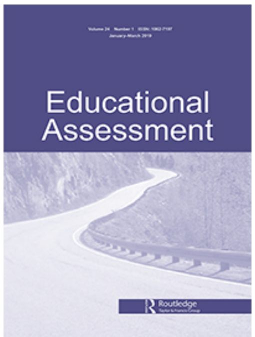


Routledge
Taylor & Francis Group

Educational Assessment

ISSN: 1062-7197 (Print) 1532-6977 (Online) Journal homepage: https://www.tandfonline.com/loi/heda20

Developing a Measure of Spanish Phonological Awareness for Preschool Age Children: Spanish Individual Growth and Development Indicators

Alisha Wackerle-Hollman, Lillian Durán, Stephanie Brunner, Jose Palma, Theresa Kohlmeier & Michael C. Rodriguez

To cite this article: Alisha Wackerle-Hollman, Lillian Durán, Stephanie Brunner, Jose Palma, Theresa Kohlmeier & Michael C. Rodriguez (2019) Developing a Measure of Spanish Phonological Awareness for Preschool Age Children: Spanish Individual Growth and Development Indicators, Educational Assessment, 24:1, 33-56, DOI: 10.1080/10627197.2018.1545570

To link to this article: https://doi.org/10.1080/10627197.2018.1545570

Published online: 23 Nov 2018.

EDUCATIONAL ASSESSMENT
2019, VOL. 24, NO. 1, 33–56
https://doi.org/10.1080/10627197.2018.1545570

Developing a Measure of Spanish Phonological Awareness for Preschool Age Children: Spanish Individual Growth and Development Indicators

Alisha Wackerle-Hollman[a], Lillian Durán[b], Stephanie Brunner[a], Jose Palma[a], Theresa Kohlmeier[c], and Michael C. Rodriguez[a]

[a]University of Minnesota – Twin Cities; [b]University of Oregon; [c]University of Wisconsin-Stout

ABSTRACT

Spanish speakers in the United States are a steadily increasing population, up by 233% since 1980. Given the growing population of dual language learners (DLLs) and the large numbers of Spanish-speaking children enrolled in pre-kindergarten programs, addressing the educational needs of preschool-aged DLLs has become a national imperative. Specifically, the intersection of this growing population and the dearth of appropriate assessment tools to evaluate DLLs early language and literacy skills creates a need for assessments that accurately measure preschool performance. This manuscript reports on the iterative design process of a measure of Spanish phonological awareness for preschool-aged DLLs: Spanish Individual Growth and Development Indicators (*S-IGDI*) Primeros Sonidos. We employed measure design framework to develop the measure and tested item function within a study of 970, 4–5 year old DLLs. Results, including item level analyses and evidence regarding construct and criterion validity are reported.

Spanish speakers in the United States (US) are a steadily increasing population, up by 233% since 1980. Projections indicate between 39 to 43 million Spanish speakers will live in the US by 2020 (Lopez & Gonzalez-Barrera, 2013). Currently, 1 in 4 preschool-age children are Latino, and in Head Start, young dual language learners (DLLs) represent 30% of national enrollment; of these, 89% speak Spanish at home (U.S. Department of Health & Human Services [USDHHS], 2013; Wildsmith, Scott, Guzman, & Cook, 2014). In Migrant Head Start, Spanish-speaking children represent 85% of enrollment (USDHHS, 2009). Given the growing population of Latinos in the US and the large numbers of Spanish-speaking children enrolled in pre-kindergarten (Pre-K) programs, addressing the educational needs of young Spanish speakers has become a national imperative (Garcia & Jensen, 2009; U.S. Census Bureau, 2010).

Current educational approaches with Spanish-English dual language learners (SE-DLL)[1] are not significantly improving English reading outcomes (National Center for Education Statistics [NCES], 2013). Since 1975, a persistent gap has been present in reading scores on the National Assessment of Educational Progress (NAEP) between White and Latino children at ages 9, 13, and 17 (NCES, 2013). The most recent long-term reading trend data indicate that although the scores for Latino students have improved, so too have the scores of White students, essentially maintaining the gap since 1975 at all age groups except at age 13 where the gap has narrowed. These data provide

CONTACT Alisha Wackerle-Hollman wacke020@umn.edu Urban Research Outreach and Engagement Center, University of Minnesota, 2001 Plymouth Avenue North, Minneapolis, MN 55411

[1] The term "Spanish-English dual language learners" (SE-DLLs) encompasses other terms frequently used, such as Limited English Proficient (LEP), bilingual, English language learners (ELL), English learners, and children who speak a Language Other Than English (LOTE).

evidence that the field has failed to implement effective practices, including appropriate assessment, for SE-DLLs. Without ample resources to support academic success, SE-DLLs experience an elevated level of risk of academic failure (Goldenberg, 2008). High quality instruction, intervention and assessment for SE-DLLs are needed to better educate this population, but research is only beginning to discern what practices and tools are most effective. We focus on one aspect of supporting SE-DLLs: developing meaningful assessments to accurately evaluate SE-DLL early literacy performance.

At the core of any evidence-based educational program is the development of assessments to measure instructional targets. Given the cultural and linguistic diversity in the US (Barrera & Liu, 2010; USCB, 2010), technically adequate measures designed to accurately capture the ability levels of a broad range of children are needed. These tools should include screening and progress monitoring measures to be used within the context of data-based decision making to support instructional modification (Brown & Doolittle, 2008; Haager, 2007; Hosp, Hosp, & Dole, 2011). Without reliable and valid inferences about children's abilities, it is difficult to gain meaningful information about student performance. We respond to these needs by describing the development and calibration of the new Spanish Individual Growth and Development Indicators (*S-IGDIs*) phonological awareness measure. *S-IGDIs* represent an innovative approach to developing a measure in Spanish that will provide early childhood practitioners with an efficient, cost-effective, and accurate tool for examining performance of preschool-aged SE-DLLs.

Supporting early literacy development

One area to focus assessment efforts for SE-DLLs is early literacy skills in both English and Spanish. Early literacy skills are those skills that are known predictors of later reading success (National Early Literacy Panel, 2008). Without the ability to read in English, SE-DLLs may struggle to achieve adequate educational outcomes across all academic content areas, as SE-DLL children in most US classrooms must be able to read in their second language to complete tasks in other domains. As such, intervening early and successfully to support the prerequisite skills necessary for learning to read in English remains integral to improving the long-term academic outcomes of SE-DLLs in the US (Goldenberg, 2008).

Exactly how to intervene to best support the long-term goal of reading in English is less understood. Researchers have demonstrated the presence of cross-linguistic transfer for SE-DLLs. Researchers show SE-DLLs use their native language skills and knowledge to support the development of English language and literacy skills, and in some circumstances also use their English language and literacy skills to support development of Spanish (c.f. Cárdenas-Hagan, Carlson, & Pollard-Durodola, 2007; Geva & Genesee, 2006; Melby-Lervåg & Lervåg, 2011). Although research findings vary in describing the nature and direction of the association between first (L1) and second languages (L2), it is generally accepted that both languages contribute in some way to how a child learns to read in the second language (Genesee, Geva, Dressler, & Kamil, 2006; Hammer, Lawrence, & Miccio, 2007; Melby-Lervåg & Lervåg, 2011). Researchers also have shown that intervention delivered in the native language has produced meaningful gains on English and Spanish early literacy and language outcomes for preschool-age SE-DLLs (Farver, Lonigan, & Eppe, 2009). Taken together, these research findings support the need to assess and support early language and literacy in L1 and L2 during the preschool years.

Spanish and english phonological awareness

We define Phonological Awareness (PA) 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 (Torgeson & Mathes, 2000). In English, PA skills such as detecting initial sounds, segmenting, blending, and elision have consistently been shown to predict reading ability when measured in preschool (Anthony & Lonigan, 2004; Muter, Hulme, Snowling, & Stevenson, 2004). Associations have been found between early Spanish PA and later Spanish reading development (Branum-Martin et al., 2006; Farver, Nakamoto, & Lonigan, 2007; Gorman & Gillam, 2003). Spanish PA skills are also predictive of reading achievement in English (Leafstedt & Gerber, 2005; Melby-Lervåg & Lervåg, 2011; Proctor, August, Carlo, & Snow, 2006; Swanson, Rosston, Gerber, & Solari, 2008). In Pre-K, PA Spanish skills such as elision and rhyming correlate with the same English skills (Dickinson, McCabe, Clark-Chiarelli, & Wolf, 2004; López & Greenfield, 2004).

Although there is sufficient evidence to suggest cross-linguistic transfer is present for PA, associations may be more complex than can be modeled by correlational studies alone (Melby-Lervåg & Lervåg, 2011). Branum-Martin et al. (2006) found through multi-level modeling that Spanish and English PA loaded best as separate factors (and potentially represent separate constructs), despite high correlations. These results support the notion that some PA skills may be language-specific and associations may vary across the PA skills that are measured. For example, Goodrich, Lonigan, and Farver (2013) found that children with higher elision scores in Spanish demonstrated higher English elision skills at the end of intervention than those children with lower initial scores in Spanish; however, the same findings were not present for blending tasks, suggesting variability in findings based on different PA skills. Further, SE-DLLs English performance may be more strongly predicted by English literacy skills than Spanish early literacy skills (Hammer, Jia, & Uchikoshi, 2011). Researchers have also identified mediators such as language of instruction and language proficiency that may influence these cross-linguistic associations (Cárdenas-Hagan et al., 2007; Goodrich, Lonigan, & Farver, 2014).

Given the lack of clarity in understanding how to best intervene with SE-DLLs relative to their native language and literacy ability, and the current evidence available regarding the cross-linguistic association between Spanish and English, it seems logical to target assessment development of early literacy skills in Spanish. In this way, we can improve our understanding of the nature of cross-linguistic transfer as well as examine student responses in ways that inform the potential for instruction and intervention in Spanish, English or both languages to support English reading development.

Understanding the need for spanish assessment tools

To provide instructional practices that maximize SE-DLL student performance in English and Spanish, it is important to help practitioners make data-based decisions with empirically sound measurement tools. Specifically, for SE-DLLs we must be able to ascertain to what degree each language has developed so that educators can respond with differentiated approaches. High-quality Spanish measures that can accurately identify candidates for language and literacy intervention are an important contribution to understanding how to best intervene with SE-DLL students.

At present, there are few high-quality, commercially available, Spanish PA early childhood assessment tools accessible to practitioners that can be used to inform instruction during the Pre-K years. We focus here on screening assessments because of their instrumental link to data-based decision making at the classroom level (McConnell, Wackerle-Hollman, & Bradfield, 2014); in differentiated instructional models where practitioners aim to meet the unique needs of each student with leveled instructional strategies, screening is essential. Of the Spanish screening assessments available, including the Get Ready to Read-Spanish (*¡Prepárate a leer!* – Revisada, *GRTR-S-R*, Lonigan, 2003), Istation ISIP (Mathes & Torgesen, 2005), and the Circle Phonological Awareness Language and Literacy System (C-PALLS, Landry, Assel, Gunewig, & Swank, 2004), none provide published information on the process of the development of the Spanish versions, making it difficult to ascertain to what degree translation of existing English items or measures was used to develop the Spanish measures (Barrueco, López, Ong, & Lozano, 2012). This is problematic for two reasons. First, limited information on current assessments prevents a review of how the assessment accesses an identified construct. Although the trajectory of Spanish PA is similar to English, it is not identical (Anthony et al., 2011), and if assessments are translations of English PA measures, then important facets of Spanish PA may not be represented because the measures may represent an English construct translated to Spanish, rather than a Spanish early literacy construct. Second, without information regarding assessment design and validation, it is difficult to determine whether factors such as length and difficulty of Spanish words chosen, complexity of distractors (incorrect options), and construct irrelevant features were addressed. Fortunately, additional tools are available in the K-12 arena that may offer a bridge to validate predictive efforts such as the Indicadores Dinámicos del Éxito en la Lectura (IDEL, Good, Baker, Knutson, & Watson, 2007) and FASTBridge Spanish Reading measures (Christ, 2013), both designed with Spanish development in mind, in contrast to existing translated tools.

