AJSLP

Research Article

Utility of a Language Screening Measure for Predicting Risk for Language Impairment in Bilinguals

Mirza J. Lugo-Neris,[a] Elizabeth D. Peña,[a] Lisa M. Bedore,[a] and Ronald B. Gillam[b]

Purpose: This study evaluated the accuracy of an experimental version of the Bilingual English Spanish Oral Screener (BESOS; Peña, Bedore, Iglesias, Gutiérrez-Clellen, & Goldstein, 2008) for predicting the long-term risk for language impairment (LI) for a matched group of preschool-aged Spanish–English bilingual children with and without LI.
Method: A total of 1,029 Spanish–English bilingual children completed the BESOS before entering kindergarten. A subset of 167 participants completed a follow-up language evaluation in 1st grade. Twenty-one of these children were identified as having LI and were matched to a group of 21 typically developing peers from the larger sample. A series of discriminant analyses were used to determine the combination of scores on the BESOS that most accurately predicted 2 years later which children presented with and without LI.
Results: The linear combination of the semantics and morphosyntax scores in the best language resulted in predictive sensitivity of 95.2% and predictive specificity of 71.4%, with an overall accuracy of 81% for predicting risk for LI.
Conclusion: A bilingual language screener administered before kindergarten can be useful for predicting risk for LI in bilingual children in 1st grade.

Language screening can be a useful tool in determining whether a preschool-age child is at risk for language impairment (LI; Washington & Craig, 2004). Gold standard procedures are, unfortunately, lacking for language screening of both monolingual and bilingual preschoolers (Dockrell & Marshall, 2015; Nelson, Nygren, Walker, & Panoscha, 2006). The effectiveness and efficiency of currently available screening methods in accurately predicting which English language learners (ELLs) are at risk for LI is unknown. The school population of ELLs has increased by more than half a million between 2002 and 2010 (Aud et al., 2013), with Spanish being the most common language spoken (Batalova & McHugh, 2010). Approximately 7% of monolingual English-speaking school-age children will have LIs that may interfere with their ability to profit from school instruction (Tomblin et al., 1997). The prevalence of LI among bilingual children appears to be similar (Gillam, Peña, Bedore, Bohman, & Mendez-Perez, 2013). Establishing the predictive effectiveness of a screener for Spanish–English bilingual children is critical given the high proportion of children in preschools and kindergartens who are ELLs. This study aims to identify the best combination of scores on an experimental version of the Bilingual English Spanish Oral Screener (BESOS; Peña, Bedore, Iglesias, Gutiérrez-Clellen, & Goldstein, 2008) that most accurately predicts which children will be at risk for LI in first grade when screened before entering kindergarten.

Preschool Language Screening

Screening is a brief procedure used to determine which children should be referred for further testing (American Speech-Language-Hearing Association, 2004; Grimes & Schulz, 2002). Preschool language screening is commonly carried out by primary care physicians at well-child visits or by educators, speech-language pathologists (SLPs), or researchers at clinics, day care facilities, preschools, and schools. Current best practice for screening preschoolers for language delays or disorders typically involves a combination of formal norm-referenced tests, developmental checklists, informal tests, and clinical judgment (Dockrell

[a]The University of Texas at Austin
[b]Utah State University, Logan

Correspondence to Mirza J. Lugo-Neris: mirzalugo@utexas.edu

Editor: Krista Wilkinson
Associate Editor: Carol Miller

Received April 23, 2014
Revision received October 19, 2014
Accepted April 10, 2015
DOI: 10.1044/2015_AJSLP-14-0061

Disclosure: Elizabeth D. Peña and Lisa M. Bedore have a financial interest in the Bilingual English Spanish Assessment. Ronald B. Gillam has a financial interest in the Test of Narrative Language.

& Marshall, 2015; Law, Boyle, Harris, Harkness, & Nye, 1998; Nelson et al., 2006).

