or both). Two researchers independently coded each response using objective guidelines; inconsistencies were reviewed and resolved through consensus. Results indicated that the majority of the responses were in Spanish language (89%) and of a verb-type (86%) suggesting that the target responses are being elicited through way items were designed. Additionally, 88% of the Spanish responses were verbs, whereas 68% of the English responses were verbs, further suggesting that the targeted construct is being addressed in the language of interest. Further analyses on particular item characteristics, such as the saliency of the action depicted in an image, the presence of multiple individuals in the images, and the effects of having more than one correct response suggested that construct underrepresentation/contamination is not necessarily introduced by clarity of the image or clarity of the prompt. Finally, types of responses were explored in relation to children's ability. Results suggested there was no significant association between characteristics of CR responses that related to meaningful patterns in the data.

Claim 13:
IGDIs-E were uniquely designed to attend to how Spanish language develops rather than by translating existing English measures.

Claim 1:
The IGDIs-E are theory-based.

Correlations between IGDIs-E measures. Finally, to examine the associations between measures across the same construct, we estimated correlations between measures. We expected that the oral language measures would correlate more highly together than with other measures of alphabet knowledge or phonological awareness. In reciprocal fashion, we expected that the phonological awareness measure would illustrate higher correlations with the alphabet knowledge measures than the oral language measures. These hypotheses were supported with the results in the correlation matrix, provided in Table 53.

Table 53. *IGDIs-E Correlations Between Measures*

		VE/EV	PS/FS	IL/LN	ID/PN
PS/FS	*r*	.197**			
	p-value	.000			
	n	330			
IL/LN	*r*	.088	.386**		
	p-value	.115	.000		
	n	325	338		
ID/PN	*r*	.642**	.137*	.159**	
	p-value	.000	.014	.005	
	n	337	319	308	
IS/SID	*r*	.063	.461**	.653**	.077
	p-value	.264	.000	.000	.173
	n	321	336	334	311

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

Note. IS/SID: Identificación de los Sonidos /Sound Identification, IL/LN:Identificación de las Letras/ Letter Naming, ID/PN: Identificación de los Dibujos/Picture Naming, VE/EV: Verbos – Expresivo/Expressive Verbs, and PS/FS: Primeros Sonidos/First Sounds

Discriminant criterion-related validity evidence. To examine the degree to which the IGDIs-E measures capture a construct that is markedly different than existing constructs. That is, to what degree do the measures discriminate between two separate constructs? Given the context of IGDIs-E measure design we believe criterion correlations with existing Spanish measures of early language and literacy will likely be low, as many of these measures were developed using a different approach examining a different construct. The construct these measures assess is frequently labeled Spanish early literacy, but more accurately can be portrayed as English early literacy, translated into Spanish. For a sub-sample of our participant pool (*n*=45-64) we correlated with three standardized measures: the Get Ready to Read- Spanish, the Preschool Language Scale-5 Spanish and the Test of Phonological Awareness in Spanish during the 2013-2014 academic year. Results are presented in Tables 54-58.

Table 54. *Descriptive Statistics for the PLS-5 Spanish, TPAS and GRTR-S*

	N	*Min*	*Max*	*M*	*SD*
PLS standard score auditory comprehension subscale	59	50	129	93.58	20.32
PLS standard score expressive comprehension subscale	53	60	135	97.09	16.33
PLS standard score total language subscale	53	51	134	95.45	18.16
TPAS raw score initial sounds	64	5	16	12.13	3.30
TPAS raw score final sounds	64	5	20	13.16	4.94
TPAS raw score rhyming words	64	5	20	13.91	3.34
Get Ready to Read total raw score	55	6	25	15.84	5.00

Table 55. *Correlations between the Get Ready to Read and IGDIs-E Measures*

		VE/EV	PS/FS	IL/LN	ID/PN	IS/SID
GRTR total score	*r*	.25	.24	.28*	-.15	.33*
	p-value	.08	.12	.05	.33	.03
	n	48	45	50	46	45

*Correlation is significant at the .05 level (2-tailed).

Table 56. *Correlations between the PLS-5 and Verbos – Expresivo/Expressive Verbs*

		PLS Auditory Comprehension	PLS Expressive Communication	PLS-5 Total Language Scale
V/VE	*r*	.29*	.12	.23
	p-value	.04	.43	.12
	n	51	47	47

*Correlation is significant at the .05 level (2-tailed).

Table 57. *Correlations between the PLS-5 and Identificación de los Dibujos/Picture Naming*

		PLS Auditory Comprehension	PLS Expressive Communication	PLS-5 Total Language Scale
ID/PN	*r*	.28	.24	.31*
	p-value	.05	.10	.04
	n	49	46	46

*Correlation is significant at the .05 level (2-tailed).

Table 58. *Correlations between the TPAS and Primeros Sonidos/First Sounds, Identificación de los Sonidos /Sound Identification, and Identificación de las Letras/ Letter Naming*

		TPAS Initial Sounds	TPAS final Sounds	TPAS Rhyming Words
PS/FS	*r*	.01	.05	.17
	p-value	.97	.70	.25
	n	49	49	49
IL/LN	*r*	.10	-.16	-.06
	p-value	.51	.30	.71
	n	43	43	43
IS/SID	*r*	-.03	-.01	-.02
	p-value	.86	.98	.89
	n	48	48	48

Note. None of the correlations are significant at the .05 level.

Measurement Invariance

Measurement invariance indicates whether an instrument measures the same trait across populations or occasions (Millsap, 2010). Scores must reflect the construct being measured and not differences due to psychometric misspecification (Millsap, 2010). When the data fit the model, the Rasch model achieves the property of invariance (Rupp & Zumbo, 2006). Underlying this property is the assumption of unidimensionality. To the extent that items tap a unidimensional construct and are not influenced by construct-irrelevant characteristics, it is important to check the stability of item calibrations over time to ensure that positions (rank order of items) and difficulties are not significantly changing.

To examine measurement invariance across seasons of the school year, we explored invariance by calibrating items based on child level fall and winter responses separately. For the purposes of this analysis children with less than 10 valid responses were excluded from the analysis. We hypothesize for these children, item level responses would include significant variance due to error which

may inappropriately influence the analyses. This analysis used data from the Year 2 study, with 464 children assessed during the fall administration and 428 children assessed during the winter administration.

In summary, we calibrated item level statistics for 65-73 items per measure (*Identificación de los Dibujos/Picture Naming* = 71, *Verbos – Expresivo/Expressive Verbs* = 70, *Identificación de las Letras/ Letter Naming* = 73, *Primeros Sonidos/First Sounds* = 73, and *Identificación de los Sonidos /Sound Identification* = 65). Correlations for item positions (rank order) between fall and winter are provided in Table 59. Results indicated, that on average, correlations were high, as expected and positions remained relatively the same. Correlations for item difficulties between fall and winter are provided in Table 60. Results indicated item difficulties remain relatively stable across seasons with respect to other item difficulties within the same calibration.

Table 59. *Correlations between Item Positions for Fall and Winter Calibrations*

Measure	*r*
Identificación de los Dibujos/Picture Naming	.92
Verbos – Expresivo/Expressive Verbs	.88
Identificación de las Letras/ Letter Naming	.88
Primeros Sonidos/First Sounds	.89
Identificación de los Sonidos /Sound Identification	.80

Table 60. *Correlations between Item Difficulties between Fall and Winter Calibrations*

Measure	*r*
Identificación de los Dibujos/Picture Naming	.91
Verbos – Expresivo/Expressive Verbs	.91
Identificación de las Letras/ Letter Naming	.90
Primeros Sonidos/First Sounds	.88
Identificación de los Sonidos /Sound Identification	.87

Score Precision

Another component of validity is the degree to which a score meaningfully and precisely represents student performance. All assessments have error because they are based on samples of behaviors, and irrelevant context conditions randomly introduce variation in item responses and total scores. As a result a person's score is made up of their true score, or their expected score over an infinite number of independent administrations of the same test, plus random error.

The SEM quantifies the precision of a measure. In general, the larger the error the less certain we are about a child's true ability. The IGDIs-E screening forms are designed to be more sensitive (i.e. smaller standard errors) around their respective

cut scores. Reciprocally, the most information is gathered, or precision is accomplished, about a child's ability location from each measure at the cut score value.