Research questions

Given the need for new measures of Spanish early literacy to improve our understanding of how to intervene with young SE-DLLs, we describe the development process and examine the psychometric properties of the *Primeros Sonidos/First Sounds (PS/FS) S-IGDIs* measure. We present the process of item design followed by large-scale validation trials to answer the following questions: (a) What are the item-level statistics of *PS/FS*? (b) To what degree do *PS/FS* items appropriately measure child Spanish PA abilities? and (c) To what extent is *PS/FS* associated with other standardized measures of Spanish language and early literacy?

Method

Instrument

Development of the spanish individual growth and development indicators

The *S-IGDIs* are a set of early literacy screening measures designed for use with SE-DLL preschoolers. Grounded in Wilson's (2005) measure design framework, the *S-IGDIs* seek to measure early literacy skills in Spanish that predict later literacy outcomes in English. *S-IGDIs* are designed to assess SE-DLL preschool performance in Spanish in complement to their English counterparts: English *IGDIs 2.0*. These assessments were developed to measure early literacy in English and Spanish as separate constructs and to provide information on student ability in each language to promote data-based decision making and improved instructional practices. *S-IGDIs* are not translations of English measures. The unique features of Spanish phonology, word structure, and semantics informed *S-IGDI* test development. This was accomplished through four components from Wilson's framework for developing measures: construct map, item design, outcome space, and measurement model (Wilson, 2005).

Defining the construct map

To begin the iterative process of measure development, the research team conducted an extensive literature review to arrive at constructs of interest, construct definitions, and research-based suggestions for item and measure design. A search using the key words "Spanish+ Assessment + Phonological Awareness" and "Early Literacy+ Spanish+ Assessment" targeting references post-2010 was employed in Google Scholar, EBSCO, Academic Search Primer, Psych INFO and ERIC, yielding a total of 39 articles, book citations, and assessment citations for review. Articles that referenced Spanish but did not include Spanish measures were removed as well as articles that included measures for only Kindergarten and older grades, reducing the pool of articles to 28. Each source was reviewed for suggested PA Spanish tasks, and presentations of each task were cataloged for a detailed review see Wackerle-Hollman et al., 2012). From the literature review, the syllable is a salient phonological unit in Spanish; blending, initial sound identification, and syllable segmenting tasks may be particularly appropriate. In addition, Spanish phonological awareness may be best represented by a unitary construct sampled by various tasks that are indicative of differing levels of performance on that construct (e.g., Anthony et al., 2011). Results of the comprehensive literature review were synthesized to inform a PA continuum of skill that functions as the foundation of the *PS/FS* measure design process. Within the continuum, we identified early childhood skills indicative of lower PA ability, such as syllable identification and blending, to skills indicative of higher levels of PA ability, such as elision (Anthony et al., 2011; Gorman & Gillam, 2003). These findings supported the notion that Spanish PA may be best represented as a unitary construct that may be sampled by tasks to assess skills that may have differing trajectories or developmental timelines.

Designing items

For each measure designed, the concept presented aligned with the construct definition and represented the targeted skill. Four initial Spanish PA measures included representations from various locations on the construct map: a rhyming task, a first sounds task, a blending task, and an elision task. All new measures of Spanish PA were designed by manipulating presentation style (e.g., receptive choice or expressive production) and/or delivery features (e.g., using images when possible to reduce cognitive burdens, etc.) present in the research literature. However, after initial pilot studies in Years 1 and 2, we focused our design efforts on one of the four measures: *Primeros Sonidos/First Sounds (PS/FS)* because the remaining measures failed to meet empirical and qualitative criteria. Information on the design, testing, and eventual elimination process for the other Spanish PA measures can be found in the *S-IGDI* technical reports (2016) available at http://innovation.umn.edu/igdi/projects/spanish-igdis/

During our design process, our construct definition of PA to define measure components. We noted that children should be able to "detect and discriminate between small units of sound," and as such we included measures that required such detection (e.g., in first sounds we asked children to identify the first sound or syllable they hear in a word).

S-IGDI PS/FS items were designed to incorporate culturally relevant factors and to include measurement review processes that carefully evaluated item-level features (Wackerle-Hollman et al., under review). As researchers suggest translated measures do not account for the ways that Spanish early literacy skills may develop differently from these skills in English (Barrueco et al., 2012; Peña, 2007; Wackerle-Hollman et al., 2012), the *S-IGDI* item design process consisted of using words from existing Spanish word corpuses including Spanish early childhood curricula, high-frequency words collected from an inventory of 100 Spanish children's picture books, and the Spanish version of the MacArthur Bates Communicative Development Inventory (MacArthur *Inventarios del Desarrollo de Habilidades Comunicativas*; Jackson-Maldonado et al., 2003). We also carefully selected and designed images to reflect culturally relevant and developmentally appropriate targets. Four national experts external to the project team were then recruited for their knowledge in bilingual assessment, Spanish language development, and early childhood research with SE-DLLs to review all items. Feedback was provided in one-on-one interviews and was incorporated into *S-IGDI PS/FS* design. Feedback included careful consideration of the selection of words and their usability with a variety of Spanish dialects and cultures within the US Latino population (for example, in Puerto Rican Spanish, *llanta* [tire] is *rueda*, and *naranja* [orange] is *china*).

Once item-level targets 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 the *PS/FS* task we excluded distractors that rhymed). We also strategically manipulated the number of distractors present on each item to include either one (distractor and target) or two (two distractors and a target). Although we recognized that including just one distractor would increase the likelihood of selecting a correct response by guessing given an odds ratio of 1/2 (.50), we reasoned that item fit for these items would be poor if guessing was present (Smith, 1994). Regarding linguistic limitations, the research team carefully considered appropriate target sounds given particular nuances of Spanish sounds. For example, when considering phonemes, no *PS/FS* items included/w/or/h/as a target because all Spanish words that begin with the grapheme/w/are English-influenced words (e.g., wafle, waterpolista) and because no Spanish words begin with a sounded/h/, as this sound is silent in Spanish. Also, we were careful to prevent words or images with similar initial sounds from appearing on the same item. For example, the graphemes/c/,/s/, and/z/all sound the same in Spanish (e.g., all make a/s/phoneme), so a zebra (*cebra*), snake (*serpiente*), and shoe (*zapato*) could never appear in the same item. We also considered the syllabic nature of Spanish words, recognizing that it is often difficult to isolate the initial phoneme (Gorman & Gillam, 2003), so we included *PS/FS* items that isolated the initial syllable (e.g.,/cha/for *chaqueta*, and/za/for *zapato*). Final items in the *PS/FS* measure targeted both initial syllable and initial phoneme.

Regarding pragmatic item design features, for *S-IGDIs* to be useful for early childhood practitioners who serve SE-DLLs, they must be easy to use and understand. These features coincide with the characteristics of general outcome measures (GOM), which are a class of measures designed to be predictors of meaningful long-term academic success while maintaining efficiency, ease of use, and feasibility (Fuchs & Deno, 1991). *PS/FS* items were designed to be easily delivered by lay-level practitioners; all prompts and response choices are printed on the back of each item and all measures require less than five minutes to deliver and score. Evidence to support the usability and feasibility of *S-IGDIs* has been collected through three primary means. First, scoring rubrics were developed and applied based on four criteria: (a) child interaction with *S-IGDIs* will produce meaningful responses through child active engagement (e.g., making eye contact, demonstrating interest, and attending to the assessors requests); (b) child interaction with *S-IGDIs* will produce a valid response pattern (e.g. the child's response should be a logical extension of knowledge based on the prompt and not off topic); (c) the *S-IGDIs* should be easy to use; (d) the *S-IGDIs* should be timely to deliver and score. Second, we obtained expert reviews from the four SE-DLL early literacy researchers previously described. Third, teacher surveys were administered addressing feasibility and usability.

Qualitative summaries of *PS/FS* were collected from four bilingual data collectors who assessed and rated 35 students. Scores ranged from 0–3, where 0 indicated the variable was unsatisfactorily achieved (not observed) and unresolvable (no clear resolution was available to repair the prompt to produce the desired behavior); a 1 indicated the variable was unsatisfactorily achieved but resolvable (a clear resolution was available to repair the prompt to produce the desired behavior); a 2 indicated the variable was satisfactorily achieved (the behavior of interest was observed, but not as anticipated in measure design), but had some reservations that were resolvable (a clear resolution was available to improve the instance of the desired behavior); and 3 indicated the variable was satisfactorily achieved without reservations. Results indicated initial trials for *PS/FS* yielded an average score of 2.8. All assessors reported *PS/FS* required 1–3 minutes to deliver and score and students appeared engaged with all items.

Four expert reviewers were given item sets to review with instructional prompts. All reviewers agreed that *PS/FS* was efficient and easy to deliver. One reviewer requested changes in the prompts to make it more accessible by lay-level assessors and another noted that the measure-administration training must be provided in Spanish and English to support bilingual classrooms. All suggested modifications were incorporated into *PS/FS* measure design.

Finally, a subset of bilingual teachers from the study (n = 20) reported on feasibility and usability via survey and interview formats. The sample included teachers with an average of seven years working in early childhood programs. Teachers reviewed the materials, instructions for administration, and scoring forms and then observed a trained data collector giving the assessment to one of their students. Following the observation, teachers completed the Feasibility and Utility survey that included two sections: ease of administration and feasibility, utility and practicality. The ease of administration section included 16 questions, where 13 questions used a four-point rating scale ranging from *strongly agree* to *strongly disagree*, and three were open ended. Example items included questions about instruction ease, administration time, scoring ease, and overall measure relevance. Results indicated teachers strongly agreed with items an average of 73% of the time (range of 45% to 90%), where only 45% strongly agreed with the item "The scoring sheets were easy for me to use." One teacher (5%) reported disagreement with this item. No teachers reported *strongly disagree* for any of the items. Overall, teachers found it easy to evaluate responses, the measure was brief to administer and score, and the measure included easy-to-understand standardized instructions. The three open-ended questions asked participants how they would make sample questions, items, and prompts easier to administer. Only seven teachers provided feedback, including suggestions such as changing some of the images, providing instructions in Spanish *and* English on the item sets, and including frequently asked questions specific to *PS/FS* right on the item cards.

The feasibility, utility, and practicality section of the survey included nine rating-scale items to evaluate the extent to which teachers could easily incorporate the assessment protocol into classroom practices and the degree to which they found resulting *S-IGDI* data useful in supporting language and early literacy instruction. All teachers reported *strongly agree* to the question "Is *PS/FS* developmentally appropriate for SE-DLLs in your classroom?" Teachers agreed that *PS/FS* scores were very useful for supporting instructional modifications, and that testing materials were interpretable for children, indicating their students would be familiar with the images provided. All teachers strongly agreed that *PS/FS* appropriately measured PA and agreed that *PS/FS* measured skills they taught in their class.