Bilingual children are a unique population when it comes to screening because their linguistic performance may be more variable relative to their monolingual peers, most likely because of the way their experience with and knowledge of languages are distributed (Valdés & Figueroa, 1994). At the preschool and kindergarten levels, professionals may underrefer bilingual children who may be at risk for LI because they believe that their difficulties with language may be associated with learning a second language (Bedore & Peña, 2008; Samson & Lesaux, 2009). Clinicians and educators benefit from the availability of formal standardized tools that aid in making reliable and objective clinical judgments about which bilingual preschoolers are at risk for LI and would benefit from watchful surveillance or referral for further evaluation.

Ideally, best practice for evaluating bilinguals for LI would be to assess their linguistic knowledge in both languages across multiple domains, thus controlling for differences in exposure to and use of each language (Bedore & Peña, 2008; Bohman, Bedore, Peña, Mendez-Perez, & Gillam, 2010; Kohnert, 2010). Obtaining the necessary information from bilingual children across two languages in a short time frame is challenging not only because of the lack of instruments available but also because of the lack of bilingual personnel available to administer such measures. A language screener is meant to be quick, efficient, and require minimal effort from both the examinee and examiner (Friberg, 2010; McCauley & Swisher, 1984; Spaulding, Plante, & Farinella, 2006; Warner, 2004; J. M. G. Wilson & Jungner, 1968).

At present, there is only one commercially available norm-referenced screener in Spanish: the Preschool Language Scales Spanish Screening Test–Fifth Edition (PLSSST-5; Zimmerman, Steiner, & Pond, 2012c). This screener is first administered in Spanish and, subsequently, incorrect items—particularly for semantics—can then be administered in English using a scripted direct translation. Scores can be derived for Spanish-only and dual-language administration. The norming sample for the PLSSST-5 includes two thirds monolingual Spanish speakers and one third Spanish–English bilinguals who are dominant in Spanish. This model for test administration can provide valuable information about children's conceptual knowledge of words across languages (Peña, Bedore, & Rappazzo, 2003). However, clinical markers for LI vary across English and Spanish, particularly in the area of morphosyntax (Bedore & Leonard, 2001, 2005; Gutiérrez-Clellen, Restrepo, & Simón-Cereijido, 2006; Gutiérrez-Clellen & Simón-Cereijido, 2007). Therefore, systematically collecting detailed information in each language could potentially help differentiate risk for impairment.

There are two major challenges for validating and interpreting a formal screening tool for bilinguals. One is that there is no standard of optimal classification or predictive accuracy to use to evaluate the screening measure. Another challenge is that, although best practices for bilinguals include sampling multiple languages across multiple domains (Bedore & Peña, 2008), there are few studies providing evidence-based direction for the clinical interpretation of obtained scores.

Considerations for the Classification Accuracy of Language Screeners

Measures of classification accuracy indicate how well a screening test discriminates between two groups of children—in this case, children who are at risk for LI or who are typically developing (TD). The sensitivity and specificity of a measure can be assessed either concurrently or predictively. *Sensitivity* refers to the proportion of individuals who are at risk for a disorder who score in the at-risk range on a screener, and *specificity* is the proportion of individuals who are not at risk who score in the normal range on a screener (Dollaghan, 2007). Most screener validation studies report concurrent classification accuracy, meaning that they administered the screener and a diagnostic reference standard within the same time frame (Dockrell & Marshall, 2015). Concurrent accuracy is informative because it evaluates how well a screener classifies children who are at risk for a disorder on the basis of a lengthier, more established measure. Knowing this information can save time by ruling out impairment and focusing clinicians' efforts on following up on those children identified at risk. Predictive accuracy involves predicting risk across longer periods of time. Studies have been unable to identify consistent factors that can accurately predict risk for LI over time (Law, Rush, Anandan, Cox, & Wood, 2012; Nelson et al., 2006; P. Wilson, McQuaige, Thompson, & McConnachie, 2013). Predictive accuracy is important because these results would indicate whether the child is likely to continue to demonstrate risk at a later age. Predictive accuracy of language screeners is difficult to establish, and few screeners provide this information.

