Early Childhood Research Quarterly 48 (2019) 155–172

Contents lists available at ScienceDirect

Early Childhood Research Quarterly

Individual Growth and Development Indicators-Español: Innovation in the development of Spanish oral language general outcome measures

Lillian K. Durán[a,*], Alisha K. Wackerle-Hollman[b], Theresa L. Kohlmeier[c], Stephanie K. Brunner[b], Jose Palma[b], Chase H. Callard[d]

[a] University of Oregon, 5261 University of Oregon, Eugene, OR 97403, USA
[b] University of Minnesota 56 E River Rd 357, Ed Sci Bldg Minneapolis, MN 55455, USA
[c] University of Wisconsin-Stout 115 Heritage Hall 712 Broadway St S, Menomonie, WI 54751
[d] Utah State University 2800 Old Main Hill Logan, UT 84322-2800

ARTICLE INFO

Article history:
Received 6 June 2017
Received in revised form 12 January 2019
Accepted 6 February 2019
Available online 15 April 2019

Keywords:
Preschool
Assessment
Spanish
Universal screening
Language
Literacy

ABSTRACT

The population of Spanish-speaking preschoolers in the United States continues to increase and there is a significant need to develop psychometrically sound early language and literacy screening measures to accurately capture children's ability in Spanish. In this paper, we describe the innovative design and calibration process of the new Individual Growth and Development Indicators-Español (IGDIs-E). We developed and tested two Spanish oral language measures: Identificación de los Dibujos/Picture Naming and Verbos (Expresivo)/Expressive Verbs with 976 Spanish-speaking preschoolers (4–5-years old; 50% female) across five states. Children were tested in Spanish in fall, winter, and spring across two academic years. Results provide evidence that the new IGDIs-E are psychometrically sound with no significant bias between genders and dialects of Spanish spoken in the United States. Cumulative results, the utility of the final measures, and the implications for data-based decision making with Spanish-speaking preschoolers is discussed.

The population of Spanish-speaking preschoolers continues to dramatically increase, and Latinos are currently the fastest-growing population in the United States. Currently 25% of children under the age of 5 are Latino (Murphey, Guzman, & Torres, 2014; U.S. Census Bureau, 2013) and 1 in 4 are growing up in poverty (Children's Defense Fund, 2017). The majority of Latino children are U.S. citizens (90% in 2013), but 70% of their families originate from Mexico, with the remaining 30% from the Caribbean, Central and South America, and Europe (Murphey et al., 2014). Large and persistent gaps in reading achievement are well documented between Spanish-speaking children and their monolingual English-speaking peers (Figueras-Daniel & Barnett, 2013; Reardon & Galindo, 2009). Although scores for both groups have increased since 2002, achievement gaps between White and Latino students have essentially remained the same for the last 30 years due to multiple factors such as ineffective educational approaches, higher rates of poverty, and challenges associated with speaking a non-majority language (Goldenberg, 2008; Hemphill, & Vanneman, 2010; National Center for Education Statistics, 2015). What is clear is that reading performance for Latino students is not increasing at the rate necessary to promote equity in literacy outcomes, which is at the core of educational achievement in the US (August & Shanahan, 2006; Goldenberg, 2008; Mancilla-Martinez & Lesaux, 2010).

1. Associations between Spanish and English oral language ability

Spanish-speaking dual language learners (DLLs) frequently enter kindergarten with low language abilities in both English and Spanish (Davison, Hammer, & Lawrence, 2011; Hoff, 2013; Jackson, Schatsneider, & Leacox, 2014). Early language development serves as a foundation for later reading achievement and children who demonstrate higher language skills during preschool are more likely to develop higher reading ability both in terms of decoding and comprehension (Lonigan & Shanahan, 2009). Even though Spanish and English oral language abilities do not result in statistically significant correlations (Melby-Lervåg & Lervåg, 2011;

* Corresponding author.
E-mail address: lduran@uoregon.edu (L.K. Durán).

https://doi.org/10.1016/j.ecresq.2019.02.001
0885-2006/

Palermo, Mikulski, Fabes, Martin, & Hanish, 2017), knowing more about growth in Spanish language development may provide teachers with an important opportunity to identify which DLLs may struggle with language learning or may be at-risk for low English reading achievement. For children who have been exposed primarily to Spanish in their homes, measuring Spanish language development becomes an important window into their general language learning ability (Castilla, Restrepo, & Perez-Leroux, 2010). For example, in a recent longitudinal analysis that modeled receptive language growth trajectories of 64 Spanish-speaking children, Jackson et al. (2014) found that the Spanish-speaking children who exhibited high levels of Spanish receptive vocabulary in preschool were likely to achieve greater growth in English receptive vocabulary through 2nd grade than those who exhibited low levels of Spanish receptive vocabulary, despite their Spanish and English receptive vocabulary skills being uncorrelated in preschool.

The mechanisms underlying associations between English and Spanish vocabulary development may be explained by the Revised Hierarchical Model (RHM; Kroll, Van Hell, Tokowicz, & Green, 2010). The RHM provides a framework for understanding how levels of first language (L1) proficiency impact cross-linguistic semantic associations. According to this theory, when children first begin to learn a new language, their L1 lexicon mediates access to conceptual knowledge in the second language (L2). For instance, when children encounter a new word in their L2, they use their L1 system to access their stored knowledge (e.g., the Spanish-speaking child hears *dog*, relates it to the Spanish word *perro*, and then accesses the concept of an animal with four legs and a tail that barks; Peña, Kester, & Sheng, 2012). As children learn more vocabulary in their L2 and have more experience using the language, L2 words start to develop their own pathways directly to the child's store of conceptual knowledge. Therefore, it is theoretically useful to teach children in their stronger language to facilitate the acquisition of new concepts and create a larger store of background knowledge that can be drawn upon to learn new words in their L2.

The RHM also provides support for assessing children in their L1 to better understand their level of L1 vocabulary, concept knowledge, and oral language ability that can be drawn on to support L2 acquisition. Evidence for this model can be found in recent work by Goodrich and Lonigan with a sample of 944 Spanish-speaking preschoolers. Significant correlations were found between the Spanish definitional vocabulary measure administered in preschool and English vocabulary tested in first grade (Goodrich & Lonigan, 2017). This same association was not found between picture naming in Spanish and English reading. Given that the definitional test is a measure of the child's conceptual knowledge of the item, it stands to reason under the RHM that an association may be present between children's conceptual knowledge of words known in Spanish and English vocabulary development. More evidence is also provided by Goodrich et al. in an earlier study in which they found that children were more likely to learn the English equivalents of words they already knew in Spanish. In their study with 212 preschool-aged children, participants defined more words correctly in English if the same word was known in Spanish (Goodrich, Lonigan, Kleuver, & Farver, 2016).

2. Associations between Spanish oral language and English reading

In the emergent literacy framework provided by Whitehurst and Lonigan (2001) both *outside-in* (contextual units, semantic units) and *inside-out* skills (language units, sound units, print units) contribute to fluent reading and comprehension. Spanish oral language can be viewed as an outside-in unit that provides children with semantic knowledge to draw from in the process of developing English reading fluency and comprehension. Some support has been found for this idea in research investigating the role of Spanish vocabulary knowledge and English reading fluency and comprehension (Proctor, August, Carlo, & Snow, 2006). In a study with 135 Spanish–English speaking 4th graders receiving Spanish and English literacy instruction a significant, but modest, interaction between Spanish vocabulary and English reading fluency and comprehension was found. Lindsey, Manis, and Bailey (2003) had similar findings in their study with 303 Spanish–English bilingual kindergarten and 1st grade students receiving literacy instruction in both Spanish and English (Lindsey et al., 2003). These findings lend support to the idea that Spanish oral vocabulary may provide outside-in support for English reading; however, the language of instruction may play an important mediating role for leveraging the cross-linguistic transfer of skills across languages (Cárdenas-Hagan, Carlson, & Pollard-Durodola, 2007).

In other studies significant correlations between Spanish oral vocabulary and English reading have not been found (Mancilla-Martinez & Lesaux, 2010; Melby-Lervåg & Lervåg, 2011). In the longitudinal study conducted by Mancilla-Martinez and Lesaux that followed 173 Spanish-speaking children from 4- to 11-years old Spanish vocabulary and word reading were not significant predictors of English reading comprehension. However, they did not report on the language of instruction received by the participants. Given conflicting findings, at the very least, the association between Spanish oral vocabulary and English reading achievement deserves further investigation with attention to potential mediators of this association such as the language of instruction and initial language proficiency in each language (Cárdenas-Hagan et al., 2007).

3. Data-based decision-making and early intervention

There is significant evidence that DLLs are at-risk for low English reading achievement, but importantly, early intervention holds promise for closing gaps early on by building the vocabulary, language, and pre-academic skills necessary to lay the foundation for later school success (Garcia & Jensen, 2009; Goldenberg, 2008). Targeted and effective instruction is guided by the accurate measurement of children's abilities (Brown & Sanford, 2011; Fien et al., 2011). For young Spanish speakers, measurement in Spanish is important given that home language environments exert the largest influence on the language development of preschool-aged children (Bohman, Bedore, Peña, Mendez-Perez, & Gillam, 2010; Peña et al., 2012). Practitioners run the risk of underestimating Spanish-speaking children's language development if they are only measured in English (Anaya, Peña, & Bedore, 2016).

All young children acquire language skills from hearing and using language (Hammer et al., 2012; Tomasello, 2005) and there is significant variability in use of Spanish and English in home and community language environments across the United States (Bedore et al., 2012; Hammer & Rodriguez, 2012; Reese, Linan-Thompson, & Goldenberg, 2008). The ratio of Spanish to English exposure, the type of Spanish spoken by the child's family based on country of origin (i.e. Puerto Rican versus Mexican Spanish), and even region within that country (i.e. differences in the Spanish used in northern versus southern Mexico), and the influence of English on Spanish in a bilingual context such as the U.S. all affect the Spanish to which young children are exposed. This variability in the amount of exposure to Spanish and the differences in the varieties of Spanish spoken in children's homes have been documented across several studies and presents a unique challenge in measurement design with Spanish-speaking children (Bedore et al., 2012; Bohman et al., 2010). Therefore, it is important to have language assessment tools that are sensitive to this variation to reliably

assess children's progress toward meaningful language and early literacy goals.

No Spanish preschool general outcome measures developed to guide instructional decision-making are publicly available that have published information on how variations in language exposure and dialectal differences were taken into consideration in the development process (Barrueco, López, Ong, & Lozano, 2012). There are several diagnostic Spanish preschool language assessments that are available for determining language delay or impairment such as the Bilingual English–Spanish Assessment (Peña, Gutierrez-Clellan, Iglesias, Goldstein, & Bedore, 2014), the Clinical Evaluation of Language Fundamentals (Wiig, Secord, & Semel, 2009), and the Preschool Language Scale (Zimmerman, Steiner, & Pond, 2012). However, there is also a need for Spanish general outcome measures that are designed to inform instruction to prevent delay or disability. In this paper, we focus specifically on the development of the oral language Individual Growth and Development Indicators-Español (IGDIs-E). The IGDIs-E is a general outcome measure (GOM) designed to guide instructional planning for Spanish-speaking preschoolers. The development of this assessment addresses a critical need in the field for Spanish oral language measures that are directly related to instruction and can improve early intervention. Although empirically sound general outcome measures are available in English, few measures are available in Spanish (Barrueco et al., 2012).

Having high quality assessments available in Spanish is important because assessing children who are dominant in Spanish only in English provides little information regarding language ability. It is important to measure children's skills in the language likely to yield the most accurate results regarding their ability levels (Hoff & Core, 2015). In fact, several national early childhood professional organizations have issued statements related to the assessment of bilingual children (American Speech and Hearing Association, 2017; Division for Early Childhood, 2010; National Association for the Education of Young Children, 2005). All concur that children should be administered assessments in their home language that are free from cultural and linguistic bias. Although these recommendations are no longer novel and are widely accepted, implementation is far from common in practice (Banerjee & Luckner, 2010). Part of the problem is that there are a limited number of preschool general outcome measures of Spanish oral language. These measures need to be affordable, psychometrically robust, and easily administered by a broad range of staff. Additionally, they should also provide information regarding a child's ability that is directly related to instruction. The IGDIs-E are designed to fill this critical gap in the field.