Across sources, results were summarized and incorporated into item design when relevant. The evidence from these three sources – qualitative rubrics, expert and teacher surveys– indicate *PS/FS* demonstrates utility and feasibility in early childhood classrooms.

Creating an outcome space

The outcome space represents the frame for item presentation and scoring. During *PS/FS* item design, we reasoned a receptive response in a multiple-choice presentation style would provide ample opportunities to administer a significant number of items while also allowing for standardization of dichotomous scoring. As such, all *PS/FS* items included two or three images; selection of the target image yielded a correct score.

When considering the outcome space, we also reviewed our intentions for soliciting responses. Responses come from preschool-age students, and given our effort to limit construct-irrelevant features, using images to represent the *PS/FS* phonemes and syllables reduces the cognitive load and contributes to a response that was less likely to be influenced by memory capacity. Further, we developed prompts that reduce cognitive dissonance by labeling all images to prevent children from drawing on an existing label of an image they may have if it is not the same label used in the item. For example, a *PS/FS* item with three images of *brazaletes* (bracelets), *cabra* (goat), *alfombra* (rug) is presented by labeling each image and then providing the prompt, "*Cual de estos dibujos empieza con*/bra/?" [which picture starts with the sound/bra/?]. We allowed responses to be provided verbally or by pointing to the correct image since pointing to the image without saying it represents knowledge of the construct. In most of our trials, students said and pointed to the image on the card or pointed to the item and then verbally labeled it when prompted. These features aligned with the *PS/FS* measure focusing on receptive understanding of the item content. We are unaware of any studies that demonstrate empirical differences in predictive power of receptive versus expressive assessment approaches; however, Anthony et al. (2011) noted that receptive measurement approaches resulted in easier items than expressive measurement approaches.

Further, our selection of the Rasch model for the selected-response (SR) items for receptive measurement in the outcome space acknowledges the differences in the location of the items (so called item difficulty, the ability required to have a 50% chance of correctly responding) and the locations (ability) of persons. The research literature on SR measures notes that the accumulation of the possibility of guessing is negligible in most cases (see Rodriguez, 2005, for a review). In addition to considering these factors, we also Used results from the outcomes space in the measurement model to evaluate the fit of items and eliminate those with poorer fit, via item-measure correlations, INFIT person fit indices, and overall item *p*-value statistics. We also examined the extent to which items functioned similarly across seasons, years, and in the presence of anchor items, including examination of item parameter drift over time and displacement when anchored with common items across forms. Finally, we evaluated differential item functioning (DIF) statistics for every item. Each item was carefully evaluated for consistency and stability (described in detail below).

Applying a measurement model

S-IGDIs employ a Rasch modeling approach to measurement (Rasch, 1980). Rasch modeling is an ideal fit for *S-IGDI* design because Rasch models are uniquely suited for measurement design and are consistent with Wilson's model. When the data fit the Rasch model, the property of invariance between items and persons is achieved, such that person location is not dependent on the specific items administered, and item location is not dependent on the persons responding to the item (Andrich, 2004; Engelhard, 2013; Sick, 2010). This property is an advancement beyond classical test theory, where person scores and item difficulty are sample specific. Specifically, the Rasch model conceptualizes responses as a function of the item locations (e.g., difficulties) and person locations (e.g., abilities). Therefore, children's scores on *PS/FS* are based on their modeled ability rather than on normative performance of the items with which they interacted.

When the data fit the model, Rasch provides units of measurement on an interval level so that there is a uniform interpretation across the scale. In addition, the Rasch model meets the requirements for invariant measurement: (a) person measurements must be independent from items the person is administered; (b) a more able person must always have a greater chance of success on a given item than a less able person; (c) difficulty of items must be independent of the sample who received the items; (d) any person must have a greater probability of success on an easy item than on a more difficult item; and (e) items and persons are located on the same scale (Andrich, 2004; Bond & Fox, 2015; Engelhard, 2013; Sick, 2010). The Rasch model requires at least 100 responses per item to adequately scale and calibrate items.

However, even with the benefits of the Rasch model, there are various contrasting perspectives in the research literature; not all measurement experts agree with the claims made by Rasch advocates. We encourage the reader to review Sick (2010) and Andrich (2004) to gain further perspective on these contrasts. Specifically, some argue that additional parameter models should be tested rather than selecting a measurement model *a priori*. Although we did not employ this approach, we acknowledge that some see value in accounting for additional variance through use of the 2PL and 3PL models. Indeed, it is true that using 2PL or 3PL models will typically account for more variance; however, the focus of test design is *not* to explain maximum variance; instead, it is to maximize accurate measurement. Further, Rasch is a more parsimonious model and Rasch person parameter estimates correlate with 2PL models at rates above .90 (de Ayala, 2013). To satisfy such interests, we conducted a model comparison with the 2- and 3-parameter logistic models, with results reported below.

Participants

Participants & setting

SE-DLL children in public and private early childhood education programs in MN, FL, CA, UT, KS, and IL were recruited for this study. In order to participate in the study, children had to speak Spanish according to parent or teacher report, be 4–5 years old, and be eligible for Kindergarten the following year. No children who were native English speakers were included in this study. Participating sites were recruited based on prior study enrollment, relationships with research staff, and email recruitment strategies. The study spanned 46 schools and 135 classrooms over two academic years (2013–2014 and 2014–2015). A total of 124 teachers participated in classroom surveys, with some teachers leading more than one classroom. Only 5% of classrooms included SE-DLL students but did not include bilingual teachers or educational support staff. Graduate research assistants and data collectors obtained signed parental consent after teachers sent consent forms home in student backpacks. Across all sites in the 2013–2014 year, 491 children returned consents to participate. A sample of 479 consented children was added in the 2014–2015 school year, for a total sample of 970 children.

We aimed to recruit a dialectically diverse sample of Spanish-speaking children from FL to obtain Caribbean, Central American, and South American Spanish dialects; and from UT, CA, MN, KS, and IL to obtain Mexican and Central American dialects. Students were recruited by teachers initially indicating which students in their classroom spoke Spanish at home. Demographic forms were completed by 75% of the sample (732 families) which included 47.6% males. Child ages ranged from 4:0 to 5:11 at the beginning of the academic year, 5.9% of the sample received special education services (according to parent report), and 57% of the sample attended Head Start and/or had weekly family incomes less than $500. A recent report from the National Research Center on Hispanic Children and Families noted that nearly 66% of the total Hispanic population in the US live in low-income households (defined as two times the federal poverty level; Wildsmith, Alvira-Hammond, & Guzman, 2016); our sample is relatively representative of US SE-DLLs in this respect. Specific to language exposure, parents reported that 69% of the sample heard Spanish from birth, 24% heard both English and Spanish, and 7% did not respond to this question. Children who were most comfortable speaking in Spanish or in a mix of Spanish and English comprised 72% of the sample. Although all native Spanish speakers, 10% of children were rated by their parent to be most comfortable speaking in English, perhaps due to the saliency of their current exposure to English in their preschool classrooms. In addition, approximately 9% of the sample reported speaking other languages at home including heritage languages such as Mam and Quechua, which are native to Central and South America. See Table 1 for additional demographic information.

At the classroom level, teachers reported language of instruction and level of language proficiency. Forty-five percent of the classrooms provided instruction in English, 53% used both

Table 1. S-IGDI sample descriptives.

Characteristic	2013–2014 (n = 396) %	2014–2015 (n = 336) %	Total (n = 732) %
Male	52.5	47.0	47.6
Receiving special education services	4.0	7.7	5.9
Latino (general)	52.1	56.0	54.1
Mexican	15.4	17.6	17.2
Puerto Rican	14.3	9.2	12.2
Caribbean	2.5	< 1.0	1.2
Central American	2.8	4.5	3.6
South American	1.5	< 1.0	< 1.0
Multiple races/ethnicities	11.5	11.8	10.7
Languages spoken to child from ages 0 to 1			
Spanish	71.0	70.8	69.1
Both	24.3	25.6	24.0
Language child uses when talking at home			
Spanish	48.5	49.1	49.5
Both	24.0	41.7	32.2
English	12.0	8.3	10.4
Other	15.5	0.0	8.6
Household weekly income			
Less than $500	63.1	66.8	65.0
$501–700	22.7	27.5	24.6
$701–900	7.7	0.0	4.4
More than $901	6.6	5.7	6.1
Mother's highest level of education			
6th grade or less	13.9	19.2	16.7
Less than 12th grade	18.0	23.3	20.3
GED	12.1	7.2	10.0
High school diploma	11.3	21.6	15.5
Some education after high school	16.6	18.5	17.5
Associate's degree	9.7	0.0	5.6
College degree (BA/BS)	13.1	5.1	9.5
Graduate/professional degree	5.4	5.1	4.8

Note: Sample sizes in this table are based on number of families who returned the family survey, not on the total consented samples. The multiple ethnicities group includes those who selected two or more ethnicities on the family survey.

languages for instruction, and 1% used Spanish only. Fourteen percent of lead teachers were bilingual, 45% of classrooms included a Spanish-speaking assistant, and 5% had no Spanish-speaking adults in the classroom. Of the 95% of classrooms containing Spanish-speaking or bilingual adults, 28% said other support staff used Spanish, 14% said Spanish interactions occurred via specialists, and 30% reported Spanish-speaking parent volunteers in the classroom. Fifty-six percent of classrooms reported that children spoke Spanish in the classroom.

Measures

S-IGDI primeros sonidos/first sounds

Six *S-IGDI* measures, including *PS/FS*, reached the final stages of development and were administered as part of a larger study. *S-IGDIs* were designed for children whose native language is Spanish and therefore are not appropriate for children whose second language is Spanish. The remaining five tasks were measures of oral language and alphabet knowledge. The *PS/FS* measure included four sample items and 21–28 test items, depending on form administered. The administrator modeled the first two sample items. The second two sample items were practice items and feedback was delivered for correct and incorrect responses. If the child responded correctly, then the administrator responded by saying, "*muy bien* (very good)." If the child responded incorrectly, then the administrator provided the correct response and re-administered the item. The measure was discontinued if the child responded incorrectly twice on a single sample item, which included 1% of the sample. If the child could not respond to the sample items even with direct feedback prompting them to repeat a correct response, then it was assumed the child did not understand the task or was unable to engage with the items. If children successfully engaged the sample items, all items were administered and no discontinue criteria were employed.

PS/FS was designed to measure a child's ability to identify the initial sounds of words independent of meaning. Items were not ordered by difficulty in item sets; instead, they were randomly selected from the item pool and then delivered in a standardized order. Administrators presented each item by pointing to and labeling each of the two or three images followed by the prompt, "*¿Cuál de estos dibujos empieza con ___?*" ("Which of these pictures begin with ___?") inserting the target sound in the blank. Item targets varied, including single phonemes (e.g.,/c/), and initial syllables (e.g.,/cha/). Administrators used a manualized list of example phonemes to facilitate standardization of sound production without the schwa sound (e.g./t/instead of/te/). Children responded by pointing to or verbally labeling the image during the untimed administration. Scores were recorded as correct (identifying the image matching the target sound) or incorrect (identifying any other image).