To increase our confidence at identifying students who are candidates for Tier 1 intervention in contrast to students who are candidates for Tier 2/3 intervention, a confidence interval of ±1 SE was computed and is noted as Tier M (previously described on page 87). To examine the degree to which the precision of the score is maximized we produced standard error of measurement values for each increment of the 15 item screening set. That is, for each score (0-15) a related SEM was computed. SEM is largest at the tails of the distribution- where students get all items wrong or all items correct because we know the least about modeling their ability because they did not vary in performance on the item set. SEM results are provided in Tables 61-65.

Table 61. *Standard Error of Measurement by Item for Verbos Expresivo/Expressive Verbs*

Raw Score	*Measure*	*SEM*	*Percentile rank*
0	-3.86	1.85	1
1	-2.61	1.04	4
2	-1.83	0.76	8
3	-1.34	0.65	13
4	-0.95	0.59	19
5	-0.63	0.55	27
6	-0.34	0.53	37
7	-0.06	0.52	49
8	0.22	0.52	58
9	0.49	0.53	66
10	0.79	0.55	75
11	1.11	0.59	83
12	1.49	0.65	89
13	1.98	0.76	94
14	2.75	1.04	97
15	4.01	1.85	99

Table 62. *Standard Error of Measurement by Item for Identificación de los Sonidos /Sound Identification*

Score	Measure	SEM	Percentile rank
0	-3.26	1.84	1
1	-2.01	1.04	4
2	-1.24	0.76	8
3	-0.75	0.65	13
4	-0.38	0.58	19
5	-0.06	0.55	27
6	0.23	0.53	37
7	0.51	0.52	49
8	0.77	0.52	58
9	1.05	0.53	66
10	1.33	0.55	75
11	1.65	0.58	83
12	2.03	0.65	89
13	2.52	0.76	94
14	3.28	1.04	97
15	4.54	1.84	99

Table 63. *Standard Error of Measurement by Item for Primeros Sonidos/First Sounds*

Score	Measure	SEM	Percentile rank
0	-3.72	1.84	1
1	-2.46	1.04	4
2	-1.70	0.76	8
3	-1.21	0.65	13
4	-0.83	0.59	19
5	-0.52	0.55	27
6	-0.23	0.53	37
7	0.05	0.52	49
8	0.31	0.52	58
9	0.59	0.53	66
10	0.88	0.55	75
11	1.19	0.58	83
12	1.57	0.65	89
13	2.06	0.76	94
14	2.82	1.04	97
15	4.08	1.84	99

Table 64. *Standard Error of Measurement by Item for Identificación de los Dibujos/Picture Naming*

Score	Measure	SEM	Percentile rank
0	-3.79	1.85	1
1	-2.53	1.04	4
2	-1.76	0.76	8
3	-1.27	0.65	13
4	-0.89	0.59	19
5	-0.57	0.55	27
6	-0.28	0.53	37
7	0.00	0.52	49
8	0.27	0.52	58
9	0.55	0.53	66
10	0.84	0.55	75
11	1.16	0.59	83
12	1.54	0.65	89
13	2.03	0.76	94
14	2.81	1.04	97
15	4.06	1.85	99

Table 65. *Standard Error of Measurement by Item for Identificación de las Letras/ Letter Naming*

Score	Measure	SEM	Percentile rank
0	-3.60	1.85	1
1	-2.34	1.04	4
2	-1.57	0.76	8
3	-1.08	0.65	13
4	-0.70	0.59	19
5	-0.38	0.55	27
6	-0.09	0.53	37
7	0.18	0.52	49
8	0.45	0.52	58
9	0.73	0.53	66
10	1.02	0.55	75
11	1.34	0.59	83
12	1.71	0.65	89
13	2.20	0.76	94
14	2.97	1.04	97
15	4.22	1.85	99

Relation to IGDIs 2.0/IGDIs Literacy

IGDIs-E were designed to complement the existing English IGDIs 2.0 (Literacy +) measures (see the English IGDIs technical manual at www.myigdis.com the technical manual on English measures). Given that IGDIs are used in US classrooms, it is important that the measures be used together to understand Spanish early language and literacy and how it supports English languages and literacy development via IGDIs-E, as well as independently assessing English literacy and language development to evaluate the degree to which the student is successfully navigating English instruction (English IGDIs 2.0/Literacy+).

Claim 9: *IGDIs-E are complementary to the English IGDIs 2.0/ Literacy + and together they can assess the overall language development of preschool SEBs.*

It is important to note that we were not interested in equating the IGDIs measures or directly comparing abilities across IGDIs scales. Specifically, best practice indicates that one assumption that must be met for creating translated or equivalent forms, or equating, to be possible is that the underlying construct must be the same (Albano & Rodriguez, 2012; Kolen & Brennan, 2014). That is, the two tasks in both languages must access the same set of skills and developmental trajectory. It is very clear that given our design model, we believe the Spanish and English constructs are different and unique.

In addition, a technical assumption that supports equating of forms is referred to as equity (see Kolen & Brennan, 2014). A heuristic for understanding this assumption is that following an equating (making two forms of a test exchangeable), it is a matter of indifference to the test-taker which form of the test to take. We find this implausible in the context of bilingual test takers, as given the heterogeneity in language exposure, students may have a preference for testing in one language over the other – thus making equating difficult to defend.

Over the course of the 2014-2015 year, we evaluated the association between Spanish and English early literacy and language performance. Correlations in performance and chi-square analyses are also reported to illustrate common occurrence of base rates by Tier level candidacy in Tables 66-68 (which of course are sample specific).

Table 66. *Correlations between IGDIs-E and English IGDIs Measures for Fall*

	Spanish Measures				
English measures	Identificacion de los Dibujos	Identificacion de las Letras	Identification de los Sonidos	Verbos Expressivo	Primeros Sonidos
Picture Naming	-.05	-.05	.123*	.02	.10
Rhyming	.02	.01	.09	-.07	.08
First Sounds	.00	.08	.10	-.04	.124*
Sound Identification	-.05	.032	.07	.02	.02

As noted, in Table 66, in fall we saw virtually no association between English and Spanish IGDIs performance.

Table 67. *Correlations between IGDIs-E and English IGDIs Measures for Winter*

	Spanish Measures				
English measures	Identificacion de los Dibujos	Identificacion de las Letras	Identification de los Sonidos	Verbos Expressivo	Primeros Sonidos
Picture Naming	.06	.09	.03	-.12*	.05
Rhyming	.05	.11*	.12*	.01	.15**
First Sounds	-.01	.17**	.20**	-.09	.15**
Sound Identification	.01	.03	.19**	-.04	.19**

Table 68. *Correlations between IGDIs-E and English IGDIs Measures for Spring*

	Spanish Measures				
English measures	Identificacion de los Dibujos	Identificacion de las Letras	Identification de los Sonidos	Verbos Expressivo	Primeros Sonidos
Picture Naming	-.12*	.06	.15**	-.01	.09
Rhyming	.08	.18**	.26**	.03	.11*
First Sounds	.10	.17**	.20**	-.02	.17**
Sound Identification	-.01	.12*	.23**	.07	.12*

Throughout winter and spring seasons correlations still remain low, but associations do emerge with the highest correlations between English and Spanish measures occurring in the phonological awareness domain.

Criterion-Related Predictive Validity Evidence

Finally, to examine the degree to which IGDIs-E measure predict performance on meaningful long-term outcomes we completed a longitudinal analysis of student performance where data was available for both their Pre-K year on the IGDIs-E measures (and English IGDIs measures) and in Kindergarten on four established measures of Spanish and English early language and literacy. This study answered the research question "To what degree do IGDIs-E and English IGDIs predict performance on Kindergarten measures of early reading?"

Claim 10: *Performance standards at each tier level were set based on empirical data, expert review panels, and information from parents and teachers, and revised based on longitudinal analysis of children's Kindergarten performance.*

This study included 160 students whom we followed from Pre-K in the 2014-2015 academic year to their Kindergartens in 2015-2016. Students were recruited from four states: MN (15.4%), FL (15.4%), CA (42.9%) and UT (26.3%). The sample include 47.2% females, and 9.7% received special education services. For this group of 160 students, we successfully were able to obtain parent reports on demographics and home language exposure for 156 students.

In the parent report families identified as primarily Latino (64%), Mexican (22%), Caribbean (5%), Central American (6%) and Multiple Ethnicities (3%). 72.4% of children spoke Spanish in the first year of life, and 25% spoke English and Spanish in the first year of life, and 2.6% only spoke English in the first year of life. In contrast to the report about the first year of development, parents reported that 55.2% of children currently use only Spanish at home, 40.9% use both at home and 3.9% use only English at home. Descriptive results are presented in Table 69.