4. Multi-tiered systems of support and general outcome measures

General outcome measures have become one of the primary types of assessment tools used in Multi-tiered Systems of Support for data-based decision making (Fuchs & Fuchs, 2006) precisely because they are brief, easy-to-use, and aligned with long-term academic outcomes (Fuchs & Deno, 1991). MTSS in early education presents an important and timely opportunity to improve outcomes for children by promoting a focus on technically sound assessment practices and effective data-driven instruction (Klingner & Edwards, 2006). Efforts to establish MTSS in early childhood settings, such as the Center for Response to Intervention in Early Childhood, are also gaining momentum (Greenwood et al., 2011). MTSS features three tiers of targeted instruction, and assessment plays a central role in the identification of children who may benefit from additional instructional support. Importantly, assessments have documented evidence of validity and reliability, provide meaningful information, and be easily interpreted by teachers. In early childhood, the Individual Growth and Development Indicators (IGDIs 2.0; McConnell, Wackerle-Hollman, & Bradfield, 2014) are one set of general outcome measures frequently used nationally to screen students' skills to determine the most appropriate intensity of language and literacy instruction (e.g. Tier 1, Tier 2/3; McConnell, Wackerle-Hollman, Roloff, & Rodriguez, 2015). However, given the increasing number of Spanish-speaking preschoolers, the need for a Spanish version of the IGDIs that complements the English measures has become increasingly important and timely.

5. Individual Growth and Development Indicators-Español (IGDIs-E)

To accurately estimate levels of Spanish oral language development measures that are based on the development of Spanish and that are not simply translations of English measures must be developed (Fien et al., 2011; Peña et al., 2012). In total, the IGDIs-E include five measures targeting Phonological Awareness, Alphabet Knowledge, and Oral Language. For the purposes of this manuscript, we focus only on the oral language measure development (see Wackerle-Hollman et al., 2019 for a description of the Phonological Awareness measure). In this paper, we explore Spanish oral language measurement by focusing specifically on the development and calibration of the oral language (OL) IGDIs-E measures. The new IGDIs-E have been designed to incorporate specific linguistic features of Spanish such as lexical frequency of words and syntax as well as cultural and dialectal considerations. To provide information on the psychometric quality and evidence of social validity of the new OL IGDIs-E, we conducted two studies. We had two phases in the first pilot study. In the first phase we assessed the viability of seven different tasks for use with Spanish-speaking preschoolers. In the second phase we field tested four of the tasks that held the most promise for scalability, technical adequacy, and developmental appropriateness. In the second study we selected the two tasks with the most promise based on qualitative and quantitative analysis that would move forward to full calibration in which we addressed two main research questions:

1 To what degree are the OL IGDIs-E psychometrically sound measures of OL ability in Spanish?
2 To what degree do the OL IGDIs-E have evidence of feasibility and utility?

6. Measure construction method

IGDIs-E design is grounded in Wilson's (2005) model for constructing measures. Wilson's model includes four components: the construct map, item design, the outcome space, and the measurement model. The construct map guides item design resulting in responses that are scored in the outcome space and transformed through a measurement model to enable inferences about the construct. Each of the four components of Wilson's model contribute to the validity argument and collectively, they present a foundation for supporting interpretations and intended uses of the measure. Validity is not a property of the assessment itself, but rather is best described as the extent to which theory and evidence support the inferences that can be made regarding children's ability levels given the proposed uses of a test (AERA, APA, & NCME, 2014; Kane, 2013). Validation is the process of gathering that evidence. The following section summarizes how Wilson's model guided the IGDIs-E measure design process.

6.1. Construct map

Wilson's first component, the construct map, is a conceptual description of the construct of interest, defined here as oral language. After a systematic literature review on definitions of oral language used in the literature, the IGDIs-E research team defined the oral language construct as the ability to produce and understand spoken words via syntax, vocabulary, and semantics for the purpose of communicating with others (Miller et al., 2006; Pena et al., 2003Pena, Bedore, & Rappazzo, 2003). Expressive language includes the use of words to express meaning and receptive language involves the ability to listen, process, and understand the meaning of spoken language. The construct map structures item development to create items that tap the full range of the construct for the target audience (children in preschool the year before kindergarten) — so that characteristics of low levels to high levels of oral language development are captured. In our initial design process we included a range of both receptive and expressive tasks to capture our oral language construct.

6.2. Item design

Our team developed measures and items indicative of Spanish oral language skills guided by this oral language construct and definition. Peña et al. (2012) highlight that oral language skills that are targeted in assessment in any language should take into consideration the (a) difficulty of items, (b) lexical frequency, and (c) syntactical structures that may highlight one-word type over another.

During item design, we carefully selected the corpus of words used for designing the oral language IGDIs-E measures. Culture influences what people talk about, which in turn influences the types and frequency of words that are learned early on in language development (Hammer & Rodriguez, 2012; Peña et al., 2012). Common foods, clothing, and household items can vary considerably across languages. If a screening instrument is translated from an English version, then there is the potential for item bias given the cultural differences between languages that influence the frequency and prototypicality of vocabulary to which young children are exposed (Peña, 2007). Therefore, our research team created a corpus of words by including words found on the Spanish MacArthur-Bates Communicative Development Inventories I and II (CDI; Jackson-Maldonado et al., 2003) which was normed on Spanish-speaking children in the US, by cataloging words from over 100 Spanish language children's books, by examining target words in existing Spanish early childhood literacy curricula (Estrellita, Creative Curriculum, DLM Express), and by cross-referencing a dictionary of frequently used words in Spanish (Davies, 2006).

The development of the OL IGDIs-E tasks also began with an iterative process exploring seven different tasks in Spanish (Pena et al., 2003; 2012). These tasks were created based on pilot work by Peña and colleagues and from previous experience designing developmentally appropriate tasks to measure the oral language development of preschool-aged children. Peña and colleagues developed tasks in Spanish and English assessing six semantic skills: associations, characteristic properties, categorization, functions, linguistic concepts, and similarities/differences (Pena et al., 2003). They administered these tests to 55 participants ranging in age from 4 to 7 years old. In Spanish, the tasks from easiest to most difficult were expressive functions, receptive functions, receptive characteristics, receptive similarities and differences, receptive linguistic concepts, receptive categorization, receptive associations, expressive categorization, expressive characteristic properties, and expressive similarities and differences, expressive linguistic concepts and expressive associations. We used this research in our initial piloting phase and we decided to develop testing tasks that were identified as the easiest based on Peña's research such as functions and receptive characteristics, given that our target population was at the lower end of the age range of their participants.

6.3. Outcome space

Wilson's third component, the outcome space, involves decisions about scoring procedures in order to ensure that appropriate inferences can be made from scores obtained on the IGDIs-E. This captures the range of responses that should be scored as correct relative to the construct. All IGDIs-E measures have a dichotomous scoring framework with 0 points for incorrect responses and 1 point for correct responses. Our team recognized the need to carefully consider decisions regarding correct and incorrect responses and all expressive responses provided by children in the study were written down by data collectors during testing and responses with rates above 5% of the total sample were reviewed. The research team consulted scholars in Spanish language development, native Spanish speakers involved in early education, and translators. Responses that were deemed correct by expert consultant review and that occurred in more than 5% of the response sample were then included in the item key as correct. For example, a picture naming item features an image of a *naranja* (orange). After initial testing, *mandarina* and *china* were added to the correct responses, as these were other common responses offered by children in our sample, particularly from our Florida sample, which included more Puerto Ricans. Our expert consultants also confirmed that *mandarina* and *china* are regionally accurate words used to describe an orange.

6.4. Measurement model

Finally, Wilson's fourth component addresses the measurement model that allows for the transformation of item responses into scores permitting inferences about the construct. We employed Rasch modeling (see Bond & Fox, 2015, for a practical treatment of Rasch measurement) to examine how IGDIs-E items perform. Rasch item parameters were estimated using Winsteps (Linacre, 2016a). The typical formulations of this logistic (log-odds) model is $log(P_{ni} / [1 - P_{ni}]) = B_n - D_i$, where P is a probability of person n responding correctly to item i, and the Rasch parameters are B_n, the ability of person n, and D_i, the difficulty of item i, in logits. The odds of correctly responding to an item are a function of the difference between the person ability and the item difficulty. Person ability and item difficulty are located on the same scale, and therefore, they are comparable. If a person's ability is greater than the item difficulty, the person has greater than a 50% probability of correctly responding; if the person's ability is lower than the item difficulty, their probability of correctly responding are less than 50%. When the model fits the data, the Rasch model provides units of measurement on an interval level, so that there is a uniform interpretation across the scale.

In order to answer questions regarding how well items map onto child-level ability, our team explored Wright item maps and the range of child ability. Wright item maps are visual depictions of how items are distributed relative to child abilities. Given that Winsteps creates the item/person location scale so that the average item location is zero, the location of the children indicates their ability levels relative to the item difficulties. On the Wright item map, a measure is adequately representing a construct when item distributions mirror the distribution of child level abilities.

This work was completed in three steps: Study 1 Phase 1 involved a pilot test of prototype measures. Study 1 Phase 2 involved a more formal field test of measures that survived the pilot phase. Study 2 involved a highly structured calibration process where the measures that functioned successfully from the field

test were formally calibrated and evaluated for measurement and psychometric quality.

7. Study 1 Phase 1: pilot test

The purpose of Study 1 Phase 1 was to complete a pilot test of prototypes of a set of initial measures. This utilized the prototypes of the measures designed as described earlier.

7.1. Methods

7.1.1. Participants

Thirty 4- to 5-year-old Spanish-speaking preschoolers identified through teacher nomination recruited from Migrant Head Start in Utah and a private bilingual preschool in Minnesota participated in the pilot.

7.1.2. Expressive measures

Expressive tasks for oral language included: (a) Identificación de los Dibujos/ Picture Naming, (b) Categorías/ Categories, (c) Funciones/ Functions, and (d) Verbos Expresivo/ Expressive Verbs.

Identificación de los Dibujos/Picture Naming is an untimed expressive measure that requires children to name images of common objects, animals, and foods. The measure is designed for use with 4- and 5-year-old children. The test begins with four sample items with the administrator first modeling the task and then the child is given two example items to practice responding. If the child responds incorrectly systematic correction procedures are provided and the child is offered another opportunity to respond. The test is discontinued if the child does not respond correctly to all sample items. The administrator presents items one at a time to the child. Each item includes one picture for the child to name. The administrator asks the child, "¿Qué es?" (What is this?). If an image has more than one name due to dialectal differences, then all possible correct answers appear on the back of the item card (items are currently presented on cards). Items for which the child produces a response that matches a response on the back of the item card are scored as correct, and all other responses (i.e., anything that does not appear on the back of the item card) are scored as incorrect (see Supplementary Appendix A for examples of items).

Categorías/Categories is a measure where children must state the category to which the three images on a card belong, or how these images "go together". This task involves the semantic ability to understand group and category membership, as well as the ability to produce spoken language. When giving the task, the administrator names each image and then asks how these items go together. The back of the card contains all possible answers to account for cases in which objects may belong to more than one category or in which there are multiple ways to word the correct answer.

Funciones/Functions is an expressive task where items provide images of household objects, toys, and everyday nouns. When displaying each item, the administrator named the image for the child and then asked, "¿Para qué sirve?" (What is this object used for?). For objects with multiple functions or with multiple verbs used that can describe the same function (e.g., un carro sirve para conducir o manejar — a car is used to drive), all potential verbs or functions appeared on the back of the card. Items were scored as correct when the child provided a verbal response that matched a response on the back of the card, and were scored as incorrect when the child produced a word or phrase that was not provided as a correct response.

Verbos Expresivo/Expressive Verbs is an untimed expressive measure and is also designed for use with 4- and 5-year-old children. This measure also includes four sample items and discontinue rules are as described above. The administrator presents each item in succession, asking the child, "¿Qué está pasando?" (What is happening?). Although attempts were made in the design process to select images portraying one clear action, multiple possible responses were included for items with images that solicited multiple verbs; each item contains one image. Items are scored as correct when the child produces a verb that is included on the back of the item card, including all conjugations of the target verbs (see Supplementary Appendix B for examples of verb items with multiple accepted correct responses).

7.1.3. Receptive measures

Receptive tasks for oral language included: (a) ¿Cuál dibujo es diferente?/ Which one doesn't belong?, (b) Verbos Receptivo/ Receptive Verbs, and (c)Vocabulario de Definiciones (Receptivo)/Definitional Vocabulary (Receptive).