Sampling scheme

PS/FS included 104 items; as such, to minimize the testing burden on participants, we employed a sampling scheme to strategically test all items at each site, allowing students to interact with a reduced number of items. Over the course of two academic years, eleven blocks were distributed across five forms using matrix sampling, and then assigned to students in an annual randomization. In Year 1, nine blocks were used, and in Year 2, students interacted with two blocks of items selected from Year 1. Each block consisted of 7–9 items, and combinations of three blocks (21–27 items) comprised a form in Year 1, whereas a combination of two blocks comprised a form in Year 2 (15 items). Blocks were sequentially assigned to forms (e.g. Form 101 included blocks A,B,C; Form 102 included blocks D, E, C etc.) and there were a total 5 forms. One block of items was common across adjacent forms so that calibration of all items could be linked and evaluated as a total item bank. Each child saw three forms (one form per season) including 4 sample items and 21–27 of 104 *PS/FS* items. Within each form for Year 1, all blocks were constructed by randomly selecting items from the item pool and then standardizing order of presentation. In Year 2, items were selected from Year 1 blocks to be administered to all student participants. Because of this design, typically employed by test developers during item development phases and during field-testing of new items, different numbers of students responded to different blocks of items, with the largest samples of responses obtained with the common (anchor) blocks used for linking forms. At the end of Year 2 all items were concurrently calibrated and treated as a single item pool.

Norm-referenced criterion measures

A strategically sampled subset of children also completed a criterion measure of early literacy in Spanish to understand the association between *S-IGDI* performance and existing, commercially available, norm-referenced measures of Spanish PA during the 2013–2014 academic year. Students were selected based on dialectical membership, gender, age distribution, program type, and geographic region to ensure that the sample included a dialectically diverse set of students with variability in performance (see Table 2). Because there are few available measures that examine the construct of Spanish early language and literacy rather than translating measures of English phonological awareness into Spanish, we expected limited correlations with *PS/FS*. The Test of Phonological Awareness in Spanish (*TPAS*, Riccio, Imhoff, Hasbrouck, & Davis, 2004) and Get Ready to Read! - Revised Spanish version (*¡Prepárate a leer!* – Revisada, *GRTR-S-R*, Lonigan, 2003) were administered to the selected sample (n = 64 for *TPAS*, n = 60 for *GRTR-S-R*). Students were strategically sampled to represent variability in age (48–60 months) and geographic region (distributed across states).

The *TPAS* consists of four subtests: Initial Sounds, Final Sounds, Rhyming Words, and Deletion. However, the Deletion subtest was normed for children ages 6+, so administrators only gave the first three subtests. During administration, three (Final Sounds) or four (Initial Sounds and Rhyming Words) sample items were delivered with feedback for correct responses

Table 2. Demographic information for the criterion assessment subsample.

		(n = 64)
	(n = 60) GRTR	TPAS
Characteristic	%	%
Male	67.3	48.4
Receiving special education services	-	8.0*
Latino (general)	83.3	68.8
Mexican	19.6*	18.0*
Puerto Rican	11.8*	16.0*
Other Latin American	5.9*	8.0*
Languages spoken to child from ages 0 to 1		
Spanish	84.0*	84.0*
Both	13.6*	13.6*
Language child uses when talking at home		
Spanish	80.0*	67.3*
Both	20.0*	14.3*
English	-	12.2*
Other	-	6.1*
Household weekly income		
Less than $500	63.3	67.2
$501–700	23.6	20.3
$701–900	5.5	3.1
More than $901	7.3	-
Mother's highest level of education		
6th grade or less	14.3*	23.9*
Less than 12th grade	22.4*	19.6*
GED	16.3*	15.2*
High school diploma	6.1*	15.2*
Some education after high school	10.2*	8.7*
Associate's degree	12.4*	4.3*
College degree (BA/BS)	14.3*	8.7*
Graduate/professional degree	4.1*	4.3*

Note: *indicates a smaller sample of 51 for TPAS and 50 for GRTR. This information comes from the family questionnaire and not all parents responded to the family questions. As such, fewer students are reported on than in the subsample.

and corrective feedback for incorrect responses. Administration rules required that the test be discontinued if a child was unable to respond to any of the sample items. During this study, we administered all test items; when a child did not respond, we asked the question again and if they continued not to respond we marked the item as incorrect. A series of 20 (Final Sounds and Rhyming Words) to 30 (Initial Sounds) items were provided to children who passed the sample items. Children responded either "*sí*" or "no" to each item, and administrators scored a 1 for a correct response and a 0 for an incorrect response. Total scores were recorded as the sum of 1-point responses for each subtest.

The *GRTR-S-R* is a 25-item flip book with one sample item, and takes 10–15 minutes to administer (Lonigan, 2003). The *GRTR-S-R* measures early literacy skills with 12 items that measure phonological awareness, including two items specific to initial phoneme awareness, however these 12 items are not available as a subtest. Administration occurred in Spanish and children saw each item in succession with the assessor scoring 1 for a correct and 0 for an incorrect response, with total score as the sum of 1-point responses. The *GRTR-S-R* does not have a discontinue criterion; all children saw all 25 items.

Procedures

Data collector training and fidelity

Two trained bilingual graduate research assistants conducted an in-person or digitally broadcasted comprehensive training with all Spanish-speaking data collectors to maintain standardized administration procedures and fidelity across participating sites. Trainings occurred twice annually due to data collector attrition and scheduling limitations and included eight sections: logistics, ethical principles for assessment, privacy and confidentiality, sampling scheme, data collection activities, scoring and entry procedures, and fidelity of implementation requirements. Data collectors watched training videos of measure administration and attended 3-hour training sessions, with specific attention to scoring. All data collectors were required to achieve 100% fidelity within three consecutive attempts, the first two allowing for feedback to support changes in performance. Fidelity trials occurred in Spanish with a team member who was a native Spanish speaker to check presentation fluency. All training materials were available in English and Spanish and designed for assessors with limited experience in early childhood assessment. Data collector country of origin varied, with approximately 14% of the data collectors from Mexico, 27% from Central or South America, and 59% identified as non-Latinos who spoke Spanish with native-like fluency. Data collector backgrounds varied to include university staff, undergraduate and graduate students, and community members who volunteered at participating preschool sites.

Testing procedures

Data collection for *S-IGDIs* occurred in fall, winter, and spring of the 2013–2014 and 2014–2015 academic years. Each student was tested up to three times in their given school year. Standardized criterion tests were administered in the spring of 2014. Assessment occurred one-on-one while the data collector sat at a table across from the child. Schools' setups varied; testing occurred in the classroom, in a room outside of the classroom, or in a quiet adjacent hallway. Total testing interactions lasted about 20 minutes per child, and children received a small trinket of their choice for participating. All data collectors were instructed to only speak Spanish with participants because young SE-DLLs demonstrate interlocutor sensitivity, such that they may be aware of which language to use with different people, and thus may not engage with an adult in Spanish if they perceive that this is not the adult's primary language (Maneva & Genesee, 2002). Participants were randomly assigned to one of the *PS/FS* forms. Those students who were selected for criterion test administration were randomly assigned to each test (*TPAS* or *GRTR-S-R*).

Results

To evaluate *PS/FS*, we aligned results with each research question. During item calibrations, we managed all data by including children who responded (with correct or incorrect response) to 60% or more of the items for a given measure. We used this conservative approach to data inclusion for item calibrations because item calibrations require a meaningful response set to reduce error or noise in the analysis. Thus, children with at least 40% of the total items scored as missing on a measure could have contributed to error in the calibration because their response pattern may have been impacted by systematic lack of understanding the task and other unknown sources of error (e.g., lack of engagement, etc.). All response patterns were included when computing student ability scores after finalized calibrations.

In addition, as noted, each child in both years of the study was assessed up to three times annually and each seasonal response set was treated independently in the calibration. Children's responses were treated as unique each season because they change in their ability levels and respond to a new set of items. If we assumed student abilities to be constant across seasons, then we would violate a basic principle of IRT, in that ability is not constant across item responses because student ability changes at each seasonal administration. As we viewed calibration within season and monitored item parameter drift (or displacement) across season and across year, it became clear that item calibration was stable regardless of how the data (students) were combined. Final concurrent-calibration was completed after all seasonal data were collected, and item responses were pooled for the 104 items. Given the sample of 970 children, the item-level results included a sample size of 2218 (which is no more than three times the total sample of children).

Item-level results and descriptive summaries of child abilities

Item-level results

To respond to the first research question, "What are the item-level statistics of *PS/FS*?", we produced Rasch item level calibrations. Item-level statistics were computed for *PS/FS* measures with Winsteps 3.67 (Linacre, 2008). Item-level output is provided in Table 3 and in Figure 1, the Wright item map. Results include standard Winsteps output variables including item difficulty (b-parameter estimates), standard error of the b-parameter estimate, correlations between the item and total test score (point-by measure correlations, PTMA), and infit and outfit measures of item fit. We expect PTMA values for items that meaningfully contribute to a test to be above .2; lower values are considered not contributing (Bond & Fox, 2015). In addition, output includes *p*-values, which represent the proportion of children who answered the item correctly. Generally, a proportion correct ranging from .2 to .8 is desired. Appropriateness of model fit was investigated based on mean square infit and outfit statistics for item. The infit statistic is more sensitive to irregular responses near the location of item, whereas the outfit statistic is sensitive uniformly across the full range of person abilities (Engelhard, 2013). Infit and outfit statistics between 0.5 and 1.5 are ideal. Items with infit and outfit greater than 2 were eliminated or revised, as is common practice in test design procedures (Haladyna & Rodriguez, 2013).

We also examined the degree to which items differentiated between ability levels. Rasch item-level output in Winsteps provides the mean ability of persons correctly and incorrectly responding to each item. Ideally, items will show adequate differentiation between scores if the mean difference in child ability between each item score unit is at least one logit (more than twice the typical score standard error). Therefore, the average ability for children who answered the item correctly should be at least one logit higher than the average ability for children who answered the item incorrectly. Items identified with weak differentiation of mean abilities between 0 and 1 scores were removed (these results are not reported here).

Finally, we examined item-level statistics to evaluate the contribution of chance or guessing on item-level responses and confirmed the assumptions of the Rasch model. If chance, or the effect of

Table 3. Item level results for S-IGDI first sounds/primeros sonidos.