Claim 14: *IGDIs-E scores can be used to inform instructional planning.*

Finally, to gauge socio-economic status we recorded mother's level of education and weekly household income. Results indicated 19.2% of families had less than a 6th grade education level, 27.4% of families had less than a 12th grade education, 4.1% had obtained a GED, 25% had achieved a high school diploma, and 15*% had achieved some education post high school. Beyond high school, 4.1% had achieved a college degree, and 5.5% had achieved a graduate degree.

Table 69. *Descriptive Demographic Information for Students who Received IGDIs in Preschool and Received the TERA, CELF, DIBELS and IDEL in Kindergarten*

Characteristic	%
Female	47.2
Special education services	9.7
Latino (general)	63.0
Mexican	21.4
Caribbean*	5.2
Central American	6.5
Multiple races/ethnicities	3.9
Region	
MN	15.4
FL	15.4
CA	42.9
UT	26.3
Languages spoken to the child from ages 0 to 1	
Spanish	72.4
Both	25.0
English	2.6
Language the child uses when talking at home	
Spanish only	55.2
Both	40.9
English	3.9
Household weekly income**	
Less than $500	63.0
$501 – 700	29.1
$701 – 900	0.0
More than $901	7.9
Mothers highest level of education***	
6th grade or less	19.2
Less than 12th grade	27.4
High school diploma/GED	28.1
Some ed. after high school/vocational program.	15.8
College degree (BA/BS)	4.1
Graduate/ professional degree	5.5

*Cuban, Puerto Rican or Dominican Republic. **N = 98. ***N = 141.

Regarding weekly household income, 63% earned less than $500.00 per week, 29.1% earned between $500.00-700.00 per week, 0% earned $700.00-901.00 per week, and 7.9% earned more than $900.00 per week.

Linear regression was used to evaluate to what degree ability scores on Spanish and English IGDIs predicted performance on the TERA-3, CELF, DIBELS and IDEL subscales. Results are presented in Tables 70-74. In the tables presented here, the slope estimate or coefficient can be interpreted as for every one unit of performance on the IGDIs, the outcome measure (TERA, CELF, etc.) increases by the coefficient unit.

Table 70. *Logistic Regression Results for Expressive Verbs (N = 112)*

	β_0 (*SE*)	β_1 (*SE*)
Spring		
CELF Expressive language usage	0.40 (0.21)	0.72 (0.18)*
Language structure	0.13 (0.19)	0.37 (0.14)*
IDEL Letter naming fluency	-1.03 (0.23)*	0.35 (0.17)*
Winter		
CELF Expressive language usage	0.57 (0.22)	0.74 (0.18)*
Language structure	0.20 (0.20)	0.36 (0.14)*
IDEL Letter naming fluency	-0.98 (0.22)*	0.34 (0.16)*
Fall		
CELF Expressive language usage	0.68 (0.23)*	0.67 (0.15)*
Language structure	0.29 (0.21)	0.34 (0.13)*
IDEL Letter naming fluency	-0.90 (0.22)*	0.25 (0.14)*

* Significant at $p < .05$.

Table 71. *Logistic Regression Results for Spanish Picture Naming (N = 113)*

	β_0 (*SE*)	β_1 (*SE*)
Spring		
CELF Expressive language usage	-0.06 (0.23)	0.68 (0.16)*
Winter		
CELF Expressive language usage	0.10 (0.22)	0.64 (0.16)*
Fall		
CELF Expressive language usage	0.33 (0.22)*	0.77 (0.18)*

* Significant at $p < .05$.

Table 72. *Logistic Regression Results for Sound ID (N = 94)*

	β_0 (*SE*)	β_1 (*SE*)
Spring		
CELF Phonological awareness subscale	-0.19 (0.24)	0.19 (0.12)
IDEL Nonsense word fluency	-1.30 (0.32)*	0.47 (0.15)*
TERA Reading quotient scale	-0.52 (0.27)	0.59 (0.16)*
Alphabet knowledge	-0.31 (0.26)	0.22 (0.13)
DIBELS Letter naming fluency	-1.83 (0.38)*	0.46 (0.16)*
Winter		
CELF Phonological awareness subscale	-0.18 (0.23)	0.24 (0.12)*
IDEL Nonsense word fluency	-1.05 (0.28)*	0.37 (0.14)*
TERA Reading quotient scale	-0.14 (0.23)	0.25 (0.12)*
Alphabet knowledge	-0.08 (0.22)	-0.01 (0.11)
DIBELS Letter naming fluency	-1.52 (0.32)*	0.31 (0.14)*
Fall		
CELF Phonological awareness subscale	0.01 (0.21)	0.13 (0.15)
IDEL Nonsense word fluency	-0.72 (0.23)*	0.31 (0.17)
TERA Reading quotient scale	0.06 (0.21)	0.20 (0.15)
Alphabet knowledge	-0.09 (0.21)	-0.06 (0.14)
DIBELS Letter naming fluency	-1.22 (0.26)	0.24 (0.18)

* Significant at $p < .05$.

Table 73. *Logistic Regression Results for Letter Naming (N = 105)*

	β_0 (*SE*)	β_1 (*SE*)
Spring		
CELF Language content index	-0.19 (0.23)	0.33 (0.15)*
Phonological awareness subscale	0.12 (0.23)	0.18 (0.14)
IDEL Letter naming fluency	-2.57 (0.48)*	1.33 (0.28)*
TERA Reading quotient scale	-0.04 (0.21)	0.23 (0.15)
DIBELS Letter naming fluency	-0.38 (0.25)	0.62 (0.18)*
Winter		
CELF Language content index	0.04 (0.20)	0.04 (0.14)
Phonological awareness subscale	0.09 (0.21)	0.48 (0.16)*
IDEL Letter naming fluency	-1.50 (0.31)*	0.68 (0.20)*
TERA Reading quotient scale	-0.04 (0.21)	0.23 (0.15)
DIBELS Letter naming fluency	-0.17 (0.23)	0.60 (0.18)*
Fall		
CELF Language content index	0.06 (0.20)	-0.01 (0.15)
Phonological awareness subscale	0.26 (0.20)	0.10 (0.15)
IDEL Letter naming fluency	-1.06 (0.24)*	0.64 (0.22)*
TERA Reading quotient scale	0.08 (0.20)	0.14 (0.15)
DIBELS Letter naming fluency	0.13 (0.21)	0.29 (0.17)

* Significant at $p < .05$.

Table 74. *Logistic Regression Results for First Sounds (N = 100)*

	β_0 (*SE*)	β_1 (*SE*)
Spring		
CELF Phonological awareness subscale	-0.33 (0.24)	0.48 (0.16)*
Core language	-0.33 (0.24)	0.43 (0.15)*
TERA Reading quotient scale	-0.33 (0.24)	0.43 (0.15)*
DIBELS Letter naming fluency	-0.22 (0.24)	0.31 (0.15)*
Winter		
CELF Phonological awareness subscale	-0.08 (0.22)	0.25 (0.13)
Core language	0.05 (0.22)	-0.01 (0.13)
TERA Reading quotient scale	0.04 (0.21)	0.37 (0.17)*
DIBELS Letter naming fluency	-0.19 (0.23)	0.34 (0.14)*
Fall		
CELF Phonological awareness subscale	0.08 (0.20)	0.33 (0.17)*
Core language	0.04 (0.20)	0.15 (0.15)
TERA Reading quotient scale	0.04 (0.21)	0.37 (0.17)*
DIBELS Letter naming fluency	0.03 (0.21)	0.28 (0.17)

* Significant at $p < .05$.

Results indicated that the constrained measures, *Primeros Sonidos, Identificacion de los Sonidos, and Identificacion de las Letras* all demonstrated significant associations (regression slopes). These slopes suggest that for every one unit change in the IGDIs-E measures, we can expect the associated coefficient change in the outcome measure. Of particular interest is the predictive power the constrained IGDIs-E measures showed in predicting English performance in Kindergarten.

The unconstrained measures, *Verbos Expressivo* and Identificacion de los Dibujos, did not show the same level of cross-linguistic associations and instead demonstrated predictive power with language, where IGDIs-E scores meaningfully predicted Spanish CELF scores.

REFERENCES

Abdi, H., & Valentin, D. (2007). Multiple correspondence analysis. In N.J. Salkind (Ed.), *Encyclopedia of measurement and statistics* (pp. 651-657). Thousand Oaks, CA: Sage Publications.