Predictive accuracy of screeners can have an impact on educational opportunities for a child who is not identified correctly as a result of being screened with that instrument. Screening programs that underrefer children at risk for LI can contribute to delayed identification and interventions at early ages (Dollaghan, 2007; Glascoe, 2001). In contrast, a very high rate of overreferral can also be problematic, as the associated costs of further evaluation may have an impact on already-strained resources in educational or clinical settings (Nelson et al., 2006). However, specific criteria for an ideal level of predictive accuracy for screeners have not been established.

More than 20 years ago, Plante and Vance (1994) proposed that diagnostic language tests should have sensitivity and specificity values of more than 80% to be considered acceptable, with 90% being optimal. At present, only about 30% of commonly used English standardized language batteries, two Spanish-only language tests, and one Spanish–English bilingual test meet this criterion (Friberg, 2010; Peña, Gutiérrez-Clellen, Iglesias, Goldstein, & Bedore, 2014; Spaulding et al., 2006; Wiig, Secord, & Semel, 2004; Zimmerman, Steiner, & Pond, 2012a).

Although Plante and Vance's recommendation for diagnostic measures is a good starting point, it is likely that the standards would be somewhat different for screening instruments. The purpose of a screener is to identify risk for a disorder (Grimes & Schulz, 2002). In educational contexts, children who fail a language screening are referred for follow-up testing and evaluations to determine whether they truly do have LI. Because screening will not result in an immediate diagnosis, there can be more flexibility in the accepted levels of specificity than in diagnostic tests, tolerating levels as low as 70%–80% (Barnes, 1982; Bright Futures Steering Committee and Medical Home Initiatives for Children with Special Needs Project Advisory Committee, 2006). Some researchers recommend that screeners should have high sensitivity, so as to maximize the number of children identified who truly have the disorder (Warner, 2004; J. M. G. Wilson & Jungner, 1968).

An important factor to consider in evaluating the accuracy of a screener is the diagnostic reference standard that is utilized to identify whether a child truly is or is not at risk for impairment (Dollaghan, 2007; Warner, 2004). It is unfortunate that there is not a single gold-standard measure for diagnosing LI in monolingual or bilingual children (Dollaghan & Horner, 2011; Law et al., 1998; Nelson et al., 2006). Test validation studies have reported a variety of reference standards, including clinical judgment, parent and teacher report, the child's current individualized education program status, participation in speech and language therapy, a battery of standardized tests, or a combination of any of the above (Allen & Bliss, 1987; Blaxley, Clinker, & Warr-Leeper, 1983; Fluharty, 2001; Gillam et al., 2013; Gutiérrez-Clellen & Simón-Cereijido, 2007; Gutiérrez-Clellen, Simón-Cereijido, & Wagner, 2008; Illerbrun, Haines, & Greenough, 1985; Jacobson & Schwartz, 2005; Merrell & Plante, 1997; Perona, Plante, & Vance, 2005; Restrepo, 1998; Shipley, Stone, & Sue, 1983; Sturner, Funk, & Green, 1996). The comparison diagnostic test can be the parent test of the screener (Zimmerman, Steiner, & Pond, 2012b, 2012c), but that is problematic because items that appear in both the screener and the parent test can result in inflated sensitivity and specificity values. Selecting different diagnostic referents can have an impact on comparisons across different validation studies of the same measure.

Our informal review of available English preschool screeners (see Table 1; Nelson et al., 2006; Sturner, Layton, Evans, Funk, & Machon, 1994) reported wide ranges of concurrent sensitivity and specificity, many of which were well below Plante and Vance's (1994) recommendation. Some of these values are available directly in test manuals, and others have been tested independently across several research studies. As a group, these English screeners underrefer monolingual English-speaking preschoolers who may be at risk for LI. We do not know how well these English screeners predict risk in bilinguals, but we can speculate that they would both under- and overidentify.