Item	Distractors		Target	DIF-SEX	DIF-DIAL	Count	P-value	b	SE(b)	INM	OUTM	PTMA
1	–	maracas	/cha/	N	Y (.00)[c]	315	.74	−0.58	.14	1.02	1.10	.30
2	–	cobija	/bi/	N	N	315	.69	−0.31	.14	1.07	1.13	.31
3	–	pato	/ro/	N	N	317	.70	−0.36	.14	0.90	0.82	.42
4	–	queso	/u/	N	N	320	.78	−0.85	.15	0.91	0.75	.38
5	cobija	lápiz	/v/	N	N	312	.57	0.36	.13	0.99	1.03	.41
6	manzana	guitarra	/z/	N	N	316	.47	0.91	.13	0.97	0.97	.48
7	muñeca	tomate	/a/	N	N	308	.42	1.15	.14	1.00	0.99	.48
8	ojo	escuela	/ma/	N	N	317	.55	0.49	.13	0.99	1.01	.42
9	piña	camisa	/o/	Y (.03)[a]	N	1352	.51	0.45	.06	0.93	0.89	.49
10	–	conejo	/tro/	N	N	318	.83	−1.25	.16	1.00	0.90	.27
11	–	plancha	/o/	N	N	758	.67	−0.46	.09	1.00	0.97	.35
12	–	escuela	/o/	N	N	318	.89	−1.80	.19	0.99	0.93	.22
13	–	pelota	/l/	N	N	318	.72	−0.50	.14	0.96	0.90	.36
14	–	arena	/ll/	N	N	755	.69	−0.59	.09	0.99	0.96	.35
15	queso	pato	/f/	N	N	1360	.51	0.41	.06	0.97	0.94	.46
16	árbol	silla	/ja/	N	N	599	.64	−0.10	.10	1.04	1.00	.34
17	bota	rana	/ll/	N	N	597	.63	−0.07	.10	0.96	0.94	.40
18	lobo	nariz	/m/	N	N	591	.44	0.93	.10	1.13	1.17	.36
19	fresa	pera	/li/	Y (.02)[a]	N	590	.56	0.28	.10	1.00	1.03	.40
20	–	leche	/fu/	Y (.01)[a]	N	598	.79	−1.03	.11	0.94	0.79	.33
21	–	bolsa	/ni/	Y (.02)[a]	N	598	.76	−0.82	.10	1.07	1.18	.24
22	–	casa	/s/	N	N	600	.72	−0.54	.10	1.01	0.95	.32
23	–	oso	/in/	N	N	594	.67	−0.27	.10	0.96	0.94	.37
24	cama	mapa	/b/	N	N	276	.54	0.29	.14	1.08	1.06	.35
25	paraguas	vestido	/c/	N	N	279	.56	0.22	.14	1.08	1.07	.34
26	baño	araña	/ra/	N	N	1323	.58	−0.01	.06	1.00	1.04	.38
27	rana	uvas	/cu/	N	N	279	.48	0.58	.14	1.03	1.11	.38
28	taza	cama	/pa/	N	N	281	.69	−0.50	.14	1.06	1.04	.28
29	abrigo	tomate	/or/	Y (.01)[b]	N	280	.59	0.03	.14	1.04	1.08	.33
30	león	silla	/glo/	N	N	283	.70	−0.57	.14	0.96	0.98	.33
31	bota	queso	/gra/	N	N	719	.65	−0.46	.09	1.01	1.02	.33
32	tortilla	cereza	/n/	N	N	576	.47	0.67	.10	1.05	1.06	.44
33	fresa	pollitos	/a/	Y(.01)[b]	N	1616	.53	0.30	.06	1.04	1.03	.41
34	rosa	círculo	/blo/	N	N	1620	.71	−0.74	.06	0.94	0.95	.39
35	arroz	reloj	/ue/	N	N	577	.68	−0.53	.10	0.99	1.01	.38
36	toalla	nubes	/l/	N	N	1608	.54	0.19	.06	1.03	1.02	.40
37	perro	galleta	/e/	N	N	573	.45	0.80	.10	1.00	0.97	.48
38	araña	maiz	/ba/	N	N	1617	.54	0.24	.06	0.88	0.86	.50
39	bus	sofa	/bol/	Y (.01)[a]	N	574	.52	0.39	.10	1.00	1.06	.44
40	almohada	vaca	/m/	N	Y(.02)[d]	290	.42	0.94	.14	1.27	1.36	.34
41	árbol	taza	/jua/	N	N	291	.78	−1.23	.16	1.00	0.98	.35
42	bolsa	globo	/ca/	N	N	296	.65	−0.41	.14	0.92	0.98	.47
43	uvas	plato	/va/	N	N	1339	.53	0.29	.06	1.10	1.13	.37
44	cama	chaqueta	/z/	N	N	295	.55	0.17	.14	0.90	0.91	.52
45	naranja	muñeca	/o/	N	N	295	.45	0.74	.14	1.18	1.25	.39
46	zapato	oreja	/ch/	N	Y (.01)[a]	295	.82	−1.58	.17	0.93	1.10	.37
47	avión	huevo	/s/	N	N	294	.49	0.53	.14	0.78	0.68	.62
48	llave	queso	/f/	N	N	1332	.52	0.33	.06	0.89	0.88	.51
49	guitarra	bandera	/ser/	N	Y(.01)[c]	1591	.56	0.16	.06	1.15	1.19	.33
50	maleta	clavos	/po/	N	N	558	.84	−1.48	.13	0.95	0.76	.36
51	tomate	globo	/sir/	N	N	557	.64	−0.15	.10	0.88	0.77	.51
52	pavo	chicle	/re/	N	N	558	.45	0.92	.10	0.85	0.82	.58
53	tenedor	mariposa	/n/	N	N	558	.54	0.39	.10	1.01	1.00	.46
54	montaña	sombrero	/ga/	N	N	554	.76	−0.93	.11	1.02	1.02	.36
55	lápiz	caballo	/m/	Y (.03)[a]	N	554	.49	0.69	.10	1.18	1.21	.38
56	plato	bicicleta	/s/	N	N	553	.58	0.21	.10	1.01	0.97	.45
57	carro	doctor	/v/	N	N	705	.67	−0.52	.09	1.02	0.84	.41
58	rosa	rodilla	/ta/	N	N	259	.67	−0.14	.15	1.01	0.97	.39
59	pelota	hoja	/be/	N	Y (.02)[a]	257	.49	0.86	.15	1.20	1.32	.33
60	libro	silla	/pla/	N	N	262	.81	−1.06	.17	0.84	0.69	.43
61	niño	hombre	/do/	N	N	265	.74	−0.58	.16	0.87	0.73	.45
62	tigre	leche	/m/	N	N	698	.53	0.26	.09	1.10	1.11	.36
63	gorro	mano	/or/	N	N	261	.69	0.73	.15	1.24	1.26	.32

(*Continued*)

Table 3. (Continued).

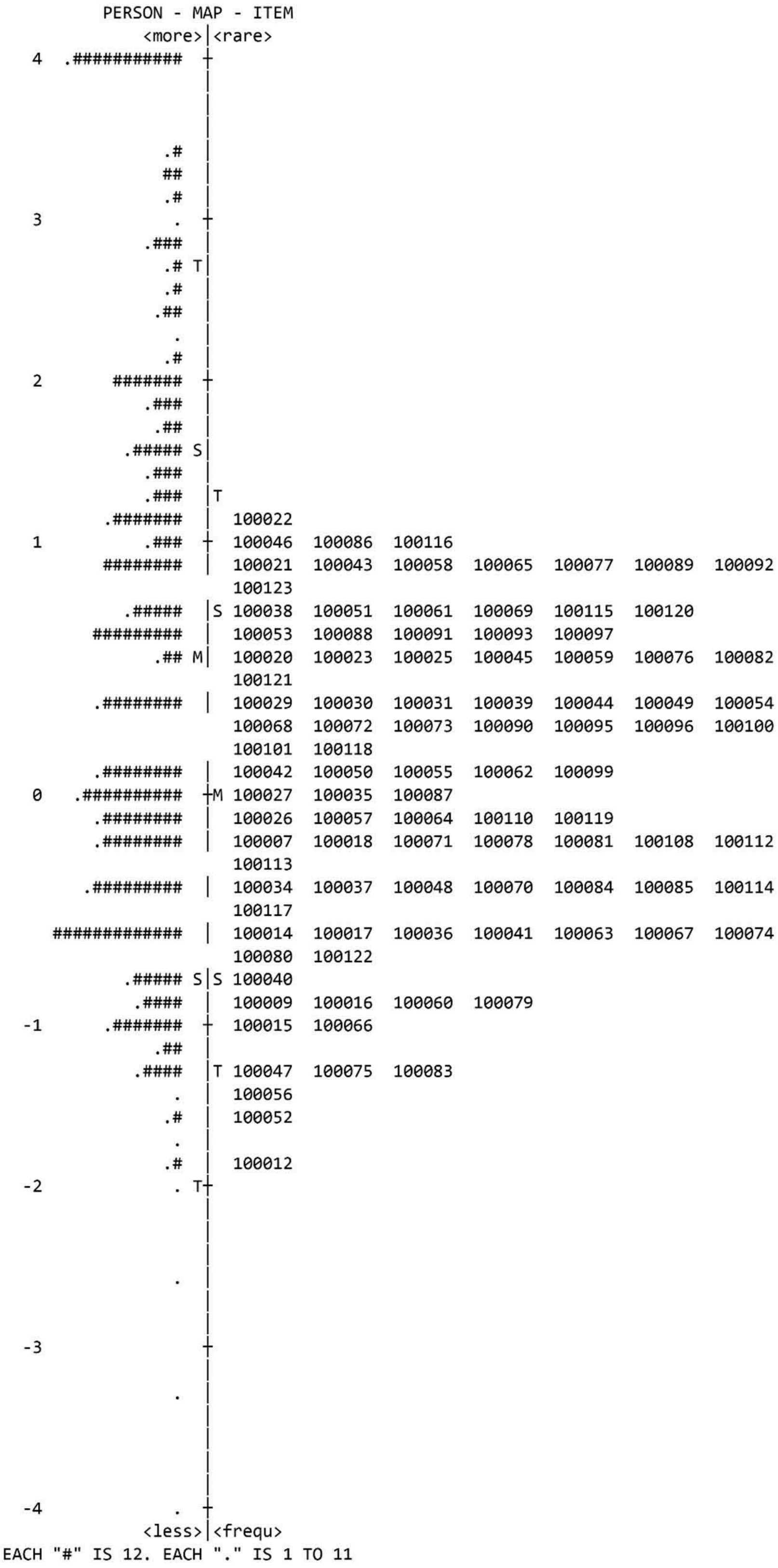
Item	Distractors		Target	DIF-SEX	DIF-DIAL	Count	P-value	b	SE(b)	INM	OUTM	PTMA
64	vestido	ratón	/ue/	N	N	1305	.66	−0.40	.07	0.95	0.92	.41
65	estufa	lluvia	/f/	Y (.02)[b]	N	261	.69	−0.27	.15	1.05	1.06	.35
66	limón	nido	/ba/	Y (.02)[b]	N	1306	.54	0.27	.07	0.93	0.93	.47
67	lampara	enfermera	/gi/	N	N	1297	.55	0.22	.07	0.88	0.84	.50
68	bate	durazno	/ca/	N	N	1294	.68	−0.54	.07	1.10	1.23	.28
69	pie	centavo	/chi/	N	N	264	.84	−1.32	.18	0.95	0.86	.33
70	clavos	avión	/m/	N	N	1309	.50	0.47	.07	1.14	1.20	.34
71	gato	taza	/s/	N	N	257	.49	0.86	.15	0.91	0.86	.52
72	leche	montaña	/ca	Y (.03)[a]	N	261	.69	−0.23	.15	0.99	0.99	.37
73	mesa	plato	/e/	N	N	261	.79	−0.83	.17	1.06	1.35	.26
74	elefante	juguete	/dur/	N	N	100	.40	0.54	.23	0.92	0.90	.44
75	mariposa	tren	/s/	N	N	100	.35	0.81	.24	1.16	1.25	.25
76	chile	zebra	/ar/	N	N	99	.44	0.32	.23	1.01	1.01	.36
77	paleta	globo	/f/	N	N	104	.39	0.50	.22	0.99	0.99	.35
78	venda	oso	/ca/	N	N	104	.46	0.17	.22	0.88	0.87	.45
79	lobo	abeja	/pa/	N	N	104	.44	0.26	.22	0.95	0.92	.40
80	trampolín	bloque	/f/	N	N	103	.45	0.24	.22	0.95	1.00	.37
81	abeja	botón	/cur/	N	N	152	.60	−0.21	.18	1.02	0.98	.35
82	globo	fleche	/be/	N	N	154	.62	−0.30	.18	0.89	0.88	.43
83	cama	playa	/o/	N	N	152	.61	−0.22	.19	0.86	0.78	.47
84	silla	carro	/ga/	N	N	154	.64	−0.40	.19	0.87	0.80	.44
85	nubes	queso	/r/	N	N	149	.44	0.66	.19	1.08	1.06	.38
86	desierto	zanahoria	/es/	N	N	101	.40	1.01	.25	1.19	1.20	.42
87	tambor	dólar	/pe/	N	N	103	.65	−0.41	.23	1.01	0.89	.39
88	pelo	llave	/f/	N	N	103	.51	0.34	.23	0.96	1.05	.48
89	casa	mano	/z/	N	N	102	.59	−0.09	.23	0.84	0.73	.52
90	doctor	gato	/o/	N	N	103	.45	0.73	.24	1.03	1.01	.49
91	pingüino	centavo	/ju/	N	N	102	.50	0.43	.24	0.96	0.89	.51
92	rata	olla	/pa/	N	N	103	.68	−0.57	.24	1.00	1.37	.35
93	mapache	lluvia	/a/	N	N	101	.42	0.88	.25	1.21	1.23	.40