Adams, M.J. (1990). *Beginning to read: Thinking and learning about print.* Cambridge, MA: MIT Press.

Al Otaiba, S., Connor, C.M., Folsom, J.S., Greulich, L., Meadows, J., Li, Z. (2011). Assessment data-informed guidance to individualize kindergarten reading instruction: Findings from a cluster-randomized control field trial. *Elementary School Journal, 111*, 535–560.

Albano, A.D., & Rodriguez, M.C. (2012). Statistical equating with measures of oral reading fluency. *Journal of School Psychology, 50*(1), 43-59.

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Anderson, R. (2004). First language loss in Spanish-speaking children. In B.A. Goldstein (Ed.), *Bilingual language development & disorders in Spanish-English speakers* (pp. 187-211). Baltimore, MD: Brooks Publishing.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42*(1), 7-16.

Anthony, J.L. & Lonigan, C.J. (2004). The nature of phonological awareness: Converging evidence from four studies of preschool and early grade school children. *Journal of Educational Psychology*, *96*(1), 43.

Anthony, J.L., Williams, J.M., Duran, L.K., Gillam, S.L., Liang, L., Aghara, R., Swank, P.R., Assel, M.A., & Landry S.H. (2011). Spanish phonological awareness: Dimensionality and sequence of development during the preschool and kindergarten years. *Journal of Educational Psychology, 103*(4), 857-876.

August, D., Carlo, M., Dressler, C., & Snow, C. (2005). The critical role of vocabulary development for English language learners. *Learning Disabilities Research and Practice, 20*(1), 50-57.

Ball, E.W. & Blachman, B.A. (1991). Does phoneme awareness training in kindergarten make a difference in early word recognition and developmental spelling? *Reading Research Quarterly, 26*(1), 49-66.

Barrueco, S., López, M., Ong, C., & Lozano, P. (2012). *Assessing Spanish-English bilingual preschoolers: A guide to best approaches and measures*. Baltimore, MD: Brookes Publishing.

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

Bedore, L.M., Peña, E., García, M., & Cortez, C. (2005). Conceptual versus monolingual scoring: When does it make a difference. *Language, Speech & Hearing Services in Schools*, *36*(3), 188-200.

Bedore, L.M., Peña, E., Griffin, M., & Hixon, J.G. (2016). Effects of age of English exposure, current input/output, and grade on bilingual language performance. *Journal of Child Language, 1*(3), 1-20.

Bedore, L.M., Peña, E.D., Joyner, E.D., Macken, D. (2011). Parent and teacher rating of bilingual proficiency and language development concerns. *International Journal of Bilingual Education and Bilingualism*, *14*(5), 489-511.

Bedore, L.M., Peña, E., Summers, C.L., Boerger, K.M., Resendiz, M.D., Greene, K., Bohman, T.M., & Gillam, R.B. (2012). The measure matters: Language dominance profiles across measures in Spanish–English bilingual children. *Bilingualism: Language and Cognition, 15*, 616-629.

Biemiller, A. (2005). Size and sequence in vocabulary development: Implications for choosing words for primary grade vocabulary instruction. In E.H. Hiebert & M.L. Kamil (Eds.), *Teaching and learning vocabulary: Bringing research to practice* (pp. 223-242). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Blashfield, R.K. (1976). Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. *Psychological Bulletin, 83*(3), 377-388.

Borzone de Manrique, A.M., & Signorini, A. (1994). Phonological awareness, spelling and reading abilities in Spanish-speaking children. *British Journal of Educational Psychology, 64*(3), 429-439.

Bradley, L. & Bryant, P.E. (1983). Categorizing sounds and learning to read – a causal connection. *Nature*, *301*(5899), 419.

Branum-Martin, L., Mehta, P.D., Fletcher, J.M., Carlson, C.D., Ortiz, A., Carlo, M., & Francis, D.J. (2006). Bilingual phonological awareness: Multilevel construct validation among Spanish-speaking kindergarteners in transitional bilingual education classrooms. *Journal of Educational Psychology, 98*, 170–181.

Brown, T.A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: Guildford.

Burchinal, M., Field, S., López, M.L., Howes, C., & Pianta, R. (2012). Instruction in Spanish in pre-kindergarten classrooms and child outcomes for English language learners. *Early Childhood Research Quarterly, 27*(2), 188-197.

Buysse, V. & Peisner-Feinberg, E.S. (Eds.). (2013). *Handbook of response to intervention in early childhood*. Baltimore, MD: Brookes Publishing.

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. *Language, Speech, and Hearing in the Schools*, *38*(3), 249-259.

Carillo, M. (1994). Development of phonological awareness and reading acquisition. *Reading and Writing, 6*(3), 279-298.

Castilla, A.P., Restrepo, M.A., & Perez-Leroux, A.T. (2009). Individual differences and language interdependence: a study of sequential bilingual development in Spanish–English preschool children. *International Journal of Bilingual Education and Bilingualism, 12*(5), 565-580.

Chiu, T.W., & Camilli, G. (2013). Comment on 3PL IRT adjustment for guessing. *Applied Psychological Measurement, 37*(1), 76-86.

Cisero, C.A. & Royer, J.M. (1995). The development and cross-language transfer of phonological awareness. *Contemporary Educational Psychology, 20*(3), 275-303.

Cizek, G.J., & Bunch, M.B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: SAGE.

Cobo-Lewis, A., Eilers, R.E., Pearson, B.Z., & Umbel, V.C. (2002). Interdependence of Spanish and English knowledge in language and literacy among bilingual children. In D.K. Oller & R.E. Eilers (Eds.), *Language and literacy in bilingual children.* Clevedon, United Kingdom: Multilingual Matters.

Coleman, M.R., Roth, F.P., & West, T. (2009). *Roadmap to pre-K RTI: Applying response to intervention in preschool settings*. New York, NY: National Center for Learning Disabilities.

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research*, *49*, 222-251.

Cunnigham, A.E. & Stanovich, K.E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology, 33*(6), 934-945.

DeAnda, S., Bosch, L., Poulin-Dubois, D., Zesiger, P., & Friend, M. (2016). The language exposure assessment tool: Quantifying language exposure in infants and children. *Journal of Speech, Language, and Hearing Research, 59*(6), 1346-1356.

De Houwer, A., & Bornstein, M. (2003, April). *Balancing on the tightrope: Language use patterns in bilingual families with young children*. Paper presented at the International Symposium on Bilingualism, Tempe, AZ.

Deno, S.L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*(3), 219-232.

Deno, S.L. (1997). Whether thou goest. Perspectives on progress monitoring. In. J.W. Lloyd, E.J. Kameenui, & D. Chard (Eds.), *Issues in educating students with disabilities,* (pp. 77-99). Mahwah, NJ: Lawrence Erlbaum.

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

Dickinson, D.K., McCabe, A., Clark-Chiarelli, N., & Wolf, A. (2004). Cross-language transfer of phonological awareness in low-income Spanish and English bilingual preschool children. *Applied Psycholinguistics, 25,* 323-347.

Dickinson, D.K., & Neuman, S.B. (2006). *Handbook of early literacy research: Volume II*. New York, NY: Guilford Press.

Dickinson, D.K., & Tabors, P.O. (2001). *Beginning literacy with language: Young children learning at home and school*. Baltimore, MD: Brookes Publishing.

Donovan, M.S., & Cross, C.T. (Eds.). (2002). *Minority students in special and gifted education.* Washington, DC: National Academies Press.

Downer, J.T., Pianta, R.C., Fan, X., Hamre, B., Mashburn, A., & Justice, L. (2012). Observations of teacher-child interactions in classrooms serving Latinos and dual language learners: Applicability of the classroom assessment scoring system in diverse settings. *Early Childhood Research Quarterly, 27*(1), 21-32.

Durán, L., & Wackerle-Hollman, A. (2016). *Language exposure evaluation report*. Minneapolis, MN: IGDILabs, University of Minnesota.

Durgunoğlu, A.Y., Nagy, W.E., & Hancin-Bhatt, B.J. (1993). Cross-language transfer of phonological awareness. *Journal of Educational Psychology, 85,* 453-465.

du Toit, M. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT user's guide*. Lincolnwood, IL: Scientific Software International.

Escamilla, K. (2000). Teaching literacy in Spanish. In R. DeVillar & J. Tinajero (Eds.), *The power of two languages* (pp. 126-141). New York, NY: McMillan/McGraw-Hill.