¿Cuál dibujo es diferente?/ Which one doesn't belong? is a semantic task evaluating children's ability to distinguish between categories. Each card contains three images, two of which belong to the same category. After the administrator names each image on the card, the child may respond by either pointing to or saying the name of the image that does not belong.

Verbos Receptivos/ Receptive Verbs is a receptive task that measures children's ability to select the image that corresponds to an action word spoken by the administrator. Each item displayed two or three images of actions. Administrators presented each item in succession and asked the child, "Señala el dibujo de ___" (point to the picture of ___), filling in the blank with the item's target verb. Children could respond by pointing to the image. Scores were recorded as correct (identifying the image that matched the target verb) or incorrect (selecting any other image).

Vocabulario de Definiciones (Receptivo)/ Definitional Vocabulary (Receptive) presents children with one image per card. The assessor names the item and offers a statement such that a defining feature of the item is required for a response. For example, the assessor would say "Este es el sol, hace calor o frío?" (This is the sun, is it hot or cold?).

7.1.4. Procedures

Trained bilingual graduate research assistants and the lead author administered the seven measures to at least 10 children per measure. A qualitative rubric was completed by the assessors to evaluate the following criteria: (a) active engagement of child, (b) valid response patterns obtained from child, (c) ease of use by administrator, and (d) timeliness of measure administration and scoring. Each of these four components were rated using a 0–3 scale, where 0 represented an unsatisfactory and unresolvable measure that did not achieve its desired outcome, and where 3 represented a superior measure that achieved its desired outcome to the highest standard and was considered for further testing.

To support the qualitative findings, we also examined descriptive statistics for each task including mean, median, standard deviation, minimum value and maximum value. Empirical standards were used to evaluate the distribution of scores and provide a cursory view of the likelihood for a task to be successful with 4–5-year-old children. We specifically investigated the number of valid responses on the assessment, the time it took to deliver the assessment, and child engagement.

7.2. Results

7.2.1. Qualitative results

If the test yielded few valid responses and children were reported to be distracted or confused during the task it was eliminated. Based on the qualitative rubrics completed by administrators ¿Cuál dibujo es diferente?, Categorias, and Vocabulario de Definiciones were eliminated. Administrators indicated that children

Table 1
Pilot study qualitative rubric results.

Task	Active engagement	Valid response patterns	Easy to use	Timely to deliver	Grand total score	Rank
Identificación de los Dibujos	3	3	3	2	11	1
¿Cuál dibujo es diferente?	2	1	2	2	7	6
Categorias	2	1	2	2	7	6
Funciones	3	3	2	2	10	4
Definiciones	3	2	3	3	11	1
Verbos Receptivo	3	3	2	3	11	1
Verbos Expresivo	3	2	2	2	9	5

Note: The grand total score is based on the sum of the four specific items in the first four columns of scores. The ranks are based on the grand total scores.

were not engaged and not responding to the ¿Cuál dibujo es diferente? and Categorias measures. They also argued that they took too long to administer. The receptive Vocabulario de Definiciones was easy to administer and the children were engaged with the task however nearly all children were able to correctly respond to every item and test administrators indicated that the task was too easy. Test administrators responded favorably to Funciones, Verbos Expresivos, and Identificación de los Dibujos (Table 1).

7.2.2. Quantitative results

Additionally, we investigated mean scores, standard deviations, and minimum and maximum values. We looked for measures that captured a range in ability as evidenced by a spread between minimum and maximum values and a standard deviation that indicated sufficient variance in performance. Definiciones was eliminated because there was functionally no variability in performance and almost every child provided correct responses to every pilot item. ¿Cuál dibujo es diferente? was eliminated because the minimum score was a 2 with the maximum score at 6. This restricted performance range did not make the task a good candidate for capturing range of child ability. Only 2 out of 10 children were able to respond to the items on the Categorías measure, which made this measure a poor candidate for scalability. Based on the quantitative results Funciones, Verbos Receptivos, Verbos Expresivos and Identificación de los Dibujos were selected for the field study. On all of these measures, over 50% of the children provided a response to the items and there was adequate variability in performance on the measure as indicated by larger standard deviations and range in the minimum and maximum values (Table 2).

8. Study 1 Phase 2: field test

The purpose of Study 1 Phase 2 was to complete a more formal field test of measures that survived the pilot phase. This involved a larger sample of children and item response theory analysis of item and measure functioning.

Table 2
Pilot study quantitative results.

Task	Empirical criteria					
	n	*M*	*Median*	*SD*	*Min*	*Max*
Identificación de los Dibujos	10	6.10	6.5	2.38	2	10
¿Cuál dibujo es diferente?	8	4.12	4.0	1.25	2	6
Categorias	2	4.13	4.0	1.25	0	7
Funciones	7	4.4	5.0	2.54	1	8
Definiciones	10	9.5	10.0	0.71	8	10
Verbos Receptivo	10	9.5	10.0	0.85	8	10
Verbos Expresivo	6	5.0	5.5	1.26	3	6

Note: *n* is the number of valid responses out of 10 students who interacted with each task.

8.1. Methods

8.1.1. Participants

Spanish-speaking children (as noted by parent or teacher report) between the ages of 4 and 5 who were going to kindergarten the following year were identified as potential participants in filed testing from various public and private pre-primary education programs in Minnesota and Utah, including Head Start and private preschools. A total of 200 children participated in the field testing, with 100 consents from Minnesota and 100 from Utah. Children were primarily from families who claimed Mexico as their country of origin with the remainder largely from Central America.

8.1.2. Measures

After we conducted this initial piloting, we were left with four viable and promising oral language measures to include in Study 1 Phase 2 for field testing: (a) Identificación de los Dibujos, (b) Funciones, (c) Verbos Expresivos, and (d) Verbos Receptivos.

8.1.3. Analysis

Item calibrations were produced using Winsteps 3.91 (Linacre, 2016a) for all items in each oral language IGDI measure. Item-total point-biserial correlations above 0.2 were considered acceptable for items to be included in the final assessment (Haladyna & Rodriguez, 2013). These are classically referred to as item-discrimination indices, indicating the extent to which items discriminate between persons with higher versus lower abilities as a function of the total score. All other items were eliminated given their lack of fit with the model. Reported item level statistics also include *p*-values which represent the proportion of children who answered the item correctly with acceptable *p*-values ranging from 0.2 to 0.8 (Haladyna & Rodriguez, 2013). Model fit was explored using the mean square infit and outfit statistics for item difficulties and person abilities. Infit and outfit statistics between 0.5 and 1.5 were considered acceptable (Linacre, 2016b). Items with infit and outfit greater than 1.5 were eliminated or revised. We also broadly explored the potential of chance or guessing in item-level responses. If there was error in items associated with guessing (model misfit), then the infit and outfit statistics would have been above acceptable levels.

8.2. Results

8.2.1. Item functioning

The person score reliabilities were adequate for Identificación de los Dibujos at 0.79, Verbos Expresivo at 0.78 and Funciones at 0.77. The person score reliability for Verbos Receptivo was not adequate at 0.35. Identificación de los Dibujos had the lowest percent of non-responses at 13% with Funciones at 27%, Verbos Expresivo at 33%, and Verbos Receptivo at 17%. All measures had a range in the minimum to maximum values between 1 or 2 correct responses to 19 out of 20. The mean Rasch scores were 1.45 for Identificación de los Dibujos, 0.82 for Verbos Expresivo, 0.67 for Funciones, and 2.62 for Verbos Receptivo (Table 3).

Table 3
Field study quantitative descriptive results.

Task	*M* raw score	*M* Rasch score	*SD*	*Min*	*Max*	Score reliability	% of non-responses
Identificación de los Dibujos	12.7	1.445	4.1	1	19	0.79	13%
Verbos Expresivo	12.0	0.824	3.6	2	19	0.78	33%
Funciones	12.1	0.669	3.7	1	19	0.77	27%
Verbos Receptivo	17.0	2.616	2.8	2	19	0.35	17%

Note: Rasch scores are in logit metric; each measure is scaled independently.

8.2.2. Expert reviewers

Additionally we included a panel of experts to review each of the oral language tasks. Leading university researchers who were all well published in Spanish language and literacy, bilingualism, and linguistics participated in the review and three of the four experts were native Spanish speakers. Each expert was provided with an online survey that provided example images of the stimulus material for each of the four assessments and specific questions about the extent to which the task was culturally, linguistically, and developmentally appropriate. We also met with the reviewers individually to gather their impressions and suggestions. We carefully summarized the results across all reviewers and used these data in our final decisions about what measures to move forward into the final development phase (see IGDIs-E Technical Manual, Wackerle-Hollman, Durán, Rodriguez, Brunner, & Palma, 2016 or a full summary of results).

This entire piloting and field study process led to the decision to move Identificación de los Dibujos and Verbos Expresivo forward to the final calibration study. After the field testing we had to consider both which tasks functioned well empirically and qualitatively. We also needed to carefully consider which tasks were scalable and would lend themselves to creating 100 items for calibration with easy and unambiguous scoring rules. Our team decided that Verbos Recptivo was too easy as most children were likely to be successful and it would have low discrimination potential. Identificación de los Dibujos functioned well across all of our metrics and was a well-supported choice for calibration. The choice between Funciones and Verbos Expresivo was more difficult because each task functioned well empirically, but each also had some challenges. Ultimately, Verbos Expresivo was selected because 100 test items could be created with less difficulty and scoring rules would be more transparent than with a test that focused on asking children to provide a function for an object. For instance, take an item as simple as a towel. Consider how many accurate functions a child might describe such as to dry, to clean, to cover, to wipe, to wash, etc. We decided that adequately capturing the range of correct responses for each item would be a difficult task that might also interfere with scoring and ultimately the accurate estimation of child ability. Asking a child to provide a discrete action in the expressive verbs measure was much more amenable to the process of measurement design.

9. Study 2: calibration

The purpose of Study 2 was to complete a formally structured calibration of measures that advanced from the pilot and field test phases. This utilized a much larger sample of children and psychometric evaluation of item and measure functioning.

9.1. Methods

9.1.1. Participants

Child participants were recruited from public voluntary prekindergarten programs, private preschools, and Head Start programs in Minnesota, Florida, California, Idaho, Kansas, Illinois, and Utah. To participate in the study, children had to have exposure to Spanish in their homes according to parent report. For the purposes of this research, our team developed a home language survey, the Language and Literacy Environment Evaluation Report (LLEER; Durán & Wackerle-Hollman, 2016), that required caregivers to indicate what languages children heard and spoke within a weekly time-block matrix. Caregivers indicated whether children heard and spoke Spanish, English, or both. According to parent report, 68% of the sample heard only Spanish from birth and 24% heard both English and Spanish. Children who were reported by parents to be most comfortable speaking in Spanish or a mix of Spanish and English comprised 72% of the sample.

Children included in the study also had to be 4 or 5 years old, and had to be eligible for kindergarten the following year. Informed consent was collected by graduate research assistants and data collectors over the course of two academic years. In the 2013–2014 school year, 491 children returned consents to participate. In the 2014–2015 school year, 485 consented participants were added. Each child participated in only one year of the study. Collapsing across both years, the calibration study included 46 schools, 134 classrooms, and 976 children. We intentionally sought a dialectically diverse sample, so we recruited from Florida to obtain Caribbean, Central and South American Spanish dialects; and from Utah, California, Minnesota, Kansas, and Illinois to obtain Mexican and Central American dialects. Across all sites that reported participants' gender, 50% of the sample was male and 50% was female. Children's ages ranged from 4:0 to 5:11, 5.8% of the sample received special education services according to parent report, and 57% of the sample met poverty guidelines for Head Start enrollment or had a weekly family income less than $500. About 1 in 4 Latino preschoolers live in poverty in the US and our sample does include an over-representation of children in poverty (Children's Defense Fund, 2017). Across both years, samples from Florida, Utah, California, and Minnesota were present, with Illinois and Kansas only present in the 2014–2015 academic year. No significant differences in ethnicity, gender, or program type were present across years so demographic information was collapsed and is presented in Table 4.