[a] The DIF contrast favored girls, such that the item produced an easier difficulty value than for boys.
[b] The DIF contrast favored boys, such that the item produced an easier difficulty value than for girls.
[c] The DIF contrast favored Mexican dialect, such that the item produced an easier difficulty value for Mexican Spanish than the combined Spanish group.
[d] The DIF contrast favored combined dialects, such that the item produced an easier difficulty value for the combined Spanish group than the Mexican Spanish group.

Note. DIF-SEX is the DIF contrast for sex, boys vs girls. DIF-DIAL is the DIF contrast for dialect, Mexican vs combined Spanish dialects. The parenthetical value in the DIF columns is the statistical significance of the DIF result (p-value). Count is the number of children responding to the item. P-value is the proportion responding correct. b indicates the Rasch b-parameter estimate (item location in logits). SE(b) is the standard error of b-parameter estimate. INM is the Infit mean square. OUTM is the Outfit mean square. PTMA is the point measure correlation as an index of item discrimination.

guessing, is playing a major role in calibration, it will produce poor fit and poorly characterize performance on the construct. Item-total correlations are attenuated when random responses (e.g., guessing) have a significant presence. If guessing is present, infit and outfit will be inappropriately high. However, since no items with poor fit were produced in the refined data set, the finalized item sets showed no detectable impact of guessing. In addition, researchers have documented long-standing controversies regarding random responses (pseudo-guessing) in a 3PL model and the value of its application (Chiu & Camilli, 2013). We posit the third parameter is not a necessary inclusion in the model given the item review process we specified.

All item-level calibration results are provided in Table 3. Initial calibration efforts showed that of the original 104 items, 11 items were not scalable because of too-small sample size and were removed from the item bank. For these items, too few item responses were collected either because the new item was added to a block late in the academic year (and therefore did not obtain enough responses) or because the new item yielded a disproportionate amount of responses listed as no response/don't know. Of the 93 scaled items, results indicated all items had adequate PTMA statistics and infit or outfit values. Six items had higher than desired *p*-values ranging from .81 to .89.



```
          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                +
                    |
                 .  |
                    |
                    |
                    |
                    |
  -4             .  +
              <less>|<frequ>
EACH "#" IS 12. EACH "." IS 1 TO 11
```

Figure 1. *Primeros sonidos/first sounds* wright item map.

Summary-level descriptives

We also computed summary-level statistics for *PS/FS* using Winsteps. In a Rasch model, the mean item difficulty is fixed at 0 logits. Item difficulties ranged from −1.79 to 1.14 as illustrated by the b-parameter variable in Table 3. Item standard errors are also presented in Table 3 and ranged from 0.05 to 0.25. The distribution of child ability scores appropriately matched the item difficulty distribution (see Figure 1, the Wright item map).

To further evaluate measure quality, we examined differential item functioning (DIF) based on sex and country of origin contrasting Mexican and non-Mexican families. Items without DIF more aptly represent the construct and are not differentially attributed to other factors, such as dialectical differences or sex. In addition, DIF is used to evaluate the extent to which measurement invariance holds across a construct-relevant group characteristic. DIF was calculated using Winsteps 4.1.0. Winsteps separately calibrates items for each dichotomously defined group (e.g., gender was compared as male vs female; dialectal representation was compared as Mexican vs all other dialects) through pairwise comparison. Each calibrated item measure (difficulty) is then statistically compared to determine if the contrast of item measures is statistically significantly different. Empirical tests reported include the Rasch-Welsch contrast statistic and the Mantel-Hanzel statistic. Items with statistically significant contrasts are flagged for potential DIF and review.

Regarding dialect, DIF was computed comparing Mexican children to non-Mexican children. Of our sample of 970, 355 children were included in the DIF analysis. Families who reported one parent as Mexican and one parent as non-Mexican (e.g. Puerto Rican, Guatemalan etc.) were excluded from this analysis (n = 97), as well as families who did not report their ethnicity (n = 13) and children who did not have scores on the *PS/FS* measure for at least two of the three seasons. The Mexican group included 142 cases and the non-Mexican group included 213 cases (represented by combined ethnicities from other dialectical groups). Results indicated statistically significant contrasts (using the Rasch-Welch test) were present for five items at $p < .01$; three contrasts favored Mexican children and two contrasts favored non-Mexican children (see Table 3). These five items were removed from the item set as a result of the DIF analysis and content-related review of the items.

Regarding sex, DIF was computed comparing boys and girls. Of our sample of 970, 732 families reported on their children's sex. Within this group 348 were boys and 384 were girls. Results indicated statistically significant contrasts were present for 11 items at $p > .01$; five contrasts favored boys and six contrasts favored girls (see Table 3). These 11 items were removed from the item set as a result of the DIF analysis and content-related review of the items.

Child ability descriptives

To answer the research question, "To what degree do *PS/FS* items appropriately measure child PA abilities?" we used item calibrations to produce seasonal ability scores for the 970 children in our sample as depicted in Figure 1. Results indicated the items appropriately match student ability levels, with 67% of students falling within the range between the easiest and most difficult items. The Rasch mean ability score for *PS/FS* was 0.54 (SD = 1.49). This statistic represents average ability of the children who interacted with items from the measure, relative to the average item location of 0.0. Larger Rasch mean scores represent higher ability. The person reliability was .78 for *PS/FS*. In a Rasch model, person reliability is often compared to a standard reliability coefficient in a classical test theory model, where the person reliability is more conservative (Linacre, 2008). However, a better approach to understanding how to estimate the consistency of the measure is to examine the conditional standard error of measurement (CSEM) for each ability level. Given that ability is a continuum, we report CSEM by ability levels with a LOESS smoothing line. The smallest CSEM is 0.42 and the largest, at the tail, is 1.81 logits (see Figure 2). We observed relatively small CSEM between −2.5 and 2.5 logits on the ability scale, where the average CSEM is 0.52 logits. Eighty-four percent of children's performances fell within this range.

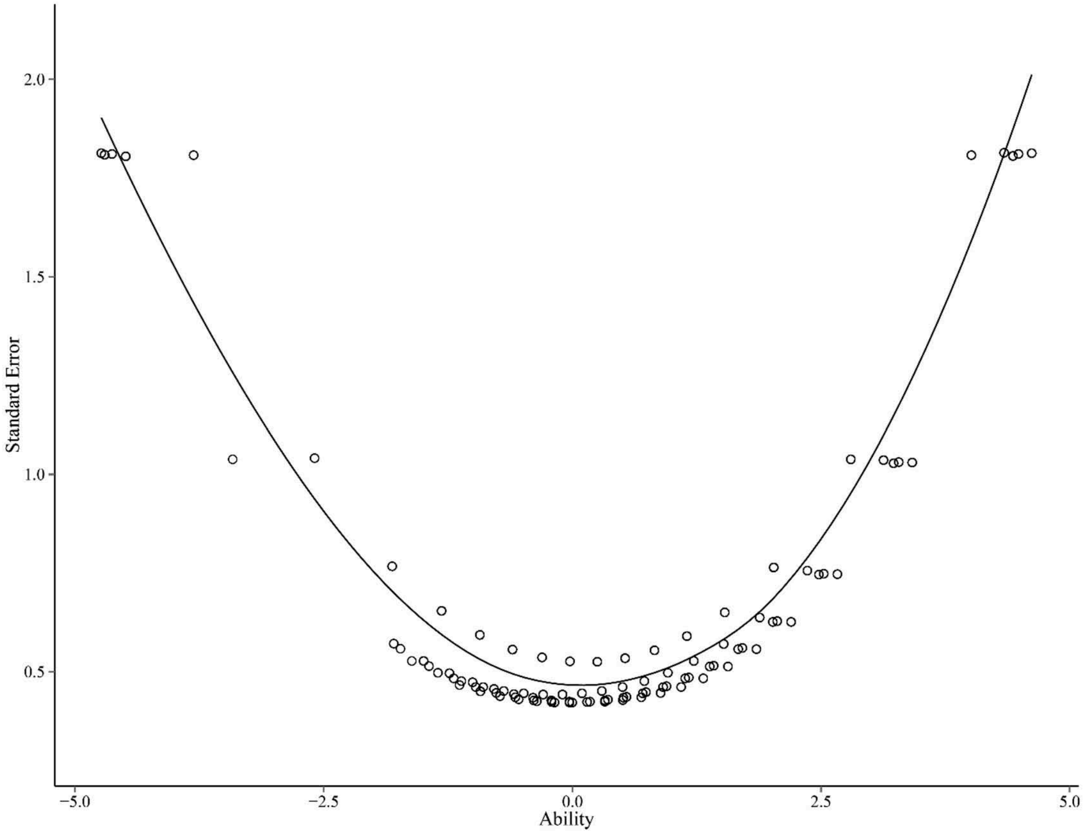


Figure 2. Standard error of measurement LOESS curve for *primeros sonidos/first sounds.*

Model fit and unidimensionality

To make sense of the tenability of the results, we must present additional results of model-fit and assumptions. The Rasch model is most appropriately characterized as a measurement design model, not a data analysis model. One typically does not go through a model fitting exercise with the Rasch model to identify the best fitting model for analyzing item response data – and this is not our purpose. Our purpose in selecting the Rasch model was to place a strong set of assumptions on item development and requirements for item retention. These requirements are naturally relaxed, as items never perfectly fit any particular model, so that items are more-likely-than-not fitting the selected measurement design model. However, to satisfy the empirical approach to model selection, items and persons were calibrated with the Rasch, 2PL (allowing item discrimination to vary), and 3PL (allowing item discrimination and lower asymptote to vary) models using Bilog Version 3 (Du Toit, 2003; Zimowski, Muraki, Mislevy, & Bock, 2003).