Farver, J.A.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*(2), 161-178.

Foy, J.G., & Mann, V. (2006). Changes in letter sound knowledge are associated with development of phonological awareness in pre-school children. *Journal of Research in Reading Banner, 29*(2), 143-161.

Fry, R., & Gonzales, F. (2008). *One in five and growing fast: A profile of Hispanic public school students*. Washington, DC: Pew Hispanic Center. Retrieved from http://www.pewhispanic.org/2008/08/26/one-in-five-and-growing-fast-a-profile-of-hispanic-public-school-students/

Fuchs, L.S., & Deno, S.L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children*, *57*(6), 488-99.

Fuchs, D., & Fuchs, L. (2006). Introduction to response to intervention: What, why, and how valid is it? *Reading Research Quarterly, 41*(1), 93-99.

Garcia, E. & Jensen, B. (2009). Early educational opportunities for children of Hispanic origins. *Social Policy Report*, *23*(2), 3-19. Retrieved from http://www.srcd.org/sites/default/files/documents/23-2_garcia.pdf

Garcia, E.E. & Miller, L.S. (2008). Findings and recommendations of the National Task Force on Early Childhood Education for Hispanics. *Child Development Perspectives, 2*(2), 53-58.

Gillam R.B., Peña E.D., Bedore L.M., Bohman, T.M., & Mendez-Perez, A. (2013). Identification of specific language impairment in bilingual children: I. Assessment in English. *Journal of Speech, Language, and Hearing Research*, *56*(6), 1813-1823.

Goldenberg, C. (2008). Improving achievement for English language learners. In S.B. Neuman (Ed.), *Educating the other America* (pp.139-162). Baltimore, MD: Brookes Publishing.

Goodrich, J.M., Lonigan, C.J., & Farver, J.M. (2013). Do early literacy skills in children's first language promote development of skills in their second language? An experimental evaluation of transfer. *Journal of Educational Psychology*, *105*(2), 414–426.

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

Greenwood, C.R., Bradfield, T., Kaminski, R., Linas, M., Carta, J.J., & Nylander, D. (2011). The response to intervention (RTI) approach in early childhood. *Focus on Exceptional Children, 43*(9), 1-22.

Greenwood, C., Carta, J., McConnell, S., Goldstein, H., & Kaminski, R. (2008). Center for Response to Intervention in Early Childhood. Kansas City, KS: University of Kansas. Retrieved from http://www.crtiec.dept.ku.edu/

Grigg, W.S., Daane, M.C., Jin, Y., & Campbell, J.R. (2003). *The nation's report card: Reading, 2002*. Washington, DC: U.S. Department of Education.

Gutierrez, K.D., Zepeda, M., & Castro, D.C. (2010). Advancing early literacy learning for all children: Implications of the NELP report for dual language learners. *Educational Researcher, 39*(4), 334-339.

Haladyna, T.M., & Rodriguez, M.C. (2013). *Developing and validating test items*. New York, NY: Routledge.

Hammer, C.S., Lawrence, F.R., & Miccio, A.W. (2007). Bilingual children's language abilities and early reading outcomes in Head Start and Kindergarten. *Language, Speech, and Hearing Services in Schools*, 38, 237-248.

Hammer, C.S., Miccio, A.W., & Rodríguez, B. (2004). Bilingual language acquisition and the child socialization process. In B. Goldstein (Ed.), *Bilingual language development and disorders in Spanish–English speakers*. Baltimore, MD: Brookes Publishing.

Hamre, B.K., & Pianta, R.C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development, 76*(5), 949-967.

Hansen, P., & Jaumard, B. (1997). Cluster analysis and mathematical programming. *Mathematical Programming, 79*(1-3), 191-215.

Hindman, A.H., & Wasik, B.A. (2015). Building vocabulary in two languages: An examination of Spanish-speaking dual language learners in Head Start. *Early Childhood Research Quarterly, 31*, 19-33.

Hoff, E. (2010). Context effects on young children's language use: The influence of conversational setting and partner. *First Language, 30*, 461-472.

Husson, F., & Josse, J. (2014). Multiple correspondence analysis. In J. Blasius & M. Greenacre (Eds.), *Visualization and the verbalization of data* (pp. 166-183). New York, NY: CRC Press.

Husson, F., Josse, J., Le, S., & Mazet, J. (2015). FactoMineR: Multivariate exploratory data analysis and data mining. R package ver. 1.29. Retrieved from https://CRAN.R-project.org/package=FactoMineR

Jackson, S., Pretti-Frontczak, K., Harjusola-Webb, S., Grisham-Brown, J., & Romani, J.M. (2009). Response to intervention: Implications for early childhood professionals. *Language, Speech, and Hearing Services*, *40*(4), 424-434.

Jackson-Maldonado, D., Thal, D., Marchman, V.A., Newton, T., Fenson, L., & Conboy, B.T. (2003). *MacArthur inventarios del desarrollo de habilidades comunicativas (inventarios): User's guide and technical manual*. Baltimore, MD: Brookes Publishing.

Jiménez González, J.E., & Garcia, C.R.H. (1995). Effects of word linguistic properties on phonological awareness in Spanish children. *Journal of Educational Psychology, 87*(2), 193-201.

Juel, C. (2006). The impact of early school experiences on initial reading. In D.K. Dickinson & S.B. Neuman (Eds.), *Handbook of early literacy research* (Vol. 2, pp. 410-426). New York, NY: Guilford.

Justice, J., Mashburn, A., Hamre, B., & Pianta, R. (2008). Quality of language and literacy instruction in preschool classrooms serving at-risk pupils. *Early Childhood Research Quarterly*, *23*(1), 51-68.

Kane, M. (2013). Validation as a pragmatic, scientific activity. *Journal of Educational Measurement*, *50*(1), 115-122.

Kaufmann, L., & Rousseeuw, P.J. (2009). *Finding groups in data: An introduction to cluster analysis*. Hoboken, NJ: John Wiley & Sons.

Kolen, M.J., & Brennan, R.L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York, NY: Springer.

Kuo, L., & Anderson, R.C. (2010). Beyond cross-language transfer: Reconceptualizing the impact of early bilingualism on phonological awareness. *Scientific Studies of Reading*, *14*(4), 365-385.

Lieven, E.V. & Tomasello, M. (2008). Children's first language acquisition from a usage-based perspective: In P. Robinson & N. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 178-206). New York, NY: Routlede.

Limbos, M. M. & Geva, E. (2001). Accuracy of teacher assessments of second-language students at risk for reading disability. *Journal of Learning Disabilities, 34*(2), 136-151.

Linacre, J.M. (2011). Winsteps (Version 3.72.0) [Computer Software]. Beaverton, OR: Winsteps.com.

Linacre, J.M. (2016). *Winsteps® Rasch measurement computer program user's guide*. Beaverton, OR: Winsteps.com. Retrieved from http://www.winsteps.com/

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.

Lonigan, C.J. (2009). *Development of a comprehensive assessment system for Spanish-speaking English learner's early literacy skills.* IES Goal 5 Early Learning Program and Policies grant. R305A090169. Retrieved from https://ies.ed.gov/funding/grantsearch/details.asp?ID=744

Lonigan, C.J., Allan, N.P., & Lerner, M.D. (2011). Assessment of preschool early literacy skills: Linking children's educational needs with empirically supported instructional activities. *Psychology in the Schools*, *48*(5), 488-501.

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.

Lonigan, C.J., Farver, J.M., Nakamoto, J., & Eppe, S. (2013). Developmental trajectories of preschool early literacy skills: A comparison of language-minority and monolingual-English children. *Developmental Psychology*, *49*(10), 1943.

Lonigan, C.J., Wagner, R.K., Torgesen, J.K., & Rashotte, C.A. (2007). *Test of Preschool Early Literacy (TOPEL)*. Austin, TX: Pro-Ed. Retrieved from https://www.proedinc.com/Products/12440/topel-test-of-preschool-early-literacy.aspx

Lopez, L.M., & Greenfield, D.B. (2004). The cross-language transfer of phonological skills of Hispanic Head Start children. *Bilingual Research Journal, 28*(1), 1-18.

Lyon, J. (1996). *Becoming bilingual: Language acquisition in a bilingual community*. Philadelphia, PA: Multilingual Matters.

MacWhinney, B. (2008). A unified model of language acquisition. In N. Ellis & P. Robinson (Eds.), *Handbook of cognitive linguistics and second language acquisition*. New York, NY: Lawrence Erlbaum.