Fourteen Spanish-speaking classroom teachers and teacher assistants at the testing sites participated in the feasibility and utility study. Six teachers and eight teaching assistants completed the usability and classroom language use surveys across four locations: Minnesota, Florida, Utah, and California. The lead teachers had an average of five years working in early childhood programs. Two teachers held graduate degrees and four teachers held four-year degrees in early childhood education or family, child, and human services. Seven teaching assistants either held a high school diploma or GED and had completed the Child Development Associate (CDA) Credential. One teaching assistant held an AA degree and a CDA.

The goal for both cohorts was to have data for fall, winter and spring for all children; given the realities of child absences, holidays, and other events, there were cases of missing data for all three seasons. Data were missing randomly (e.g., children were sick that day) and there were a few cases in which kids stopped participating (e.g., because they moved); however these cases were minimal. We treated child participants (particularly for calibrations) as unique cases between seasons (so as if it was a new participant each sea-

Table 4
Sample demographic characteristics by state.

State	*N*	Program type	Regional representation (percentage of sample by region)	% male
Minnesota	29	Private/ scholarship/ dual immersion	Central American: 10% South American: 3% Caribbean: 3% Mexican: 28%	34%
	135	Head Start	Identified as Hispanic/Latino: 80%	49%
Florida	249	Public VPK	Central American: 5% South American: 3% Caribbean: 58% Mexican: 5% Identified as Hispanic/Latino: 70%	46%
Utah & Idaho	176	Head Start	Central American: 1% South American: 0% Caribbean: 3% Mexican: 31% Identified as Hispanic/Latino: 81%	41%
California	257	Head Start	Central American: 4% South American: 0% Caribbean: 0% Mexican: 23% Identified as Hispanic/Latino: 82%	55%
Kansas	72	Public pre-kindergarten	Central American: 0% South American: 0% Caribbean: 0% Mexican: 17% Identified as Hispanic/Latino: 100%	N/A
Illinois	58	Public pre-kindergarten	Identified as Hispanic/Latino: 78%	N/A

Note: Private/scholarship programs are tuition-based and offer scholarships to families based on need. Dual immersion programs offer instruction in both English and Spanish and sometimes dictate enrollment ratios of native English speakers and native Spanish speakers. VPK is voluntary prekindergarten, which is a free prekindergarten program offered to families of 4-year-olds in the state of Florida. The sites in Kansas and Illinois decided to keep their students de-identified; as such, we could not obtain child gender for these participants. Aggregated data regarding regional representation was provided by the Kansas and Illinois sites.

son) and therefore, missing data were not found to compromise the analysis. This was important for the purposes of item calibration. The Rasch model (as with most IRT models) estimates the probability of a correct response as a function of person ability and item difficulty. As 4-year-old children grow from one season to the next, their oral language abilities are changing. Therefore, a child's responses to items in the fall are based on one ability and their responses in the winter (or spring) are based on a different ability — thus children are treated as having different abilities across seasons in order to properly estimate item difficulties.

This approach to resampling the same children was important to support calibration of all new items introduced at each season. To calibrate items in any operational testing program, test takers would always be considered as new participants, to account for potential changes in their trait levels. This may be viewed as a resampling procedure, but it is essential to not confound changes in ability with item difficulty — as person ability and item location are estimated simultaneously in IRT. Nor would it make sense to fix person ability to calibrate items in a later season, since ability most likely changes and would bias item calibrations due to the inappropriately fixed person abilities. It is not a longitudinal model, as growth is not being directly estimated.

Finally, we acknowledge that there is a relatively wide range of ages of children in preschool the year before kindergarten, including 4- and 5-year-old children. A one-year difference could encompass one-fourth of the child's life. However, these differences do not necessarily interfere with calibration of item difficulties or person abilities. The items are designed to be appropriate for children in this age range and with a variety of levels of preschool experience. The calibration process assumes that children will vary in their abilities and the construct map embodies this variation — whether it is a function of age or language experience or any other personal context/condition, as long as that variation is construct relevant, the calibration should successfully place all items and children on the same scale and produce scores that allow for appropriate inferences to the construct (given the results of the evaluation of measurement invariance presented below).

9.1.2. Measures

In order to quantify the association between IGDIs-E performance and existing norm-referenced measures of Spanish language, a strategically sampled subset of children ($n = 59$) also completed the Preschool Language Scales-Fifth Edition Spanish (PLS-5 Spanish; Zimmerman et al., 2012) during the 2013–2014 academic year. Students were selected from each participating state with 12 from MN, 13 from FL, 25 from CA, and 9 from UT. They were selected based on country of origin, gender, and program type so that the subsample reflected the diversity present in our larger sample. The PLS-5 Spanish includes Auditory Comprehension (AC) and Expressive Communication (EC) subscales which combine to produce a Total Language Score. The PLS-5 Spanish was normed on 1150 Spanish-speaking children from all regions of the United States and in Puerto Rico representing a broad range of Spanish dialects with both simultaneous and sequential bilinguals. Reliability coefficients range from 0.91 to 0.92 for 4- to 5-year-olds.

Participating teachers completed a feasibility and usability survey that included five sections: ease of administration, clarity of decision making criteria, feasibility, utility and practicality, and demographics. The survey consisted of 20 four-point rating scale items ranging from *strongly agree* to *strongly disagree*. Nine open-ended questions elicited recommendations to improve administration, scoring, or test materials.

9.1.3. Procedures

Two trained graduate research assistants conducted in-person or online comprehensive trainings for out-of-state sites so that all data collectors could maintain standardized administration procedures and fidelity across sites. Due to data collector attrition and scheduling limitations, training was completed twice annually. Training included eight sections: logistics, ethical principles for assessment, privacy and confidentiality, sampling scheme, data

collection activities, scoring and entry procedures, and delivering the IGDIs-E assessment with fidelity. Importantly, training was also conducted in Spanish for data collectors whose primary language was Spanish. All data collectors were required to achieve 100% fidelity on administration procedures and scoring within three consecutive attempts, the first two allowing for feedback. Data collector country of origin varied across sites with approximately 14% of the data collectors being of Mexican origin, 27% from Central or South America, and 59% identified as non-Latinos who spoke Spanish with native fluency and pronunciation gained, in most cases both through college level study of Spanish and extensive travel in Spanish-speaking countries.

As Rasch calibration requires at least 100 responses per item to appropriately calibrate ability levels and item difficulties (de Ayala, 2009), we designed a sampling scheme that allowed strategic testing of all items in a way that minimized testing burden on participants, allowing them to receive a reduced number of items. Over the course of two years, 15 groupings of items were distributed across nine different forms. For each measure, each form consisted of four sample items and 25–28 operational items. A linking approach linked forms such that common items appeared across forms allowing all items to be calibrated as a total item bank (through concurrent calibration). On each form, about 25% of items were linking items (which varied slightly across sets of forms in a matrix sampling approach), about 75% were unique items. A different group of children was recruited each year and tested seasonally in fall, winter, and spring of the 2013–2014 and 2014–2015 academic years. Each season, children were randomly assigned to one of the IGDIs-E forms.

Teachers were given the survey in the spring of 2014 by the on-site graduate research assistants. Teachers had as much time as they needed to complete the survey and 100% of the surveys were returned from the 14 participating teachers.

9.1.4. Analysis

Item calibrations were produced using Winsteps 3.91 (Linacre, 2016a, 2016b) for all items in each oral language IGDI measure, similar to what was done during the field test phase. We employed the same criteria and guidelines presented in the analysis plan for the field test. In addition to examining item statistics and item fit to the Rasch model, we also examined the dimensionality of the measures, as unidimensionality is an assumption of the Rasch model. Two methods were used to evaluate dimensionality, including the evaluation of the Rasch residuals from the model with principal components analysis. Principal components analysis of Rasch residuals (Wright, 1996) is a method used to investigate unidimensionality of a measure. Principal components analysis of Rasch residuals looks at patterns within the data that are not explained by the Rasch measures (i.e. item difficulties and person abilities). If there is systematic variation in the residuals, then there might exist unexplained dimensions interfering with the measurement of the intended constructs. In addition, a confirmatory factor analysis (CFA) was completed for five forms of Identificación de los Dibujos and eight forms of Verbos Expresivos with sufficient data, using Mplus (version 7.0; Muthén & Muthén, 2012). Three indices provide different aspects of model-data fit, including the root mean-squared error of approximation (RMSEA), the extent to which the model fits reasonably well in the population; comparative fit index (CFI), the relative fit to a more restricted baseline model; and the Tucker–Lewis index (TLI), which compensates for the effect of model complexity. Multiple indicators of fit should be examined (Muthén & Muthén, 2012). The general criteria for model-data fit are as follows (Sanford & Brown, 2011): RMSEA < .05 is good fit, RMSEA < 0.08 is adequate fit; CFI > 0.95 is good fit, CFI > 0.90 is adequate fit; TLI > 0.95 is good fit, TLI > 0.90 is adequate fit.

Additional analyses included the examination of item maps which illustrate the alignment of item difficulties and person abilities — to evaluate the extent to which items are located and providing information relative to the ability levels of the children in the study. Person score reliabilities from the Rasch model were examined and reported.

An examination of measurement invariance was conducted using differential item functioning (DIF) analysis. To support score interpretation across diverse communities of Spanish-speaking preschoolers and acknowledge variation in language usage, we conducted differential item functioning (DIF) analyses to explore potential item bias based on dialectal representation and sex. First, we explored dialectal differences. There are many different dialects of Spanish represented in the United States and it is important to examine the degree to which there may be bias present between groups represented in our sample. Due to sample size constraints of our participants, we were only able to separate children who spoke a Mexican dialect of Spanish from children who spoke any other dialect (e.g., Caribbean), as DIF analysis requires independent calibration of items based on group membership. Dialectal groups from other regions did not include over 100 children per group, so we could not produce 100 item responses per item for each dialectal group. Items were flagged for DIF if the contrast resulted in a p-value of less than 0.05 and the magnitude of the Rasch contrast DIF was 0.64 or greater, the magnitude defined by ETS as moderate to large DIF (Linacre, 2016a, 2016b; Zwick, 2012). DIF analysis was conducted using the Rasch-Welch probabilities associated with the contrasts (differences) in item locations (difficulties).

Finally, additional analyses were conducted to evaluate the association of the measures with an related measure and teacher reports were gathered to obtain feedback regarding the utility and feasibility of each measure.

9.2. Results

To address the first research question exploring whether the IGDIs-E are psychometrically sound measures of oral language, we explored item statistics, dimensionality, child level abilities, item maps, person reliability, measurement invariance using differential item functioning, and correlations with the PLS-5.

9.2.1. Identificación de los Dibujos item performance and model fit

For Identificación de los Dibujos, 117 items were calibrated based on 2178 cases. Each student response to each item represented a case. The numbers of item responses varied because of absences, children discontinuing the assessment, and 11 items were deleted prior to final calibration due to a low number of responses (insufficient for calibration). For the remaining 106 items, mean-squared infit and outfit statistics fell within the optimal ranges (i.e., 0.5–1.5) for all but five items (rosa/rose, arroz/rice, bate (de beisbol)/baseball bat, limón/lemon or lime, and cinta/tape). Ranges for p-values and item-total correlations are displayed in Table 5. In all, 81% of item p-values fell in the desired range and 99% of items had adequate point-biserial correlations (105 of 106 calibrated items).

9.2.2. Verbos Expresivo item performance and model fit

For Verbos Expresivo, 111 items were calibrated based on 2455 cases, again representing each individual student response. Similar to Identificación de los Dibujos, 18 Verbos Expresivo items were removed prior to calibration because they had poor item fit statistics, including gatear/crawl, balancear/balance, zambullir(se)/dive, buscar/search, guiñar/wink, estornudar/sneeze, and bailar/dance. After calibration on the remaining 93 items, another four items were removed because the infit and outfit statistics fell outside

Table 5
Item-level statistics.

	N	Rasch item difficulty range	SE range	Mean-square infit	Mean-square outfit	Point-biserial item-total correlation	Item *p*-value range
Picture Naming/ Identificación de los Dibujos	106	−3.27 to 3.59	0.06–0.46	0.65–1.37	0.28–1.79	0.20–0.75	0.11–0.93
Expressive Verbs/Verbos – Expresivo	93	−4.28 to 4.73	0.05–0.55	0.70–1.35	0.31–3.00	0.21–0.63	0.02–0.90

Note: Rasch item difficulties (location measures) are reported in logits, with $M = 0$, generally ranging from approximately −4 to +4.