There is a need for an efficient and accurate screening measure that samples clinical markers of both English and Spanish. When administered to Spanish speakers (or Spanish-dominant bilinguals), the PLSSST-5 has concurrent sensitivity of 95% and specificity of 79% using its parent test as the reference standard (Preschool Language Scales–Fifth Edition Spanish; Zimmerman et al., 2012a). This result demonstrates a higher sensitivity value and a lower specificity value than those reported for the English measures administered to monolingual English speakers (see Table 1). We currently do not know what the predictive accuracy of the PLSSST-5 would be for a sample of more balanced or English-dominant bilingual children or how using a different reference standard would affect its accuracy. Because of the lack of available bilingual tools, to test across languages clinicians often turn to commercially available instruments in English (see Table 1) and combine them with additional formal or informal measures in Spanish (i.e., language sampling).We do not know how accurate these published screeners or informal measures are for Spanish–English bilinguals.

Interpretation of Screening Scores

Establishing the optimal cutoff value for clinical interpretation of risk may also have an impact on predictive accuracy in bilingual screening (Warner, 2004). If the cutoff is set too low, it may fail to detect children who are at risk for LI, and if it is too high, then too many children will be misidentified as being at risk and unnecessarily referred for follow-up. For monolingual children, cutoffs are sometimes set on the basis of arbitrary points set by test developers or school districts (i.e., 1.5–2.0 *SD*s below the mean). Cutoffs can also be empirically derived on the basis of mean differences between the average score of a TD group and a group with LI. For example, for diagnosing LI in monolinguals, Tomblin, Records, and Zhang (1996) established a cutoff of 1.14 *SD*s below the mean on a composite score of several standardized measures of English language ability (standard scores below 82.9).

Establishing appropriate cut-points for bilingual screening instruments is important. A single cutoff value applied across languages or domains may not accurately classify all children across the bilingual experience continuum (Bedore et al., 2012). We do not know whether arbitrary cutoffs are applicable to bilinguals because the scores of culturally and linguistically diverse children with and without LI typically do not show clear separation. This overlap in performance makes it challenging to differentiate risk for impairment (Oetting, Cleveland, & Cope, 2008; Spaulding et al., 2006; Tomblin et al., 1996).

Bilingual children can present with different levels of linguistic performance depending on the domains tested (Bedore et al., 2012). It is not uncommon for monolingual TD children to show different profiles of ability across language domains (i.e., semantics, morphosyntax). However, with bilinguals, profiles of differential skills can vary by domain and by language. This is sometimes referred to as *mixed dominance* (Bedore et al., 2012). For example, a TD bilingual child may exhibit strengths in semantic skills in one language and morphosyntax in the other. Because

Table 1. Commonly used English preschool language screeners.

Screening test	Language	Classification accuracy	
		Sensitivity	Specificity
Fluharty Preschool Speech and Language Screening Test (Allen & Bliss, 1987; Blaxley et al., 1983; Fluharty, 2001; Illerbrun et al., 1985)	English	36–65[a]	93–96[a]
Sentence Repetition Screening Test (Sturner et al., 1996)	English	76	92
Test for Examining Expressive Morphology (Merrell & Plante, 1997; Perona et al., 2005; Shipley et al., 1983)	English	88.1–90.0[a]	85.4–95.0[a]
Preschool Language Scales Screening Test–Fifth Edition (Zimmerman et al., 2012b)	English	86	96

[a]Ranges are reported because validation studies have reported differing values for classification accuracy across multiple studies and samples.

bilingual children in the United States are a heterogeneous group with varying levels of exposure to and use of each language (Bedore et al., 2012; Valdés & Figueroa, 1994), they may display these mixed dominance profiles because of differences in the rate and order of acquisition of each of the languages they are learning. Mixed dominance can further cloud the clinical picture when trying to predict risk for LI because children who are acquiring English as a second language may make errors that could overlap with those made by English monolinguals with LI (Gutiérrez-Clellen et al., 2008; Paradis, 2005). A single cutoff across all domains and languages tested may not differentiate which patterns belong to the typical progression of second language acquisition versus those that characterize LI.