Mancilla-Martinez, J., & Lesaux, N.K. (2010). Predictors of reading comprehension for struggling readers: The case of Spanish-speaking language minority learners. *Journal of Educational Psychology, 102*(3), 701-711.

Mancilla-Martinez, J., Gámez, P.B., Vagh, S.B., & Lesaux, N.K. (2016). Parent reports of young Spanish-english bilingual children's productive vocabulary: A development and validation study. *Language, Speech, and Hearing Services in Schools, 47*(1), 1-15.

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*(4), 214-224.

Mashburn, A.J., Pianta, R.C., Hamre, B.K., Downer, J.T., Barbarin, O.A., Bryant, D., Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child Development, 79*(3), 732-749.

McBride-Chang, C. (1999). The ABCs of the ABCs: The development of letter-name and letter-sound knowledge. *Merrill-Palmer Quarterly: Journal of Developmental Psychology*, *45*(2), 285-308.

McConnell, S.R. & Greenwood, C.R. (2013). General outcome measures in early childhood and the Individual Growth and Development Indicators. In V. Buysse & E. Peisner-Feinberg (Eds.), *Handbook of response to intervention in early childhood* (pp.143-154). Baltimore, MD: Brookes Publishing.

McConnell, S.R., & Missall, K.N. (2008). Best practices in monitoring progress for preschool children. In A. Thomas & J. Grimes (Eds.), *Best practices in school*

psychology (5th ed., pp. 561-573). Washington, DC: National Association of School Psychologists.

McMaster, K., & Espin, C. (2007). Technical features of curriculum-based measurement in writing: A literature review. *Journal of Special Education, 41*(2), 68-84.

Melby-Lervåg, M., & Lervåg, A. (2011). Cross-linguistic transfer of oral language, decoding, phonological awareness and reading comprehension: A meta-analysis of the correlational evidence. *Journal of Research in Reading*, *34*(1), 114-135.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*(11), 1012-1027.

Miller, J.F., Heilmann, J., Nockerts, A., Iglesias, A., Fabiano, L., & Francis, D.J. (2006). Oral language and reading in bilingual children. *Learning Disabilities Research and Practice, 21*(1), 30-43.

Millsap, R.E. (2010). Testing measurement invariance using item response theory in longitudinal data: An introduction. *Child Development Perspectives, 4*(1), 5-9.

Mitzel, H.C., Lewis, D.M., Patz, R.J., & Green, D.R. (2001). The bookmark procedure: Psychological perspectives. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249–281). Mahwah, NJ: Lawrence Erlbaum.

Muthén, L.K., & Muthén, B.O. (2012a). Mplus. (Version 7). [Computer software]. Los Angeles, CA: Authors.

Muthén, L.K., & Muthén, B.O. (2012b). *Mplus user's guide* (7th ed.). Los Angeles, CA: Authors.

NELP: National Early Literacy Panel. (2008). *Developing early literacy: Report of the National Early Literacy Panel - A scientific synthesis of early literacy development and implications for intervention*. Jessup, MD: National Institute for Literacy.

National Reading Panel (2000). *Teaching Children to read: Report of the National Reading Panel – An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Bethesda, MD: National Institutes of Health.

O'Grady, W., Lee, M., & Kwak, H.Y. (2009). Emergentism and second language acquisition. In W.C. Ritchie & T.K. Bhatia (Eds.), *The new handbook of second language acquisition* (pp. 69-88). Bingley, United Kingdom: Emerald Group Publishing.

Oller, D.K., & Eilers, R.E. (Eds.). (2002). *Language and literacy in bilingual children.* Clevedon, United Kingdom: Multilingual Matters.

Páez, M.M., Tabors, O.P., & López, L.M. (2007). Dual language and literacy development of Spanish-speaking preschool children. *Journal of Applied Developmental Psychology, 28*(2), 85-102.

Paradis, J. (2010). The interface between bilingual development and specific language impairment. *Applied Psycholinguistics*, *31*(2), 227-252.

Peña, E.D. (2007). Lost in translation: Methodological considerations in cross-cultural research. *Child Development*, *78*(4), 1255-1264.

Peña, E., Bedore, L.M., & Rappazzo, C. (2003). Comparison of Spanish, English, and bilingual children's performance across semantic tasks. *Language, Speech, and Hearing Services in the Schools, 34*, 5-16.

Peña, E.D., Gutierrez-Clellen, V.F., Iglasias, A., Goldstein, B.A., & Bedore, L. (2014). *Bilingual Input-Output Survey (BIOS)*. Baltimore, MD: Brookes Publishing.

Peña, E.D., & Halle, T.G. (2011). Assessing preschool dual language learners: Traveling a multiforked road. *Child Development Perspectives, 5*(1), 28-32.

Peña, E.D., Kester, E.S., & Sheng, L., (2012). Semantic development in Spanish–English bilinguals: Theory, assessment, and intervention. In: B.A. Goldstein (Ed.), *Bilingual language development and disorders in Spanish–English speakers* (pp. 131-149). Baltimore, MD: Brookes Publishing.

Pettito, L. A., Katerlos, M., Levy, B. G., Guana, K., Tetretault, K., & Ferraro, V. (2001). Bilingual signed and spoken language acquisition from birth: implications for the mechanisms underlying early bilingual language acquisition. *Journal of Child Language, 28*(2), 453-496.

Phillips, B.M., Clancy-Menchetti, J., & Lonigan, C.J. (2008). Successful phonological awareness instruction with preschool children: Lessons from the classroom. *Topics in Early Childhood Special Education, 28*(1), 3-17.

Polikoff, M.S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice, 29*(4), 3-14.

Priest, J.S., McConnell, S.R., Walker, D., Carta, J.J., Kaminski, R.A., McEvoy, M.A., Good, R.H., Greenwood, C.R., & Shinn, M.R. (2001). General growth outcomes for young children: Developing a foundation for continuous progress measurement. *Journal of Early Intervention, 24*, 163–180.

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*(1), 159-169.

Quiroz, B.G., Snow, C.E., & Zhao, J. (2010). Vocabulary skills of Spanish—English bilinguals: Impact of mother—child language interactions and home language literacy support. *International Journal of Bilingualism, 14*(4), 379-399.

Raîche, G. (2005). Critical eigenvalue sizes (variances) in standardized residual principal components analysis. *Rasch Measurement Transactions, 19.1*, 1012. Retrieved from. http://www.rasch.org/rmt/rmt191h.htm

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. (expanded edition, originally published 1966, Copenhagen, Danish Institute for Educational Research). Chicago, IL: The University of Chicago Press.

Raynolds, L.B., & Uhry, J.K. (2010). The invented spellings of non-Spanish phonemes by Spanish–English bilingual and English monolingual kindergarteners. *Reading and Writing*, *23*(5), 495-513.

Reynolds, A.J., & Temple, J.A. (1998). Extended early childhood intervention and school achievement: Age thirteen findings from the Chicago longitudinal study. *Child Development, 69*(1), 231-246.

Rodriguez, M.C. (2011). *Technical guidance report #4: The Rasch model*. Minneapolis, MN: Center for Response to Intervention in Early Childhood, University of Minnesota.

Rolstad, K., Mahoney, K., & Glass, G.V. (2005). The big picture: A meta-analysis of program effectiveness research on English language learners. *Educational Policy*, *19*(4), 572-594.

Ruiz-Felter, R., Cooperson, S.J., Bedore, L.M., & Peña, E.D. (2016). Influence of current input-output and age of first exposure on phonological acquisition in early bilingual Spanish-English-speaking kindergarteners. *International Journal of Language and Communication Disorders, 51*(4), 368-383.

Rupp, A.A., & Zumbo, B.D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement, 66*(1), 63-84.

Scheele, A.F., Leseman, P.P.M., & Mayo, A.Y. (2010). The home language environment of monolingual and bilingual children and their language proficiency. *Applied Psycholinguistics, 31*(1), 117-140.

Schweinhart, L.J., DeBruin-Parecki, A., & Robin, K.B. (2004). *Preschool assessment: A guide to developing a balanced approach*. New Brunswick, NJ: Nation Institute for Early Education Research.

Shinn, M. (Ed.). (1998). *Curriculum-based measurement: Assessing special children.* New York, NY: Guilford Press.

Sick, J. (2010). Rasch measurement in language education (part 5): Assumptions and requirements of Rasch measurement. *JALT Testing and Evaluation SIG Newsletter, 14*(2), 23-29.