Table 6
Principal component analysis of residuals results.

	Verbos (Expresivo)/Expressive Verbs ($n = 93$)			Identificación de los Dibujos/Picture Naming ($n = 106$)		
Variance sources	Eigenvalue	Observed	Expected	Eigenvalue	Observed	Expected
Total raw variance in observations	151.6	100%	100%	173.3	100%	100%
Raw variance explained by measures	58.6	38.7%	38.5%	67.3	38.8%	38.8%
Raw variance explained by persons	27.3	18.0%	17.9%	38.5	22.2%	22.2%
Raw variance explained by items	31.3	20.7%	20.6%	28.7	16.6%	16.6%
Raw unexplained variance (total)	93.0	61.3%	61.5%	106.0	61.2%	61.2%
Unexplained variance in 1 st contrast	1.6	1.1%	1.8%	1.8	1.1%	1.7%

optimal range (0.5–1.5): pintar/paint, caer(se)/fall, serrar o serruchar/saw, and rugir o gruñir/growl or snarl. Seventy-four percent of item *p*-values fell in the desired range and 100% of items had adequate point biserial correlations (Table 5).

The evidence from successful Rasch calibration of both Identificación de los Dibujos and Verbos Expresivo provides strong validity evidence based on response processes (AERA, et al., 2014). This indicates that items function consistently with Rasch model, which is consistent with the construct map approach estimating the underlying trait continuum for the two measures.

9.2.3. Dimensionality

We explored principal components analysis of the Rasch residuals for both IGDIs-E measures using all items from the current calibration. Principal components analysis reports illustrate the observed variance explained in each measure. Resulting statistics offer a breakdown of the total variance explained by the measure, the children (persons), and the items. This variance, as in factor analysis, can be standardized and represented with eigenvalues. We observed small eigenvalues for Identificación de los Dibujos of 1.68 and Verbos (Expresivo)/Expressive Verbs at 1.85, indicating that the first dimensions in the results had the strength of less than two items. This result provided evidence that there was minimal systematic variation in the residuals — in other words, the data unlikely contain multiple dimensions. Principal components analysis results are presented in Table 6 for each measure.

In addition, the CFA was completed on multiple forms of each measure. Based on the criteria described in the analysis section, each operational form of measure resulted in good to adequate fit to a unidimensional model (mostly in the good range), meeting the unidimensionality assumption of the Rasch model (full results are available in the technical report; Wackerle-Hollman et al., 2016).

The evidence from successful CFA of both Identificación de los Dibujos and Verbos Expresivo provides strong validity evidence based on internal structure (AERA et al., 2014). This indicates that items are strongly associated with a common factor, which is consistent with the construct map approach estimating unidimensional latent traits, providing a strong basis for score interpretation.

9.2.4. Child-level abilities, item alignment, and person reliability

To empirically examine how Spanish–English bilingual children performed on Verbos Expresivo and Identificación de los Dibujos, we used the final concurrent item calibrations to estimate ability scores for each season for the 976 children in our sample. In Identificación de los Dibujos child abilities ranged from −4.01 to 4.18 across seasons, with a mean score of 0.19 and a standard deviation of 1.72. Verbos Expresivo mean child ability levels ranged from −4.77 to 3.69 with the mean performance at −1.58 and a standard deviation of 2.27.

Figs. 1 and 2 depict the Wright item maps for Verbos Expresivo and Identificación de los Dibujos respectively. In these figures, item difficulties and child level abilities are on the same scale. A measure is meaningfully capturing the observed range of child level abilities when the items are distributed across the same range as child ability distributions. These figures illustrate that the IGDIs-E measures have adequate construct coverage and meaningfully map to the majority of student abilities.

In a Rasch model, person reliability is comparable to a standard reliability coefficient in classical test theory (Linacre, 2016a, 2016b), the proportion of score variance that is true-score variance, not error variance. Winsteps estimates an adjusted variance which is consistent with parameter or true variance, and thus estimates the ratio of true to observed score variance. For Identificación de los Dibujos, score reliability is about 0.91; for Verbos Expresivo, score reliability is similarly about 0.91.

An important difference is that score precision varies in IRT models across score levels. A more appropriate approach to estimate score consistency of a measure is to examine the conditional standard error of measurement (CSEM) for each ability level. In item response models, each ability level is estimated with a different level of precision, based on the information available at that ability level (essentially based on having items located at that ability level). Smaller standard error of measurement indicates increased precision in the score and less error, whereas higher standard error of measurement values indicate more error and less precision in capturing child ability. See Figs. 3 and 4 for the CSEM curves for Identificación de los Dibujos and Verbos Expresivo. The CSEM curves indicate that the smallest standard error of measurement is near the average item location (0.0) and much larger standard error of measurement exist for scores lower than −4.0 (and higher than +4.0). We see the lowest levels for Identificación de los Dibujos slightly lower than the 0.0 level, suggesting slightly more precision for scores located nearer the lower ability levels, with similar results for *Verbos Expresivo*. We also observed slight variation in these curves (the dotted curves) based on variation across forms. These levels of measurement error are comparable

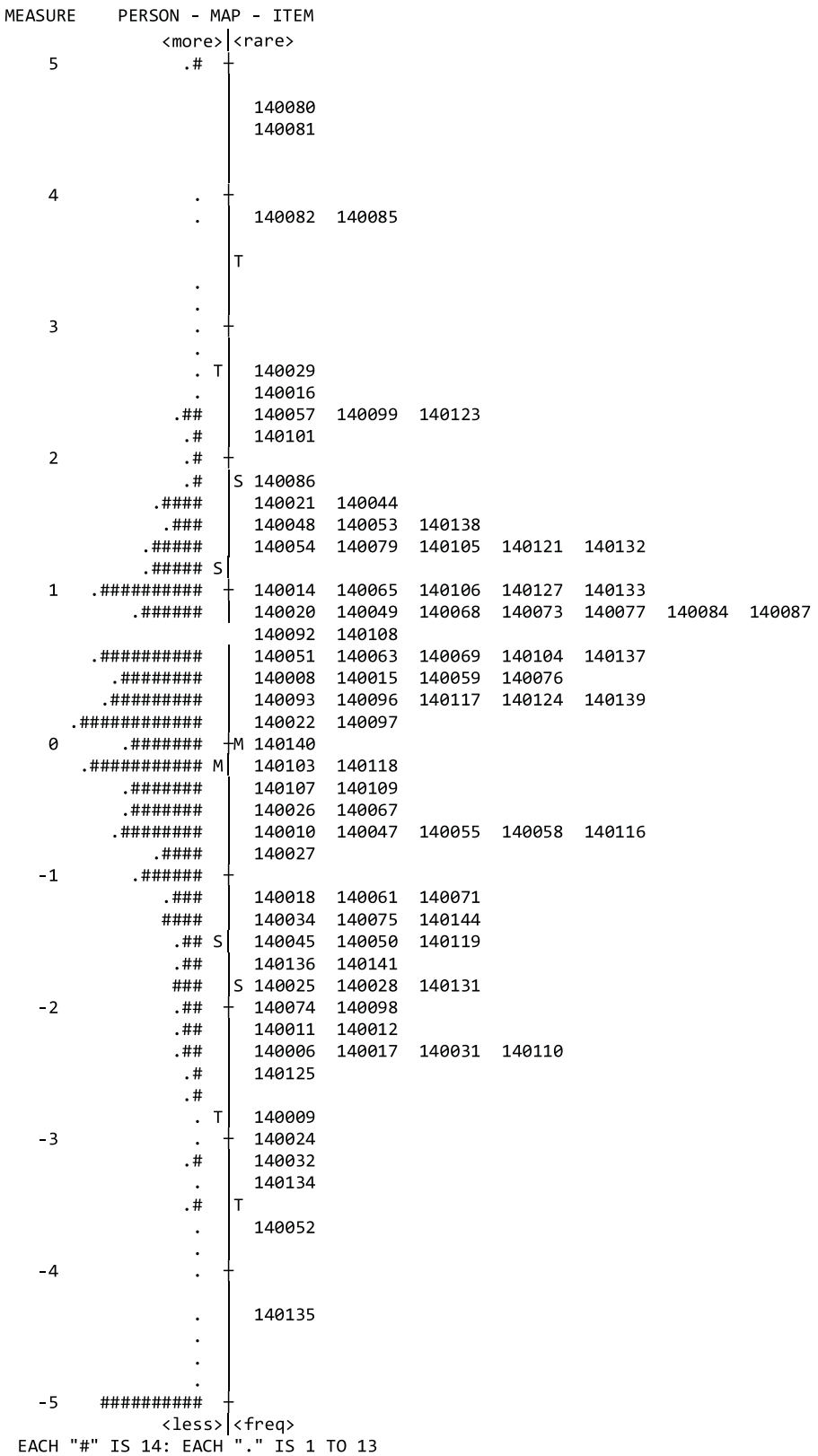

Fig. 1. *Verbos (Expresivo)/Expressive Verbs* Wright item map, including 2618 cases and 93 items.

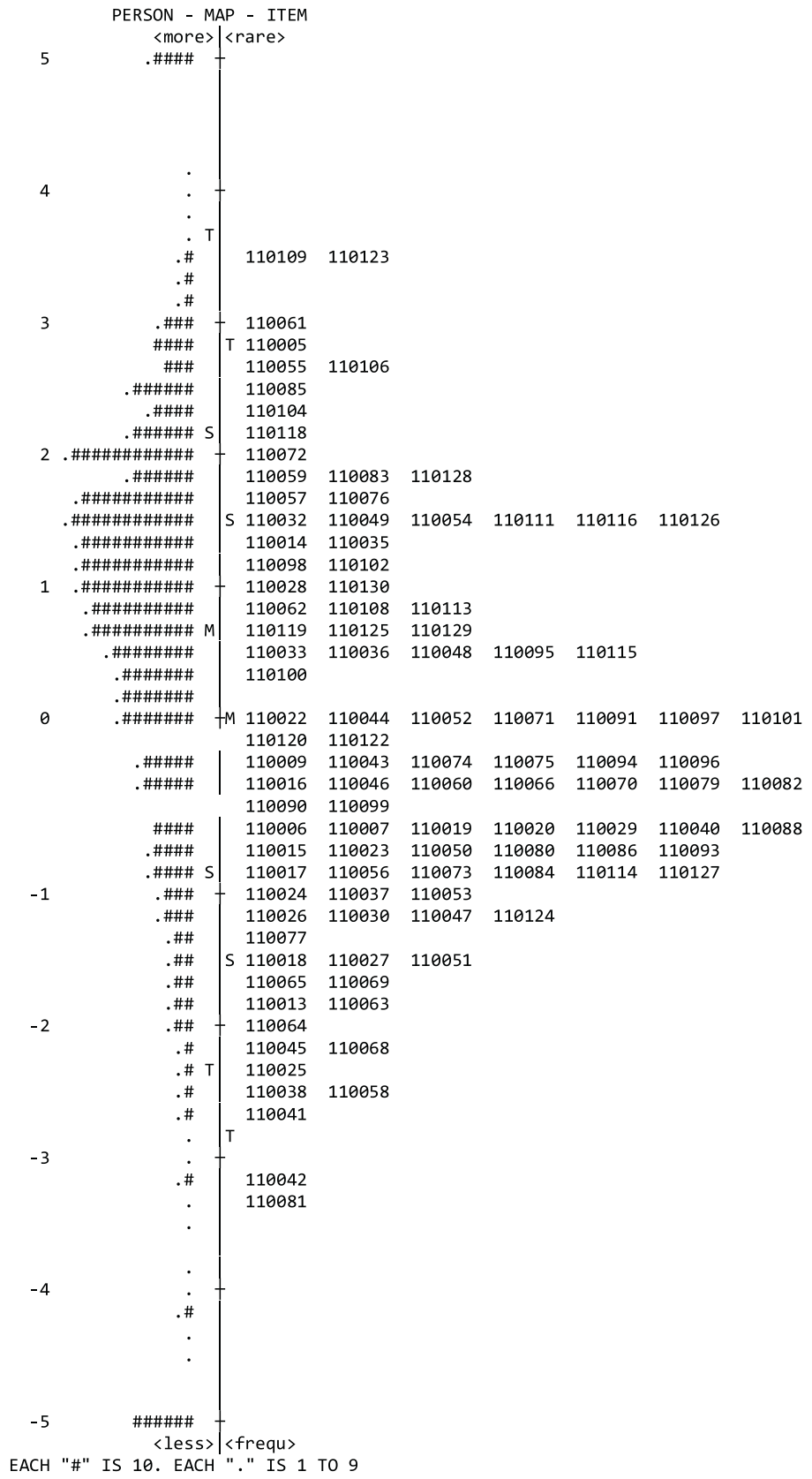

Fig. 2. *Identificación de los Dibujos/Picture Naming* Wright ítem map, with 2283 cases and 106 items.