To simplify the results of these multiple models, employing the same data, we examined correlations among person locations (person ability scores) and item locations (b-parameter estimates). The correlations for person locations were .98 (Rasch vs. 2PL), .98 (Rasch vs. 3PL), and .99 (2PL vs. 3PL). The correlations among item locations were .99 (Rasch vs. 2PL), .96 (Rasch vs. 3PL), and .97 (2PL vs. 3PL). We also noted that the standard errors of the item location parameters were smaller for the Rasch model (0.13), slightly larger for the 2PL model (0.15), and larger yet for the 3PL model (0.24), suggesting relatively more precise estimation of item locations with the Rasch model.

An important assumption in the Rasch model, and all unidimensional IRT models, is that 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), 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. It is generally agreed that multiple indicators of fit should be examined (Muthén & Muthén, 2012b). The general criteria for model-data fit are as follows. Model fit is indicated such that: RMSEA < .05 is Good Fit; CFI > .95 is Good Fit; CFI > .90 is Adequate Fit; TLI > .95 is Good Fit; TLI > .90 is Adequate Fit (Brown, 2006). Based on these criteria, each form of the PA measure resulted in good fit to a unidimensional model, meeting the unidimensionality assumption of the Rasch model (see Table 4).

A more typical method to examine the degree to which the data met the Rasch assumption of unidimensionality, includes examination of item level infit and outfit statistics and scree plots of eigenvalues using a principal component analysis (PCA) of Rasch residuals (components of the item variances and covariances not explained by the Rasch model). Infit and outfit statistics were appropriate, with no items producing values greater than 1.37 or less than 0.68. Residuals are presented in statistical contrasts after the primary variance due to items and abilities has been accounted. For a secondary dimension to exist, it needs to exceed a value of 2 to eclipse noise in the data (Linacre, 2008). Results from *PS/FS* are presented in in Table 5. Comparing the ratio of Rasch variance by items (6%) with the unexplained variance of the first contrast (1.5%) indicates that the Rasch dimension is about 4 times as large as the second dimension. An Eigenvalue of 1.9 further suggests the presence of a single dimension.

Associations among measures

To respond to the research question, "To what extent is *PS/FS* associated with other standardized measures of Spanish early language and literacy?" we computed Pearson correlations between *PS/FS* child ability Rasch scores (where scores were computed without the 11 items dropped because of poor item level statistics) and *TPAS* initial sounds subscale and *GRTR-S-R* total scores. For the *TPAS* we hypothesized we would see weak correlations with the Rhyming and final sound subtests as Rhyming is not a salient part of the PA construct in Spanish and final sounds have few, if any studies

Table 4. Confirmatory factor analysis results for primeros sonidos/first sounds by form.

Form	# of Items	# of Children	RMSEA	CFI	TLI
101	25	315	.027	.956	.952
102	25	279	.031	.945	.940
103	25	298	.031	.979	.977
104	25	266	.035	.949	.944
105	15	1051	.042	.963	.957

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

Table 5. PCA results for primeros sonidos/first sounds (n = 93).

	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%

Note. The variance explained by measures is also partitioned into the variance explained by persons and variance explained by items. The total variance is partitioned into variance explained by measures and unexplained variance.

to substantiate their contribution to the construct. Therefore, we were particularly interested in correlations between *PS/FS* and the *TPAS* initial sounds subtest. Descriptive results on the *TPAS* indicated raw scores ranged from 0–17 (maximum score potential of 30), with 23% of students achieving a raw score of 0 and a mean raw score of 9.79. Results indicated the correlation between *PS/FS* and *TPAS* initial sounds demonstrated virtually no association, at r = .005. The correlation between *PS/FS* and *GRTR-S-R* was also low, at r = .24, as expected.

Discussion

We summarized the iterative development process and the collection of validity evidence for the new *S-IGDI* PA measure: *Primeros Sonidos/First Sounds*. *PS/FS* is a measure that joins only a few existing preschool early literacy and language tools in the field that were developed without translating an existing English measure. Translated measures may not appropriately capture the identified construct because they cannot address conceptual and theoretical differences between languages. Results from this study contribute evidence to support the importance of measuring Spanish early literacy skills with empirically and conceptually sound assessment tools. Specifically, in the development of *PS/FS*, there were unique features relevant to how Spanish develops, including the relative difficulty of the phoneme versus syllable, and the salience and availability of each syllable or phoneme in Spanish language.

Results suggest *PS/FS* demonstrated adequate item-level quality; the items produced adequate infit and outfit statistics, and a conservative DIF analysis indicated 16% of items showed statistical DIF across either contrast. Dimensionality testing offered evidence to align with previous findings that suggest Spanish phonological awareness is a unitary construct.

Our results also revealed low correlations with an existing measure of Spanish phonological awareness. Even though we hypothesized these correlations would be low, this finding is worth additional attention. The *TPAS* is the only standardized test available that specifically tests PA in preschool-age Spanish-speaking students; however, we had deep reservations about the validity of its inferences. First, the *TPAS* manual provides limited information about how items were developed and reports that norms were not collected with 4-year-olds (which represents over half of our sample), suggesting any score interpretation for four-year-old DLLs is an extrapolation of older student's performance. Second, the *TPAS* items are yes/no, agree/disagree protocols and have no discontinue criteria. A test with 100% of items in this format may have a less robust alignment with performance on Spanish PA because children may have limited opportunity to illustrate knowledge of the construct and children are encouraged to provide a response even if they express that they don't know the answer. Third, 23% of students achieved a score of zero on the *TPAS*. This skewed distribution may have impacted the correlation results due to the truncated variance present in *TPAS* scores. Finally, the *TPAS* includes subtests that do not align with the Spanish PA construct. Given these constraints, it may be the case that the low correlations between *PS/FS* and the *TPAS* may be due to limited validity evidence of the *TPAS* measure.

Our findings indicate *PS/FS* may be a promising new measure for at least two reasons. First, as noted, the measure was designed by attending to unique facets of Spanish early literacy and language development. Evidence indicates assessment for SE-DLLs must consider the unique linguistic features of Spanish and the cultural differences amongst Spanish-speaking populations in the US (Peña, 2007). Using these findings as a catalyst, we devoted specific attention to cultural and linguistic variables most salient to Spanish language to design the measure. Second, *PS/FS* demonstrated robust empirical results in the Rasch model at the item level. Few early literacy assessments attend to the quality of assessments at this level, and we are aware of no publicly-available Spanish early literacy screening tools that consider how Spanish develops and evaluates item-level statistics. The dearth of Spanish measures in the field may be indicative of the lag between evidence supporting the need to measure Spanish early literacy development and efforts to distill such science into practical tools that can be used in applied early childhood settings.

It is important to note, however, that the process of collecting validity evidence to support the use of *PS/FS* is not yet complete. Although we described the development and calibration process, it is not enough to simply develop a bank of items. How such measures and data are used is an important part of improving instructional practices for SE-DLLs. The *S-IGDI PS/FS* item bank was developed to be applied within a multi-tiered system of support (MTSS) to construct sets of items for screening, which, in turn, can inform data-based decision making and differentiated instruction. In complement to the work presented here, we also have pursued establishing benchmarks for screening performance in an MTSS model (see Wackerle-Hollman et al., under review), growth metrics (Wackerle-Hollman, Durán, Palma, Miranda, & Batz, 2018), and pragmatic application of English and Spanish IGDI scores (see Wackerle-Hollman et al., 2016). Moreover, high-quality Spanish phonological awareness tasks offer much potential in the field. The item sets can support ongoing research to explore associations between Spanish early intervention and child outcomes to promote long-term reading proficiency in English and Spanish. With well-designed measures of Spanish early literacy and language that have evidence to support claims of validity, educators will be better prepared to adjust instruction to meet their students' unique needs. Further, improved assessments can lead to improved intervention, which in turn can have the potential to improve the reading outcomes of SE-DLLs in US classrooms.

Limitations

Although the evidence provided here constitutes a valuable contribution to emerging research on Spanish early literacy and language PA measures, a few limitations could be addressed in the future. First, access to information about prior preschool experience was limited. Such experiences may have impacted the level of phonological awareness skill SE-DLLs bring to their Pre-K year, thus affecting the nature of their responses. Future studies could benefit from analyses that consider covariates that may impact child-level performance on such measures including prior preschool experience, dosage of such an experience, and language of instruction in which such experiences were provided. These variables could be addressed through DIF or entered as covariates in regression models that predict performance over time. Similarly, experience with such tests may play a role in child responses processes.

Second, the *S-IGDI PS/FS* results provided here only minimally contribute to aiding in understanding dialectical differences and level of language proficiency, two of the leading conceptual issues in Spanish assessment (Wackerle-Hollman et al., under review). During the two years that these studies occurred, we made an effort to include four dialectical groups: Mexican Spanish, Caribbean Spanish, Central and South American Spanish, and Mixed Spanish representation groups, because the same image may have a different word associated with it unique to each dialect. However, our available sample only permitted DIF comparisons between Mexican and non-Mexican children. In future studies, the field could benefit from more detailed comparisons of dialectical groups on Spanish early literacy and language measures like *PS/FS* to illustrate the degree to which performance differences attributed to dialect may be confounded by item functioning.

In sum, our results showcase the process of development and validation of the *S-IGDI PS/FS* as a PA measure of Spanish language and early literacy. Our results support ongoing efforts to build a practical, contemporary model of Spanish language and early literacy assessment for SE-DLLs. This work, along with ongoing research on dual language approaches to intervention, professional development for teachers of SE-DLLs, and continued exploration of the association between Spanish and English language and literacy acquisition are important contributions to this salient and quickly expanding need in early childhood education.

Conflict of Interest

Dr. Alisha Wackerle-Hollman holds equity in, and receives a consulting income from Early Learning Labs LLC, a company which has a license from the University of Minnesota to commercialize Individual Growth and Development Indicators (IGDIs). The University of Minnesota also has equity and royalty interests in ELL. These relationships have been reviewed and managed by the University of Minnesota in accordance with its conflict of interest policies.

Acknowledgments

This work was supported in part by grant R305A120449, Research and Development of Spanish Individual Growth and Development Indicators (S-IGDIs): 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 metro area, central Florida, rural Utah and Idaho, and northwestern California. 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. 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.

Funding

This work was supported by the Institute of Education Sciences [R305A120449]

References

Andrich, D. (2004). Controversy and the rasch model: A characteristic of incompatible paradigms? *Medical Care*, *42* (1), 7–16. doi:10.1097/01.mlr.0000103528.48582.7c

Anthony, J. L., & Lonigan, C. J. (2004). The nature of PA: Converging evidence from four studies of preschool and early grade school children. *Journal of Educational Psychology*, *96*(1), 53–55. doi:10.1037/0022-0663.96.1.43

Anthony, J. L., Williams, J. M., Duran, L. K., Gillam, S. L., Liang, L., Aghara, R., … Landry, S. H. (2011). Spanish PA: Dimensionality and sequence of development during the preschool and kindergarten years. *Journal of Educational Psychology*, *103*(4), 857–876. doi:10.1037/a0025024

Barrera, M., & Liu, K. K. (2010). Challenges of general outcome measurement in the RTI progress monitoring of linguistically diverse exceptional learners. *Theory into Practice*, *49*, 273–280. doi:10.1080/00405841.2010.510713

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

Bond, T. G., & Fox, C. M. (2015). *Applying the rasch model: Fundamental measurement in the human sciences*. New York, NY: Routledge.