Signorini, A. (1997). Word reading in Spanish: A comparison between skilled and less skilled beginning readers. *Applied Psycholinguistics, 18*(3), 319-344.

Slavin, R.E., & Cheung, A.C. (2005). A synthesis of research on language of reading instruction for English language learners. *Review of Educational Research, 75*(2), 247-284.

Slavin, R. & Madden, N.A. (2011). Measures inherent to treatments in program effectiveness reviews. *Journal of Research on Educational Effectiveness, 4*(4), 370-380.

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

Stoner, G., Carey, S.P., Ikeda, M.J., & Shinn, M.R. (1994). The utility of curriculum-based measurement for evaluating the effects of methylphenidate on academic performance. *Journal of Applied Behavior Analysis, 27*(1), 101-113.

Tomasello, M. (2005). *Constructing a language: A usage-based theory of language acquisition.* Cambridge, MA: Harvard University Press.

Torgeson, J.T., & Mathes, P.J. (2000). *A basic guide to understanding, assessing, and teaching phonological awareness.* Austin, TX: Pro Ed.

National Center for Education Statistics. (2003). *Status and trends in the education of Hispanics* (NCES 2003–008). Washington, DC: U.S Department of Education. Retrieved from http://nces.ed.gov/pubs2003/2003008.pdf

Vaughn, S., Linan-Thompson S., Pollard-Durodola, S.D., Mathes, P.G., & Cardenas-Hagan, E. (2006). Effective interventions for English language learners (Spanish-English) at risk for reading difficulties. In D.K. Dickinson & S.B. Neuman (Eds.), *Handbook of early literacy research* (Vol. 2, pp. 185-197). Austin, TX: PRO-ED.

Vaughn, S., & Fuchs, L.S. (2003). Redefining learning disabilities as inadequate response to instruction: The promise and potential problems. *Learning Disabilities Research and Practice, 18*(3), 137-146.

Wayman, M.M., Wallace, T., Wiley, HI., Tichá, R., & Espin, CA. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education, 41*(2), 85-120.

Whitehurst, G.J. & Lonigan, C.J. (2001). Emergent literacy: Development from prereaders to readers. In S.B. Neuman & D.K. Dickinson (Eds.), *Handbook of early literacy research* (pp. 11-29). New York, NY: Guilford Press.

Wickham, H. & Francois, R. (2015). dplyr: A grammar of data manipulation. R Package Version 0.4.3. Retrieved from http://CRAN.R-project.org/package=dplyr

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wright, B.D. (1996) Local dependency, correlations and principal components. *Rasch Measurement Transactions, 10*(3), 509-511.

Zehler, A., Fleischman, H., Hopstock, P., Stephenson, T., Pendzick, M., & Sapru, S. (2003). *Descriptive study of services to LEP students and LEP students with disabilities: Policy report: Summary of findings related to LEP and SPED-LEP students*. Washington, DC: Development Associates.

Zimmerman, I.L., Steiner, V.G., & Pond, R.E. (2014). *The Preschool Language Scales-5 Home Communication Questionnaire*. San Antonio, TX: Pearson Education.

Zimowski, M.F., Muraki, E., Mislevy, R.J., & Bock, R.D. (1996). BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items [Computer software]. Chicago, IL: Scientific Software International.

Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement*. Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/RR-12-08.pdf

.

Appendix A. Language Exposure Evaluation Research

Language Exposure Evaluation Report (LEER)

About Your Child

1. What is your relationship to your child?

- ☐ Mother
- ☐ Father
- ☐ Grandparent
- ☐ Other relative
- ☐ Foster parent
- ☐ Other – *Please describe:* _____

2. From the ages of 0 to 1 year, was there, English, Spanish or both spoken to your child at home?

- ☐ English
- ☐ Spanish
- ☐ Both

3. Does your child know any **other language** in addition to Spanish and English?

- ☐ Yes → *Please specify this other language here:* _____
- ☐ No

4. How would you describe your child's nationality? *Please check all that apply:*

- ☐ Cuban
- ☐ Mexican
- ☐ Puerto Rican
- ☐ Central American Country: _____
- ☐ South American Country: _____
- ☐ Dominican
- ☐ Other → *Please describe:* _____

Current Language Use

We are interested in how much English and Spanish your child hears and speaks. First, think about weekdays (Monday-Friday) and then think about weekends (Saturday-Sunday).

Monday-Friday What languages does your child HEAR?

Morning Routine (awake to 9)	*Early Afternoon (9 to 1)*	*Mid Afternoon (1 to 4)*	*Evening (4 to bedtime)*
☐ Spanish ☐ English ☐ Both	☐ Spanish ☐ English ☐ Both	☐ Spanish ☐ English ☐ Both	☐ Spanish ☐ English ☐ Both

Saturday and Sunday What languages does your child HEAR?

Morning Routine (awake to 9)	Early Afternoon (9 to 1)	Mid Afternoon (1 to 4)	Evening (4 to bedtime)
☐ Spanish ☐ English ☐ Both	☐ Spanish ☐ English ☐ Both	☐ Spanish ☐ English ☐ Both	☐ Spanish ☐ English ☐ Both

Monday-Friday What languages does your child SPEAK?

Morning Routine (awake to 9)	Early Afternoon (9 to 1)	Mid Afternoon (1 to 4)	Evening (4 to bedtime)
☐ Spanish ☐ English ☐ Both ☐ Other _____	☐ Spanish ☐ English ☐ Both	☐ Spanish ☐ English ☐ Both	☐ Spanish ☐ English ☐ Both

Saturday and Sunday What languages does your child SPEAK?

Morning Routine (awake to 9)	Early Afternoon (9 to 1)	Mid Afternoon (1 to 4)	Evening (4 to bedtime)
☐ Spanish ☐ English ☐ Both	☐ Spanish ☐ English ☐ Both	☐ Spanish ☐ English ☐ Both	☐ Spanish ☐ English ☐ Both

More than 12 boxes marked Spanish-only = Spanish Dominant
More than 12 boxes marked Both = Balanced Bilingual
More than 12 boxes marked English = English Dominant

Appendix B.
Item Revisions and Removal Tables

Table B1. *Identificación de los Dibujos/Picture Naming*

Measure	Item ID	Original Target	Description of Revision or Removal Criteria			
			Cut: Low PBSC or In/Out Fit	Cut: DIF	Cut: Other (Description)	After Year:
Identificación de los Dibujos/ Picture Naming	110012	Bus/Autobús/Camión			Mostly English responses	
	110016	Arbol (tree)		✓		
	110028	Maleta/Equipaje (suitcase)		✓		2
	110029	Tomate (tomato)		✓		3
	110035	Abrigo/Chaqueta	✓			2
	110036	Arroz (rice)		✓		2
	110039	Pavo/guajolote	✓			2
	110044	Pelota		✓		3
	110049	Tigre		✓		3
	110052	Escoba		✓		2
	110057	Bate (de beisbol)	✓			2
	110059	Limon	✓			2
	110060	Pies (feet)		✓		3
	110067	Maracas	✓			2
	110078	Tazón/plato hondo	✓			3
	110083	Hormiga (ant)	✓			2
	110084	Chile		✓		2
	110087	Tobillo (ankle)	✓			3
	110089	(banco, escano) bench	✓			3
	110103	jugo (de naranja) orange juice	✓			3
	110105	lima(s), limon(es)	✓			3
	110112	Olla, puchero, marmite (pot)	✓			3
	110117	Cascara, concha, caparazon (shell)	✓			3
	110121	Calcetin(es) (socks)	✓			3

Table B2. *Verbos – Expresivo/Expressive Verbs*

Measure	Item ID	Original Target	Revised ID & New Target (if applicable)	Description of Revision or Removal Criteria			
				Cut: Low PBSC or In/Out Fit, high or low p-value	Cut: DIF	Cut: Other (Description)	After Year:
Verbos – Expresivo/ Expressive Verbs	140008	Tocar la guitarra (to play the guitar)			✓		2
	140016	Esquiar (to ski)		✓			2
	140017	Pintar (to paint)		✓			2
	140099	Batear (to bat)		✓			2
	140023	Abrir (to open)	140107: added new image in Year 2	✓			3
	140030	pasear el perro, andar con el perro (to walk the dog)	140041	✓			2
	140032	caer(se) (to fall)		✓			3
	140033	abrochar(se), (to tie)	amarrar(se)	✓			2
	140034	Volar (to fly)	140042	✓			2
	140101	amontonar, juntar las hojas, rastrillar (to rake)		✓			3
	140031	Cocinar (to cook)				Duplicate	2
	140032	Dar (to give)		✓			2
	140037	Cavar (to dig)		✓			2
	140038	Coser (to sew)	140114; new image in Year 2	✓			3
	140039	acariciar el perro (to pet)		✓			2
	140040	tomar fotos (to take a photo or picture)		✓			2
	140043	Mirar (to look)		✓			3
	140047	hacer compras, comprar (to shop)			✓		2
	140048	deslizarse, resbalar (to slide)			✓		3
	140050	Peinarse (to comb)			✓		3