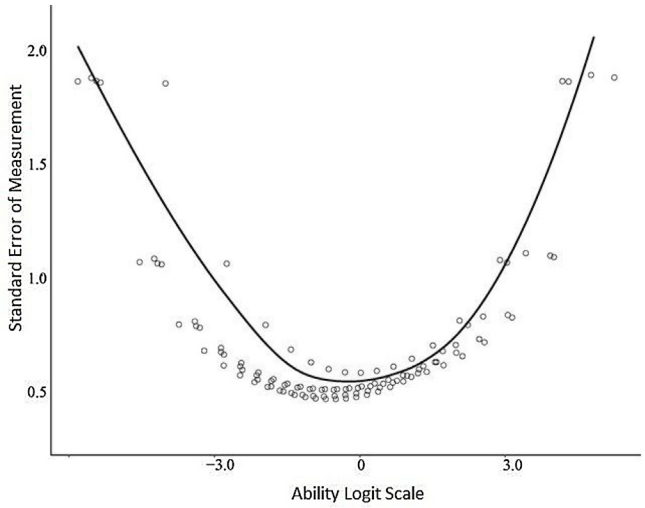

Fig. 3. Standard error of measurement curve for Identificación de los Dibujos/Picture Naming.

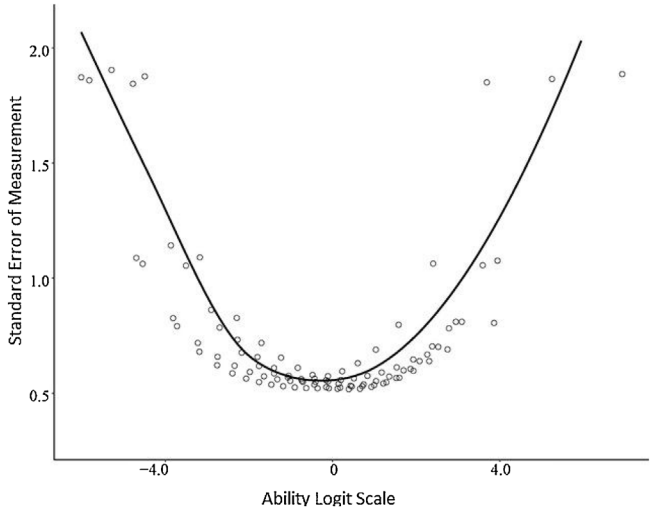

Fig. 4. Standard error of measurement curve for Verbos (Expresivo)/Expressive Verbs.

with those found on large-scale assessments of similar length (Wu, 2010).

9.2.5. Measurement invariance

We explored potential item bias based on dialectal representation using the DIF statistics generated from the Rasch model, using the criteria and methods described in the analysis section. In Identificación de los Dibujos, eight items favored Mexican dialects (were easier for children with Mexican dialects) and nine items favored non-Mexican dialects. In Verbos Expresivo, seven items were found that favored Mexican dialects and six that favored non-Mexican dialects (see Table 7).

Another DIF analysis was conducted using the same procedures exploring item performance differences based on sex. In Identificación de los Dibujos, 17 items demonstrated DIF with seven items favoring girls. In Verbos Expresivo, seven items showed DIF with three items favoring girls (see Table 8). Across both measures, items that demonstrated DIF were removed from the final item pool.

9.2.6. Associations with a related measure

Correlations between IGDIs-E measures and PLS-5 standard scores are presented in Table 9. The PLS-5 is a broad measure of expressive and receptive Spanish language development and is a significantly different measure than the IGDIs-E in terms of how the items on the test were developed, how they are delivered, and the intended purpose of the assessment. We did not anticipate that correlations would be high, but the PLS-5 is one of the most common assessments used in early childhood education for the identification of children with speech and language impairment. Children were selected at random to receive the PLS-5. Given the small numbers it is impossible to compare differences between groups such as children with different levels of proficiency or differences between states. Overall, the correlations between IGDIs-E measures and PLS-5 standard scores were relatively weak. However, two correlations achieved statistical significance: the correlation between Identificación de los Dibujos and the PLS-5 Total Language standard score ($r = 0.31$), and the correlation between Verbos Expresivos and the *PLS*-5 AC standard score ($r = 0.29$).

Validity evidence based on associations with other variables (AERA et al., 2014) have long been the primary source of validity evidence — with too much reliance on this source of evidence. The broader conceptualization of validation, as a validity argument in support of (or repudiation of) the interpretation and use argument for the test scores of interest, presents a more defensible evaluation of the proposed interpretations and uses.

9.2.7. Teacher reports on usability and feasibility

Fourteen teachers responded to the feasibility and utility survey after having the opportunity to interact with and observe administration of the IGDIs-E measures. We completed a descriptive analysis to summarize the teachers' responses on the rating-scale items on the survey. Teachers' narrative responses were synthesized and investigated for themes to inform revisions of the testing or scoring procedures and instructions. Results indicated teachers strongly agreed ($n = 9$) or agreed ($n = 5$) to questions regarding ease of use for Identificación de los Dibujos and Verbos Expresivos. Similarly, 14 teachers noted they *strongly agreed* (10) or *agreed* (4) that the OL measures are an accurate portrayal of Spanish early literacy skills. Teachers reported that administration of Identificación de los Dibujos and Verbos Expresivo were easy to understand and to implement. Two survey questions assessed the scoring ease of understanding and meaningfulness of scores. Eleven teachers *strongly agreed* and three *agreed* that the tests were easy to understand. Six teachers *strongly agreed* and 8 *agreed* that the scores on the IGDIs-E were meaningful. All 14 teachers found the two measures developmentally appropriate for Spanish-speaking preschoolers and also *strongly agreed* (12) or *agreed* (2) the measure was engaging and provided familiar content. Teacher opinions and experiences suggest IGDIs-E oral language measures are appropriate for Spanish-speaking preschoolers. Importantly, teachers acknowledged that the measure was easy to deliver and score.

10. Discussion

Our team undertook the task of developing a Spanish oral language preschool general outcome measure. Technically sound and culturally and linguistically appropriate measures are needed to support early education practitioners to more accurately identify the oral language ability of Spanish-speaking preschoolers in the United States. The primary purpose of the oral language IGDIs-E is to identify children in need of additional language support in preschool with the idea that effective and targeted early intervention and data-based decision-making will prevent the development of later reading problems and language delays. This measure development and calibration study provides initial evidence that IGDIs-E can accurately and adequately evaluate Spanish oral language ability. Below we highlight the contribution of our work to Spanish

Table 7
Differential item functioning results for dialect (Mexican vs. Non-Mexican).

Item ID	Item content	Difficulty for non-Mexican	Difficulty for Mexican	DIF contrast	*p*-Value	Favors
Identificación de los Dibujos/Picture Naming						
110036	Arroz	−0.19	1.63	−1.82	0.000	Non-Mex
110014	Muñeca/mona/bebe de juguete	1.64	1.17	0.46	0.042	Mex
110016	Árbol	0.40	−0.48	0.88	0.015	Mex
110027	Llave	−1.26	−0.69	−0.58	0.045	Non-Mex
110028	Maleta/equipaje/bulto	0.82	1.96	−1.13	0.000	Non-Mex
110029	Tomate/jitomate	1.00	−0.13	1.13	0.000	Mex
110044	Pelota/bola/bolita/balón	0.99	−0.65	1.64	0.000	Mex
110046	Camisa/camiseta/playera	−0.77	−0.19	−0.58	0.005	Non-Mex
110049	Tigre	1.33	2.11	−0.78	0.002	Non-Mex
110052	Escoba	1.32	0.03	1.29	0.000	Mex
110053	Lámpara/luz/luces	−1.20	−1.89	0.69	0.028	Mex
110057	Bate de beisbol/bate	1.53	2.40	−0.87	0.015	Non-Mex
110060	Pie/pies	−0.56	0.30	−0.86	0.017	Non-Mex
110075	Rana/sapo	0.06	0.54	−0.48	0.040	Non-Mex
110080	Ratón/rata	0.09	−0.52	0.61	0.014	Mex
110083	Hormiga	1.33	2.93	−1.60	0.000	Non-Mex
110084	Chile	1.27	−1.10	2.37	0.000	Mex
Verbos (Expresivo)/Expressive Verbs						
140008	Tocar la guitarra	0.66	−0.26	0.92	0.004	Mex
140010	Jugar fútbol/patear	−0.56	−1.16	0.60	0.010	Mex
140017	Pintar	−3.41	−2.17	−1.24	0.010	Non-Mex
140021	Soplar	1.13	1.69	−0.56	0.038	Non-Mex
140047	Hacer compras/comprar	−1.63	−0.76	−0.88	0.014	Non-Mex
140048	Deslizar/resbalar	2.14	0.77	1.37	0.000	Mex
140067	Quemar/incendiar	−0.36	−1.04	0.67	0.004	Mex
140068	Saludar	1.29	0.53	0.76	0.006	Mex
140071	Morder	−1.60	−1.13	−0.47	0.018	Non-Mex
140074	Romper/quebrar	−2.46	−1.71	−0.76	0.007	Non-Mex
140079	Cubrir/cobijar/tapar	1.81	0.61	1.21	0.009	Mex
140084	Aspirar	1.04	0.15	0.89	0.024	Mex
140097	Tirar/aventar/lanzar	−0.61	0.09	−0.70	0.042	Non-Mex

Table 8
Differential Item Functioning Results for Sex (Boys v. Girls).

Item ID	Item content	Difficulty for girls	Difficulty for boys	DIF contrast	*p*-Value	Favors
Identificación de los Dibujos/Picture Naming						
110006	Vaca	0.38	−0.23	0.61	0.036	Boys
110007	Almohada/almohadilla	−0.21	0.36	−0.57	0.006	Girls
110035	Abrigo/chaqueta/chamarra	1.67	0.63	1.04	0.000	Boys
110037	Toalla	−1.07	0.03	−1.10	0.001	Girls
110019	Tren/ferrocarril	0.38	−0.32	0.70	0.002	Boys
110028	Maleta/equipaje/bulto	1.20	1.97	−0.77	0.000	Girls
110043	Espejo	0.02	0.72	−0.70	0.001	Girls
110044	Pelota/bola/bolita/balón	0.40	−0.36	0.75	0.001	Boys
110046	Camisa/camiseta/playera	−0.23	−0.68	0.45	0.013	Boys
110048	Martillo	1.92	1.07	0.85	0.000	Boys
110049	Tigre	2.13	1.38	0.75	0.000	Boys
110054	Vestido/traje	0.68	2.37	−1.70	0.000	Girls
110057	Bate de beisbol/bate	2.96	1.47	1.49	0.000	Boys
110063	Cuchara	−1.81	−1.21	−0.60	0.030	Girls
110075	Rana/sapo	0.55	−0.03	0.58	0.006	Boys
110076	Camión/troca	1.97	1.33	0.65	0.004	Boys
110079	Mariposa	−0.31	0.42	−0.73	0.001	Girls
Verbos (Expresivo)/Expressive Verbs						
140010	Jugar fútbol/patear	−0.68	−1.39	0.70	0.001	Boys
140101	Amontonar/juntar las hojas/ rastrillar/recoger hojas	1.42	2.81	−1.39	0.005	Girls
140048	Deslizar/resbalar	1.55	0.86	0.69	0.003	Boys
140050	Cepillar	−0.33	0.17	−0.49	0.015	Girls
140065	Mendigar/pedir/dar comida/dar de comer	0.99	0.39	0.60	0.049	Boys
140074	Romper/quebrar	−1.69	−2.25	0.55	0.017	Boys
140076	Subir/cerrar/abrochar	−0.07	0.58	−0.65	0.002	Girls

Table 9
Correlations between OL S-IGDI measures and preschool language scales.