Branum-Martin, L., Mehta, P., Fletcher, J. M., Carlson, C. D., Ortiz, A., Carlo, M., & Francis, D. J. (2006). Bilingual PA: Multilevel construct validation among Spanish-speaking kindergarteners in transitional bilingual education classrooms. *Journal of Educational Psychology*, *98*, 170–181. doi:10.1037/0022-0663.98.1.170

Brown, J., & Doolittle, J. (2008). A cultural, linguistic, and ecological framework for response to intervention with English language learners. *Teaching Exceptional Children*, *40*(5), 66–72. doi:10.1177/004005990804000509

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

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. doi:10.1044/0161-1461(2007/026)

Chiu, T. W., & Camilli, G. (2013). Comment on 3PL IRT adjustment for guessing. *Applied Psychological Measurement*, *37*(1), 76–86. doi:10.1177/0146621612459369

Christ, T.; Associates. (2013). *Formative Assessment for Teachers (FAST) aReading*. Minneapolis, MN: FastBridge Learning. Retrieved from www.fastbridge.org

de Ayala, R. J. (2013). *Theory and practice of item response theory*. New York, NY: Guilford Press.

Dickinson, D. K., McCabe, A., Clark-Chiarelli, N., & Wolf, A. (2004). Cross-language transfer of PA in low-income Spanish and English bilingual preschool children. *Applied Psycholinguistics*, *25*, 323–347. doi:10.1017/S0142716404001158

Du Toit, M. (Ed.). (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Lincolnwood, IL: Scientific Software International.

Engelhard, G. (2013). *Invariant measurement: Using rasch models in the social, behavioral, and health sciences*. New York, NY: Routledge.

Farver, J. A. M., Lonigan, C. J., & Eppe, S. (2009). Effective early literacy skill development for young Spanish-speaking English language learners: An experimental study of two methods. *Child Development*, *80*(3), 703–719. doi:10.1111/j.1467-8624.2009.01292.x

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

Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children*, *57*, 488–499. doi:10.1177/001440299105700603

Garcia, E., & Jensen, B. (2009). *Early educational opportunities for children of hispanic origins* (Social Policy Rep. No. 23). Ann Arbor, MI: Society for Research in Child Development. doi:10.1002/j.2379-3988.2009.tb00059.x

Genesee, F., Geva, E., Dressler, C., & Kamil, M. (2006). Synthesis: Cross-linguistic relationships. In D. August & T. Shanahan (Eds.), *Developing literacy in second-language learners: Report of the national literacy panel on language-minority children and youth*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Geva, E., & Genesee, F. (2006). First-language oral proficiency and second-language literacy. In D. August & T. Shanahan (Eds.), *Developing literacy in second-language learners: Report of the national literacy panel on language-minority children and youth*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

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

Good, R., Baker, D., Knutson, N., & Watson, J. (2007). *Indicadores dinámicos del éxito en la lectura*. Eugene, OR: Dynamic Measurement Group.

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. doi:10.1037/a0031780

Goodrich, J. M., Lonigan, C. J., & Farver, J. M. (2014). Children's expressive language skills and their impact on the relation between first-and second-language phonological awareness skills. *Scientific Studies of Reading*, *18*(2), 114–129. doi:10.1080/10888438.2013.819355

Gorman, B., & Gillam, R. (2003). PA in Spanish: A tutorial for speech-language pathologists. *Communication Disorders Quarterly*, *25*(1), 13–22. doi:10.1177/15257401030250010301

Haager, D. (2007). Promises and cautions regarding using response to intervention with English language learners. *Learning Disability Quarterly*, *30*(3), 213–218. doi:10.2307/30035565

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

Hammer, C. S., Jia, G., & Uchikoshi, Y. (2011). Language and literacy development of dual language learners growing up in the United States: A call for research. *Child Development Perspectives*, *5*(1), 4–9. doi:10.1111/j.1750-8606.2010.00140.x

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. doi:10.1044/0161-1461(2007/025)

Hosp, J., Hosp, M., & Dole, J. (2011). Potential bias in predictive validity of universal screening measures across disaggregation subgroups. *School Psychology Review*, *40*(1), 108.

Jackson-Maldonado, D., Thal, D. J., Fenson, L., Marchman, V. A., Newton, T., & Conboy, B. (2003). *MacArthur inventarios del desarollo de habilidades comunicativas*. Baltimore, MD: Brookes.

Landry, S., Assel, M., Gunewig, S., & Swank, P. (2004). *Circle phonological awareness language and literacy screener*. Houston, TX: Ridgeways.

Leafstedt, J. M., & Gerber, M. M. (2005). Crossover of phonological processing skills a study of Spanish-speaking students in two instructional settings. *Remedial and Special Education*, *26*(4), 226–235. doi:10.1177/07419325050260040501

Linacre, J. M. (2008). *Winsteps (Version 3.68)*. [Computer Software]. Chicago, IL: Winsteps.

Lonigan, C. J. (2003). *Technical report on the development of the NCLD Spanish-language Get Ready to Read! Screening tool*. Retrieved from http://www.getreadytoread.org

López, L. M., & Greenfield, D. B. (2004). The cross-language transfer of phonological skills of hispanic head start children. *Bilingual Research Journal*, *28*, 1–18. doi:10.1080/15235882.2004.10162609

Lopez, M. A., & Gonzalez-Barrera, A. (2013). *What is the future of Spanish in the United States?* Washington, DC: Pew Research Center. Retrieved from http://www.pewresearch.org/fact-tank/2013/09/05/what-is-the-future-of-spanish-in-the-united-states/

Maneva, B., & Genesee, F. (2002). Bilingual babbling: Evidence for language differentiation in dual language acquisition. In *Boston University Conference on Language Development 26 Proceedings* (pp. 383–392). Somerville, MA: Cascadilla Press.

Mathes, P., & Torgesen, J. (2005). *Istation's Indicators of Progress (ISIP) Español* technical report, Version 4. Dallas, TX: Istation.

McConnell, S. R., Wackerle-Hollman, A. K., & Bradfield, T. A. (2014). Early childhood literacy screening. In R. Kettler, T. Glover, C. Albers, & K. A. Feeney-Kettler (Eds.), *Universal screening in educational settings: Identification, implications, and interpretation* (pp. 141–170). Washington, DC: American Psychological Association.

Melby-Lervåg, M., & Lervåg, A. (2011). Cross-linguistic transfer of oral language, decoding, PA and reading comprehension: A meta-analysis of the correlational evidence. *Journal of Research in Reading*, *34*, 114–135. doi:10.1111/j.1467-9817.2010.01477.x

Muter, V., Hulme, C., Snowling, M. J., & Stevenson, J. (2004). Phonemes, rimes, vocabulary, and grammatical skills as foundations of early reading development: Evidence from a longitudinal study. *Developmental Psychology*, *40*, 665–681. doi:10.1037/0012-1649.40.5.665

Muthén, L. K., & Muthén, B. O. (2012a). *Mplus (Version 7.0)*. [Computer Software]. Los Angeles, CA: Authors.

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

National Center for Education Statistics. (2013). *The nation's report card: A first look: 2013 mathematics and reading* (NCES 2014-451). Washington, DC: U.S. Department of Education, Institute of Education Sciences.

National Early Literacy Panel. (2008). *Developing early literacy: Report of the National Early Literacy Panel*. Jessup, MD: National Institute for Literacy. Retrieved from http://www.nifl.gov/publications/pdf/NELPReport09.pdf

Peña, E. D. (2007). Lost in translation: Methodological considerations in cross-cultural research. *Child Development*, *78*, 1255–1264. doi:10.1111/j.1467-8624.2007.01064.x

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

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

Riccio, C. A., Imhoff, B., Hasbrouck, J. E., & Davis, G. N. (2004). *Test of phonological awareness in Spanish*. Austin, TX: Pro-Ed.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*(2), 3–13. doi:10.1111/j.1745-3992.2005.00006.x

Sick, J. (2010). Rasch measurement in language education part 5: Assumptions and requirements of rasch measurement. *JALT Testing & Evaluation SIG Newsletter*, *14*(2), 23–29. Retrieved from http://jalt.org/test/news-date.htm

Smith, R. (1994). A comparison of the power of rasch total and between-item fit statistics to detect measurement disturbances. *Educational and Psychological Measurement*, *54*(1), 42–55. doi:10.1177/0013164494054001005

Swanson, L. H., Rosston, K., Gerber, M., & Solari, E. (2008). Influence of oral language and phonological awareness on children's bilingual reading. *Journal of School Psychology*, *46*(4), 413–429. doi:10.1016/j.jsp.2007.07.002

Torgeson, J. K., & Mathes, P. G. (2000). *A basic guide to understanding, assessing, and teaching PA*. Dallas, TX: Pro-ed.

U.S. Census Bureau. (2010). *Hispanic or Latino by type: 2010*. Washington, DC: Author. Retrieved from http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=DEC_10_SF1_QTP10&prodType=table

U.S. Department of Health and Human Services. (2009). *Dual language learning: What does it take. Head Start dual language report*. Washington, DC: U.S. Office of Head Start. Retrieved from https://drupaldemo3.cleverex.com/sites/default/files/pdf/dual-language-learning-what-does-it-take.pdf

U.S. Department of Health and Human Services, Administration for Children and Families, Office of Head Start. (2013). *Report to congress on dual language learners in head start and early head start programs*. Washington, DC: Author. Retrieved from http://www.acf.hhs.gov/sites/default/files/opre/report_to_congress.pdf

Wackerle-Hollman, A., Brunner, S., Duran, L., McConnell, S., Palma, J., Kohlmier, T., … Rodriguez, M. (2012). *Technical Report #1: The Development of Early Literacy Skills in Bilingual and Spanish-speaking Preschool-age Children: A Literature Review*. Retrieved from http://innovation.umn.edu/igdi/wp-content/uploads/sites/37/2018/08/TechnicalReport1.pdf

Wackerle-Hollman, A., Durán, L., Palma, J., Miranda, A., & Batz, R. (2018). *The performance of Spanish-English bilingual preschoolers on Spanish and English progress monitoring measures*. Paper presented at the Society for the Scientific Study of Reading Annual conference, Brighton, UK.

Wackerle-Hollman, A., Durán, L., Rodriguez, M., Brunner, S., Palma, J., Kohlmeier, T., & Callard, C. (2016). *Spanish individual growth and development indicators technical manual*. Minneapolis, MN: University of Minnesota.

Wackerle-Hollman, A., Durán, L., Rodriguez, M., Brunner, S., Palma, J., & Raikes, A. (under review). Understanding preschool Spanish early literacy and language skills in the context of bilingual language exposure profiles. *International Journal of Bilingualism*.

Wildsmith, E., Alvira-Hammond, M., & Guzman, L. (2016). *A national portrait of hispanic children in need*. Bethesda, MD: National Research Center on Hispanic Children and Families. Retrieved from http://www.childtrends.org/wp-content/uploads/2016/02/2016 15HispChildrenInNeed.pdf

Wildsmith, E., Scott, M. E., Guzman, L., & Cook, E. (2014). *Family structure and family formation among low-income hispanics in the U.S.* Bethesda, MD: National Research Center on Hispanic Children and Families. Retrieved from http://www.hispanicresearchcenter.org/wp-content/uploads/2018/04/Family-Formation-Brief-V2.pdf

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

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *Bilog-MG (Version 3.0)*. [Computer Software]. Lincolnwood, IL: Scientific Software International.