Measure	Item ID	Original Target	Revised ID & New Target (if applicable)	Description of Revision or Removal Criteria			
				Cut: Low PBSC or In/Out Fit, high or low p-value	Cut: DIF	Cut: Other (Description)	After Year:
	140054	llevar (to carry a bag)				Ambiguous image	2
	140056	gatear (to crawl)		✓			3
	140058	servir, llenar el vaso		✓			2
	140060	rasurar, afeitar(se)		✓			3
	140070	Balancear (to balance)		✓			3
	140074	romper, quebrar (to break)			✓		3
	140076	subir/cerrar (la cremallera/cierre/ziper)			✓		3
	140079	cubrir, cobijar (to cover)			✓		2
	140082	to saw (serrar, serruchar)		✓			3
	140083	zambullir(se) - to dive		✓			3
	140084	aspirar (to vacuum)			✓		3
	140086	girar, darse vueltas			✓		3
	140088	to slide				Duplicate	
	140089	to swing				Duplicate	
	140090	to wake up				Duplicate	
	140091	buscar (to look for)		✓			3
	140094	dejar caer, soltar la pelota (to drop)		✓			3
	140095	to curl hair (rizar, enrizar)		✓			3
	140099	to bat (batear)				Duplicate	2
	140143	buscar (to look for)		✓			3
	140111	tejer (to weave)		✓			3
	140112	guiñar (to wink)		✓			3
	140113	sembrar (to plant a seed)		✓			3
	140115	marchar (to march)		✓			3
	140122	agitar,sacudir, temblar (to shake)		✓			3

Measure	Item ID	Original Target	Revised ID & New Target (if applicable)	Description of Revision or Removal Criteria			
				Cut: Low PBSC or In/Out Fit, high or low p-value	Cut: DIF	Cut: Other (Description)	After Year:
	140123	to growl or snarl (rugir, gruñir)		✓			3
	140126	encintar o pegar (to tape)		✓			3
	140128	to squirt (chorrear, rociar,echar un chorro)		✓			3
	140129	to sneeze (estornudar)		✓			3
	140130	to spill (derramar)		✓			3
	140141	to dance (bailar)		✓			3

Table B3. *Identificación de las Letras/ Letter Naming*

				Description of Revision or Removal Criteria			
Measure	Item ID	Original Target	Revised ID & New Target (if applicable)	Cut: Low PBSC or In/Out Fit, high or low p-value	Cut: DIF	Cut: Other (Description)	After Year:
Identificación de las Letras/ Letter Naming	150006	P			✓		3
	150008	E			✓		3
	150012	f		✓			2
	150013	j		✓			2
	150028	i	/e/ in English is same as /i/ in Spanish	✓			2
	150031	J		✓			2
	150033	i		✓			2
	150037	H		✓			2
	150046	P			✓		3
	150047	I	Visually similar letters		✓		3
	150055	y			✓		3
Identificación de las Letras/ Letter Naming	150059	z			✓		3
	150060	G			✓		3
	150062	O			✓		3
	150066	Q	potentially visually similar to foils and similar sound	✓			2
	150068	y	visually similar foils	✓			2
	150074	N	Foil is a lowercase L	✓			2
	150085	N	all letter names have two syllables	✓			2

				Description of Revision or Removal Criteria			
Measure	Item ID	Original Target	Revised ID & New Target (if applicable)	Cut: Low PBSC or In/Out Fit, high or low p-value	Cut: DIF	Cut: Other (Description)	After Year:
	150087	q		✓			2
	150092	Z		✓			2
	150097	e	Foils too similar	✓			3
	150102	g	visually similar foils both have "tails" that hang below the line	✓			
	150105	d	Visually similar foils	✓			

Table B4. *Identificación de los Sonidos /Sound Identification*

				Description of Revision or Removal Criteria			
Measure	Item ID	Original Target/item	Revised ID & New Target (if applicable)	Cut: Low PBSC or In/Out Fit, high or low p-value	Cut: DIF	Cut: Other (Description)	After Year:
Identificación de los Sonidos /Sound Identification	160008	Ñ		✓			2
	160020	w f i	Foil /i/ sound similarity to English e	✓			2
	160021	i		✓			2
	160025	z				No consensus on how to pronounce this sound (/s/, or /ks/)	2
	160028	ñ			✓		3
	160029	g		✓			2
	160030	j		✓			2
	160035	A			✓		3
	160040	W – foil M visually similar to w				/w/ is not a sound in Spanish - all words in Spanish that start with /w/ are English-influenced	2
	160045	V	One of the foils is a capital I, not a lowercase L		✓		3
	160047	Q		✓			2
	160049	ll		✓			2
	160051	x				No consensus on how to pronounce	2

				Description of Revision or Removal Criteria			
Measure	Item ID	Original Target/item	Revised ID & New Target (if applicable)	Cut: Low PBSC or In/Out Fit, high or low p-value	Cut: DIF	Cut: Other (Description)	After Year:
						this sound (/s/, or /ks/	
	160052	j		✓			2
	160053	y		✓			2
	160059	M	Foil visually similar and similar sound			Poor fit	2
	160066	q	Foil /P/visually similar ("q" flipped over)	✓			2
	160067	V			✓		3
	160069	C		✓			2
	160073	J	Foil potentially similar sound if child thinks of /j/ in English		✓		3
	160078	X				No consensus on how to pronounce this sound (/s/, or /ks/	2
	160084	Z				No consensus on how to pronounce this sound (/s/) and poor fit	3
	160085	q		✓			3
	160086	D		✓			3
	160099	Q		✓			3
	160102	j		✓			3
	160107	e		✓			3

				Description of Revision or Removal Criteria			
Measure	Item ID	Original Target/item	Revised ID & New Target (if applicable)	Cut: Low PBSC or In/Out Fit, high or low p-value	Cut: DIF	Cut: Other (Description)	After Year:
	160108	t		✓			3
	160114	r		✓			3

Table B5. *Primeros Sonidos/First Sounds*

				Description of Revision or Removal Criteria			
Measure	Item ID	Original Target/item	Revised ID & New Target (if applicable)	Cut: Low PBSC or In/Out Fit, high or low p-value	Cut: DIF	Cut: Other (Description)	After Year:
Primeros Sonidos/First Sounds	100002	Lapiz	100107 Changed target to be a syllable rather than a phoneme (/ga/ for gato instead of /l/ for lapiz).				
	100004	Cama	100106 Changed target to be a syllable rather than a phoneme (/ca/ for camisa instead of /a/ for abuela).	✓			2
	100019	Elefante/Abuela		✓			2
	100022	Abrigo/tomate	All 3 foils are syllable words	✓			2
	100031	Conejo/vestido	feature - "v" and "p" sound very similar	✓			2
	100034	Pato/Cama/Taza	Features pato and taza are both brown	✓			2
	100046	Mantequilla/almohada/vaca	Pillow image looks like butter	✓			2
	100052	chile/oreja/zapato			✓		3
	100055	Cereza/bandera/quitarra			✓		3
	100065	Venado/hoja/pelota/			✓		3
	100069	Oreja/mano/gorro2		✓			2

100076	Martillo/avion/clavos			✓		3
100078	Caballo/montaña/leche			✓		3
100079	Helado/plato		✓			2
100080	Chaqueta/maracas			✓		3
100081	Rodilla/pato			✓		3
100083	Troca/conejo		✓			2
100088	León/pelota			✓		3
100094	Arcoiris/nieve/fuego			✓		2
100098	Tiburón/granja/corbata					
100102	Trineo/hueso/escribir				Verb as foil	2
100103	Nieve/libro/dar				Verb as foil	2
100103	Cuchara/búfalo/volar				Verb as foil	2
100109	Mochila, pelota, banana		✓			3
100111	Pastel/libro/bota	Changed ID to 100124 because ID 100111 already exists	✓			3