	PLS-5 auditory comprehension	PLS-5 expressive communication	PLS-5 total language score
Identificación de los Dibujos/Picture Naming	.28	.24	0.31*
Verbos – Expresivo/ Expressive Verbs	.29*	.12	0.23

Note: All PLS-5 scores used for correlations were standard scores.
* $p < 0.05$.

language general outcome measures in three main areas: innovation is assessment design in Spanish, providing detailed and publicly available information on the psychometric properties, and supporting the assessment of Spanish-speaking preschoolers in Spanish.

10.1. Innovation in assessment design

The work described here is distinguished by its explicit foundation on Spanish language development for Spanish-speaking preschoolers in the United States. There is a dearth of information available on adequate and appropriate processes for developing assessments in languages other than English. An important and innovative feature of our work is that we did not simply translate the English version of the IGDIs, but instead started the process with Spanish language development within the context of the United States as our construct of interest. This led the research team down the path of carefully selecting culturally and linguistically appropriate vocabulary, considering syntactical and grammatical features unique to Spanish, undertaking a comprehensive piloting and field testing process, and including a range of correct responses to our items in the interest of reducing dialectal bias and increasing the accuracy of our ability estimates. We argue that this makes the measures reported here uniquely suited to their primary purpose — measuring the Spanish early language and literacy ability of preschool children being raised in the United States.

A variety of techniques were used in the measure design process, including careful analysis of differential item functioning between children whose families reported being from Mexico versus other Spanish-speaking areas including Puerto Rico and Central and South America. In total, 17 out of 106 Identificación de los Dibujos items and 13 out of 95 Verbos Expresivos items demonstrated DIF and we removed them from the item pool. For some items our research team could discern a rationale for evidence of DIF. For example the "chile" item in Identificación de los Dibujos favored Mexican populations. Chiles are more common in Mexican rather than in Central or South American or Caribbean cuisine. However, for other items there was no discernable rationale for why the item might demonstrate DIF. The item "arbol/tree" in Identificación de los Dibujos also favored Mexican populations, however a tree is arguably a fairly neutral vocabulary target. Approximately 13–19% of our items demonstrated DIF based on Mexican versus non-Mexican sample differences. The number and the specific items with DIF was surprising to our team and emphasizes the need to complete DIF analyses and to not simply rely on expert opinion about what items might work best with certain populations. Without empirically evaluating DIF on oral language measures it is difficult to discern if student performance differences are due to true ability or to bias within the test design.

Another way we captured dialectal differences in the Spanish spoken across populations in the US was to gather child responses during data collection, and after expert and project team review we added many new responses to our scoring key. For example on our Identificación de los Dibujos test the item "baño/bathroom" now has four correct responses including "inodoro, retrete, excusado and baño." The fact that we included multiple correct responses in the measure design process contributes evidence that the IGDIs-E can be used effectively with the range of Spanish-speaking preschoolers found in the United States.

10.2. Empirical item evaluation

We also evaluated each item based on fit statistics. As mentioned in the results section a total of 22 Verbos Expresivo items were removed because they had poor item fit statistics and in Identificación de los Dibujos a total of 16 items were removed. It is not possible to pinpoint precisely why these items did not function well. We speculate that for some items the image we selected simply did not elicit the desired response and for some items children may have been unfamiliar with the specific verb or noun we were targeting. The critical point of this empirical evaluation of the functioning of each item is that all items in the final IGDIs-E screening set have adequate fit statistics and even though items were removed we had developed a large enough initial item pool where over 100 items still remained in each test.

10.3. Psychometric properties

Limited information is available on the psychometric properties of other widely used Spanish early language and literacy screening measures (Barrueco et al., 2012). This is concerning because it is difficult to know how much confidence can be placed in the scores achieved by children and the meaning of performance benchmarks. The IGDIs-E were found to have excellent item statistics with 81% of the *p*-values of all Identificación de los Dibujos items falling in the desired range and 99% of items having adequate point-biserial correlations. In the Verbos Expresivos measure 74% of item *p*-values fell in the desired range and 100% of items had adequate point biserial correlations. In addition, the IGDIs-E items demonstrate strong single-factor dimensionality, as illustrated by the CFA (consistent model-data fit statistics), contributing evidence that the scores from the oral language IGDIs-E measures reflect levels of ability on a single continuum (for each measure). These empirical indices illustrate that the IGDIs-E fit the Rasch model appropriately.

Regarding score use and interpretation, our evaluation of the CSEM shows that IGDIs-E achieve CSEM values similar to high-stakes test ranging from approximately 0.3 to 0.5 logits, depending on the child's ability level (see Fig. 3). These confidence bands given end-users a meaningful context for interpreting scores and the degree to which they can have confidence in a student's observed score. In addition, teachers and end-users also provide information on the feasibility and the utility of the measures, demonstrate their likelihood to be used appropriately and for scores to be meaningfully interpreted in ways that align with instructional decision.

10.4. Assessing preschoolers in Spanish

It is necessary to have technically adequate assessments available in Spanish in order to appropriately assess Spanish-speaking preschoolers. The purpose of general outcome measurement is to effectively target children's instructional levels. However, we run the risk of over-identifying Spanish-speaking children when we only assess them in English with no attention to their language proficiency. Children with limited English proficiency cannot demonstrate their true ability levels in a language they are still in the process of learning. To mitigate these issues, it is important to assess Spanish-English bilinguals in both languages, and IGDIs-E was constructed in response to this need. The IGDIs-E measures are intended to be given alongside the English IGDIs to more accurately and equitably screen Spanish-speaking children. Supplying teachers with information about a child's full language profile can reduce the risk of escalating children through tiers of instruction and potentially decreases inappropriate referrals for special education evaluations. Even in the context of English-only instruction it is important to understand children's' ability in Spanish as an important indicator of their general learning ability and the foundational knowledge they bring to the task of learning English and mastering pre-academic content.

10.5. Policy implications

Currently most publicly funded preschool programs are conducted primarily in English in spite of the fact that there is significant evidence that preschool instruction in Spanish does not hinder English development and it has the added benefit of promoting higher rates of growth in Spanish (Goldenberg, 2008). Several studies have found that children in English-only early education settings demonstrated decreased rates of growth in Spanish (e.g. Barnett, Yarosz, Thomas, Jung, & Blanco, 2007; Durán, Roseth, Hoffman, & Robertshaw, 2013; Farver, Lonigan, & Eppe, 2009). When English-only programs have been compared to two-way immersion programs, advantages of instruction in both languages have been found for children who are native Spanish speakers on Spanish and English receptive vocabulary and early literacy achievement measures such as Spanish rhyming and English phoneme deletion skills (Barnett et al., 2007). Other advantages of instruction in both languages have been found on Spanish measures of receptive and expressive language and letter–word identification in kindergarten (Durán et al., 2013).

If instruction is provided in Spanish it is critical to also have appropriate measures in Spanish to decide whether or not the Spanish instruction is having the desired effect. Therefore, the IGDIs-E could play an important role in evaluating the performance of children in Spanish instructional environments and addressing a significant need for more empirical evidence regarding the efficacy of these programs. The IGDIs-E are designed to be sensitive to the effects of instruction and could be part of an important process to develop effective Spanish early language and literacy intervention and instructional approaches.

10.6. Limitations

Although we intentionally sought to recruit a representative sample of Spanish speakers in the United States that included children from Mexico, the Caribbean, and Central and South America, we were not able to locate large enough samples of non-Mexican populations to investigate differences between these specific groups. However, during data collection, all expressive responses provided by children were recorded and considered, so lists of correct responses in both Identificación de los Dibujos and Verbos Expresivos reflect responses frequently provided by children across all of our samples. We conducted DIF analysis between Mexican and non-Mexican dialects, but in future work larger samples of other dialects should be recruited to take a more fine-grained approach to analyzing other cultural and linguistic differences in responses we might find on items in the IGDIs-E.

Fifty-seven percent of our sample also lived in poverty and this is a little over twice the national average of 21.4% (U.S. Census Bureau, 2013). However, given the range of abilities in the participants, this should not have affected the calibration of our items. Although we know there is an interaction between poverty and language development with children growing up in poverty generally demonstrating lower language development (Sinatra, Zygouris-Coe, & Dasinger, 2012), this overall performance difference should not impact item calibration. In future work it will be important to explore the performance of Spanish-speaking children growing up in families above the poverty line in the US to explore differences in ability levels.

Interestingly, we found only small correlations with the PLS-5. The PLS-5 is a broad measure of Spanish language development and it is possible that the language constructs accessed in this study by the IGDIs-E were not the same as those targeted in the PLS-5. Indeed, the PLS-5 does not describe their underlying construct as Spanish early language and literacy in ways that are unique or different from English. In this way, it may be the case that a low correlation between the measures could be expected as they are measuring two separate constructs. Given the correlations between the PLS-5 and IGDIs-E were relatively low, it may be useful to conduct additional studies with expanded sample sizes to represent specific correlations disaggregated by program type, home language profile, and dialect, to further explore if correlations maintain the same pattern or show differential correlations based on each identified variable. More research may also be needed to investigate associations with other widely used Spanish language assessments in the field to better understand the correlations we may find between the IGDIs-E and other broad oral language assessments.

Measures for use within a MTSS framework also need to be easy and efficient to administer given the broad range of training and background found in early care and education personnel and the limited time allocated to assessment. We included fourteen Spanish-speaking teachers in this study to review the IGDIs-E administration procedures to provide feedback on the feasibility and usability of the measure. Importantly, the majority of teachers reported that the IGDIs-E were easy to understand, deliver, and score. Although clearly more data are needed to fully investigate the utility of the IGDIs-E in the field, this preliminary evidence suggests the measure itself does not present major challenges or barriers to large-scale implementation. However, future studies will need to be conducted to better understand what supports are necessary for the wide scale administration of the IGDIs-E (such as the availability of bilingual staff and training needs) and the subsequent need for support to interpret the testing results for each child and to translate these results to meaningful and effective approaches to intervention.

In this study, we could not account for some variables in the child's environment that might also influence children's performance over the year, such as quality of instruction. The quantity and quality of exposure to English and Spanish need to be considered within the context of the United States as children have been found to have a broad range of exposure to both languages (Bohman et al., 2010). Scores achieved in both languages may be more likely to yield the most accurate information about the development of Spanish-speaking preschoolers within the United States. Future research should compare performance in both languages while including home language exposure and language of instruction in the models to explore these associations and how they might affect performance on the IGDIs-E. There is significant evidence that a rapid shift to English-only settings often found during the preschool years can cause a decrease or plateau in Spanish language skills (Anderson, 2012; Hammer, Lawrence, & Miccio, 2008; Paéz, Tabors, & Lopez, 2007). Without accounting for the influence of these environmental influences we run the risk of inaccurately identifying delay in Spanish when in reality differences in performance may be attributed to lack of Spanish language support rather than child specific low ability levels.

11. Conclusion

We provided a review of the procedures of the development and calibration of the IGDIs-E tasks. The IGDIs-E are a promising measure for the early identification of young Spanish speakers in need of more systematic and explicit instruction to address language and literacy instructional needs before they escalate to full referrals for special education services. Early intervention has the potential to address achievement gaps early on and prevent school failure in a population of students that has historically been marginalized. Meaningful and technically sound screening measures are a critical component and often a first step in early education to provide targeted and effective intervention.

Acknowledgements

This work was supported in part by grant R305A120449, Re-search and Development of Spanish Individual Growth and Development Indicators: Early Literacy identification measures for Spanish–English bilingual children, from the Institute of Education Sciences, U.S. Department of Education. The authors would like to thank colleagues who assisted with this project including participating childcare centers and programs in the Minneapolis/St. Paul area, central Florida, the Bay Area in California, and rural Utah. However, the opinions and recommendations presented in this paper are those of the authors alone, and no official endorsement from the Institute of Education Sciences should be inferred.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.ecresq.2019.02.001.

References

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

American Speech and Hearing Association. (2017). *Issues in ethics: Cultural and linguistic competence* Retrieved from. http://www.asha.org/Practice/ethics/Cultural-and-Linguistic-Competence/

Anaya, J. B., Peña, E. D., & Bedore, L. M. (2016). Where Spanish and English come together: A two dimensional bilingual approach to clinical decision making. *ASHA Perspectives SIG 14*, *1*(1), 1–14.

Anderson, R. T. (2012). First language loss in Spanish-speaking children: Patterns of loss and implications for clinical practice. In B. A. Goldstein (Ed.), *Bilingual language development and disorders in Spanish–English speakers* (pp. 187–212). Baltimore, MD: Brookes.

August, D., & Shanahan, T. (Eds.). (2006). *Developing literacy in second language learners (report of the national literacy panel on language-minority children and youth)*. Mahwah, NJ: Lawrence Erlbaum.

Banerjee, R., & Luckner, J. L. (2010). Assessment practices and training needs of early childhood professionals. *Journal of Early Childhood Teacher Education*, *34*, 231–248.

Barnett, W. S., Yarosz, D. J., Thomas, J., Jung, K., & Blanco, D. (2007). Two way and monolingual English immersion in preschool education: An experimental comparison. *Early Childhood Research Quarterly*, *22*, 277–293.

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

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

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 Services in Schools*,

Children's Defense Fund. (2017). *State of America's Children Yearbook 2017*.

Durán, L., & Wackerle-Hollman, A. (2016). *Language and literacy environement exposure report. Unpublished survey*. Minneapolis, MN: University of MN.

Bohman, T. M., Bedore, L. M., Peña, E. D., Mendez-Perez, A., & Gillam, R. B. (2010). What you hear and what you say: Language performance in Spanish–English bilinguals. *International Journal of Bilingual Education and Bilingualism*, *13*, 325–344.

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed). New York, NY: Routledge.

Brown, J. E., & Sanford, A. K. (2011). *RTI for English language learners: Appropriately using screening and progress monitoring tools to improve instructional outcomes* Retrieved from. Washington, DC: U.S. Department of Education, Office of Special Education Programs, National Center on Response to Intervention. http://www.rti4success.org/images/stories/pdfs/rtiforells.pdf

Castilla, A. P., Restrepo, M. A., & Perez-Leroux, A. T. (2010). Individual differences and language interdependence: A study of sequential bilingual development in Spanish English preschool children. *International Journal of Bilingual Education and Bilingualism*, *12*(5), 565–580.

Davies, M. (2006). *A frequency dictionary of Spanish: Core vocabulary for learners*. New York, NY: Routledge.

Davison, M. D., Hammer, C., & Lawrence, F. R. (2011). Associations between preschool language and first grade reading outcomes in bilingual children. *Journal of Communication Disorders*, *44*, 444–458.

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

Division for Early Childhood of the Council for Exceptional Children. (2010). *Responsiveness to ALL children, families, and professionals: Integrating cultural and linguistic diversity into policy and practice (Position Statement)* Retrieved from. Missoula, MT: Author. http://www.dec-sped.org/uploads/docs/about_dec/position_concept_papers/Position%20Statement_Cultural%20and%20Linguistic%20Diversity_updated_sept2010.pdf

Durán, L. K., Roseth, C., Hoffman, P., & Robertshaw, M. B. (2013). An experimental study comparing predominantly English and transitional bilingual education on Spanish-speaking preschoolers' early literacy development: Year three results. *Bilingual Research Journal*, *36*(1), 6–34.

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.

Fien, H., Smith, J. L. M., Baker, S. K., Chaparro, E., Baker, D. L., & Preciado, J. A. (2011). Including English learners in a multitiered approach to early reading instruction and intervention. *Assessment for Effective Intervention*, *36*, 143–157.

Figueras-Daniel, A., & Barnett, W. S. (2013). *Preparing young Hispanic dual language learners for a knowledge economy. Preschool Policy Brief. Issue 24*. New Brunswick, NJ: National Institute of Early Education Research, Rutgers.

Fuchs, L. S., & Deno, S. L. (1991). Paradigmatic distinctions between instructionally relevant measurement models. *Exceptional Children*, *57*(6), 488–500.

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

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

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

Goodrich, J. M., & Lonigan, C. J. (2017). Language independent and language-specific aspects of early literacy: An evaluation of the common underlying proficiency model. *Journal of Educational Psychology*, *109*(6), 782–793.

Goodrich, J. M., Lonigan, C. J., Kleuver, C. G., & Farer, J. M. (2016). Development and transfer of vocabulary knowledge in Spanish-speaking language minority preschool children. *Journal of Child Language*, *43*, 969–992.

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

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

Hammer, C. S., & Rodriguez, B. L. (2012). Bilingual language acquisition and the child socialization process. In B. Goldstein (Ed.), *Bilingual language development & disorders* (pp. 31–46). Baltimore, MD: Paul H. Brooke Publishing.

Hammer, C. S., Komaroff, E., Rodriguez, B. L., Lopez, L. M., Scarpino, S. E., & Goldstein, B. (2012). Predicting Spanish–English bilingual children's language abilities. *Journal of Speech, Language, and Hearing Research*, *55*, 1251–1264.

Hammer, C., Lawrence, F. R., & Miccio, A. W. (2008). Exposure to English before and after entry into Head Start: Bilingual children's receptive language growth in Spanish and English. *International Journal of Bilingual Education and Bilingualism*, *11*, 30–56.

Hemphill, F. C., & Vanneman, A. (2010). *Achievement gaps: How Hispanic and White students in public schools perform in mathematics and reading on the National Assessment of Educational Progress (NCES 2011-459)*. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Hoff, E. (2013). Interpreting the early language trajectories of children from low-SES and language minority homes: Implications for closing achievement gaps. *Developmental Psychology*, *49*.

Hoff, E., & Core, C. (2015). What clinicians need to know about bilingual development. *Seminars in Speech and Language*, *36*, 89–99.

Jackson, C. W., Schatshneider, C., & Leacox, L. (2014). Longitudinal analysis of receptive vocabulary growth in young Spanish English-speaking children from migrant families. *Language, Speech, and Hearing Services in Schools*, *45*, 40–51.

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

Kane, M. (2013). The argument based approach to validation. *School Psychology Review*, *42*(4), 448–457.

Klingner, J. K., & Edwards, P. E. (2006). Cultural considerations with response to intervention models. *Reading Research Quarterly*, *41*, 108–117.

Kroll, J., Van Hell, J., Tokowicz, N., & Green, D. (2010). The revised hierarchical model: A critical review and assessment. *Bilingualism: Language and Cognition*, *13*(3), 373–381.

Linacre, J. M. (2016a). *Winsteps® (Version 3.91) [Computer Software]* Retrieved from. Beaverton, Oregon: Winsteps.com. http://www.winsteps.com/

Linacre, J. M. (2016b). *Winsteps® Rasch measurement computer program user's guide* p. 433. Beaverton, Oregon: Winsteps.com. Retrieved from. http://www.winsteps.com/

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

Lonigan, C. J., & Shanahan, T. (2009). Developing early literacy: Report of the national early literacy panel. Executive summary. a scientific synthesis of early

literacy development and implications for intervention. *National Institute for Literacy*.
Mancilla-Martinez, J., & Lesaux, N. K. (2010). Predictors of reading comprehension for struggling readers: The case of Spanish-speaking language minority learners. *Journal of Educational Psychology*, *102*(3), 701–711.
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.
McConnell, S. R., Wackerle-Hollman, A. K., Roloff, T. A., & Rodriguez, M. (2015). Designing a measurement framework for response to intervention in early childhood programs. *Journal of Early Intervention*, *36*, 263–280.
Melby-Lervåg, M., & Lervåg, A. (2011). Cross-linguistic transfer of oral language, decoding, phonological awareness and reading comprehension: A meta-analysis of the correlational evidence. *Journal of Research in Reading*, *34*(1), 114–135.
Miller, J. F., Heilmann, J., Nockerts, A., Iglesias, A., Fabiano, L., & Francis, D. J. (2006). Oral language and reading in bilingual children. *Learning Disabilities Research & Practice*, *21*, 30–43.
Murphey, D., Guzman, L., & Torres, A. (2014). *America's Hispanic children: Gaining ground, looking forward*. Bethesda, MD: Child Trends Hispanic Institute.
Muthén, L. K., & Muthén, B. O. (2012). *MPlus: statistical analysis with latent variables-user's guide*.
National Association for the Education of Young Children. (2005). *Screening and assessment of young English-language learners: Supplement to the NAEYC and NAECS/SDE joint position statement on early childhood curriculum, assessment, and program evaluation* Retrieved from. National Association for the Education of Young Children. http://www.naeyc.org/files/naeyc/file/positions/ELL_Supplement_Shorter_Version.pdf
National Center for Education Statistics. (2015). *The condition of education 2015 (NCES 2015-144), English Language Learners*. Washington, DC: U.S. Department of Education.
Palermo, F., Mikulski, A. M., Fabes, R. A., Martin, C. L., & Hanish, L. D. (2017). Cross-language associations and changes in Spanish-speaking preschoolers' English and Spanish academic abilities. *Applied Psycholinguistics*, *38*(2), 347–370.
Páez, M. M., Tabors, P. O., & López, L. M. (2007). Dual language and literacy development of Spanish-speaking preschool children. *Journal of Applied Developmental Psychology*, *28*, 85–102.
Peña, E. D. (2007). Lost in translation: Methodological considerations in cross-cultural research. *Child Development*, *78*, 1255–1264.
Peña, E. D., Gutierrez-Clellan, V. F., Iglesias, A., Goldstein, B. A., & Bedore, L. (2014). *Bilingual english-Spanish assessment*. Baltimore, MD: Brookes Publishing.
Peña, E. D., Kester, E. S., & Sheng, L. (2012). Semantic development in Spanish-English bilinguals: Theory, assessment, & intervention. In B. Goldstein (Ed.), *Bilingual language development & disorders* (pp. 131–152). Baltimore, MD: Paul H. Brooke Publishing.
Pena, E., Bedore, L. M., & Rappazzo, C. (2003). Comparison of Spanish, English, and bilingual children's performance across semantic tasks. *Language, Speech, and Hearing Services in Schools*, *34*, 5–16.
Proctor, C. P., August, D., Carlo, M. S., & Snow, C. (2006). The intriguing role of Spanish language vocabulary knowledge in predicting English reading comprehension. *Journal of Educational Psychology*, *98*(1), 159–169.
Reardon, S. F., & Galindo, C. (2009). The Hispanic-white achievement gap in math and reading in the elementary grades. *American Educational Research Journal*, *46*, 853–891.
Reese, L., Linan-Thompson, S., & Goldenberg, C. (2008). Variability in community characteristics and Spanish-speaking children's home language and literacy opportunities. *Journal of Multilingual and Multicultural Development*, *29*, 271–290.
Sanford, A. K., & Brown, J. E. (2011). RTI for English language learners: Appropriately using screening and progress monitoring tools to improve instructional outcomes.
Sinatra, R., Zygouris-Coe, V., & Dasinger, S. B. (2012). Preventing a vocabulary lag: What lessons are learned from research. *Reading & Writing Quarterly*, *28*(4), 333–357.
Tomasello, M. (2005). *Constructing a language: A usage-based theory of language acquisition*. Boston, MA: Harvard University Press.
U.S. Census Bureau. (2013). *USA quick facts* Retrieved from. http://quickfacts.census.gov/qfd/states/00000.htmle
Wackerle-Hollman, A., Durán, L., Brunner, S., Kohlmeier, T., Palma, J., & Rodriguez, M. (2019). Spanish individual growth and development indicators: the development and validation of phonological awareness tasks. *Educational Assessment*, *24*(1), 33–56.
Wackerle-Hollman, A., Durán, L., Rodriguez, M., Brunner, S., & Palma, J. (2016). *Spanish individual growth and development indicators technical manual*. Minneapolis, MN: University of MN.
Whitehurst, G. J., & Lonigan, C. J. (2001). Emergent literacy: Development from prereaders to readers. *Handbook of early literacy research*, *1*, 11–29.
Wiig, E. H., Secord, W. A., & Semel, E. (2009). *Clinical evaluation of language fundamentals-preschool*. Spanish Edition [CELF-P2 Spanish].
Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.
Wright, B. (1996). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *3*, 3–24.
Wu, M. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice*, *29*(4), 15–27.
Zimmerman, I. L., Steiner, V., & Pond, R. E. (2012). *Preschool language Scale-5*. Birmingham, AL: Psych Corp.
Zwick, R. (2012). *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement* Retrieved from. Princeton, NJ: Educational Testing Service. https://www.ets.org/Media/Research/pdf/RR-12-08.pdf