In addition to the issue of setting appropriate cutoffs, a related challenge in screening bilingual children is that there is no clear agreed-upon way to combine the obtained scores from multiple languages or domains for clinical interpretation (Core, Hoff, Rumiche, & Señor, 2013; Thordardottir, Rothenberg, Rivard, & Naves, 2006). Bilingual children who score very high or very low in one or both languages would be relatively easy to classify as having an impairment or not, but the determination process is not so straightforward for children with mixed dominance profiles or those with differing skills across domains. Thus, it is important to establish an empirical method to establish cutoffs that will facilitate the interpretation of test scores across the multiple domains and languages tested.

Development of the BESOS

In an effort to obtain a brief measure of linguistic knowledge of Spanish–English bilingual children living in the United States, a team of researchers has developed the BESOS (Peña et al., 2008). This experimental screening tool has been utilized in studies identifying clinical markers of LI or enrollment in a response to an intervention program (Bedore et al., 2012; Gillam et al., 2013; Greene, 2012; Peña, Gillam, Bedore, & Bohman, 2011). Derived from the most discriminating items in the experimental Bilingual English Spanish Assessment (BESA; Peña et al., 2014) item set, the BESOS consists of semantics and morphosyntax subtests in Spanish and English. A bilingual examiner administers each language section individually, allowing for responses in either language on semantics, resulting in a conceptual score. Four subtest scores are derived. The norming sample includes bilingual children growing up in the United States who have a broad range of exposure and use of both Spanish and English. Validation studies are still underway, but preliminary validation data for the BESOS indicate concurrent sensitivity of 90% and specificity of 91%.

We have begun exploring ways to combine scores across domains and languages to obtain the clearest and most complete picture of a bilingual child's linguistic knowledge and ability (Peña & Bedore, 2011; Peña et al., 2014). For example, Peña and Bedore (2011) have illustrated how using a single cut-point on the four BESOS subtests may yield different classification accuracy depending on the languages included in the analysis. Participants were bilinguals who were dominant in English (e.g., used English 60%–80% of the time) or in Spanish (e.g., used Spanish 60%–80% of the time). When using a cutoff score of 1 *SD* below the mean on the child's reported dominant language scores, they obtained 100% sensitivity, but specificity was 57% for English-dominant children and 70% for Spanish-dominant children. Comparing performance in each language and then selecting a child's best score on each administered subtest (semantics and morphosyntax) resulted in an improved specificity of 81%. Although they did not empirically derive the cutoff, selecting the best score in each domain resulted in acceptable classification accuracy on the basis of Plante and Vance's (1994) recommendation.

The Present Study

Previous studies have not identified optimal levels of sensitivity and specificity, appropriate cutoffs, or combination of scores that best represent a bilingual child's linguistic ability on language screening tests. Studies also have not established the ability of screening tools to predict longer term risk for LI. The purpose of this study was to identify the combination of scores that maximized the predictive (long-term) classification accuracy of the BESOS for a matched group of children with and without LI.

The specific research questions were as follows:

1. What is the predictive classification accuracy of the BESOS subtests for a group of matched TD/LI children in first grade when administered prior to entering kindergarten?
2. What is the combination of scores on the BESOS in preschool that best discriminates group membership (LI/TD) in first grade?

Method

Participants

Forty-two participants were selected from a data set of a longitudinal study of diagnostic markers of LI (Gillam et al., 2013). In the larger study, 1,029 participants were recruited from three different school districts in Utah and Texas. The BESOS was administered before the children started kindergarten. The 42 participants who were the focus of this study were chosen from a subset of 167 children who scored at the 30th percentile or below on at least two of the four subtests of the BESOS and had at least 20% of exposure to and use of both English and Spanish. Participants were then followed for 2 years, and a determination of the presence of LI was made in first grade.

Twenty-one children had LI on the basis of data collected in first grade in the larger study. These children were matched to 21 children with TD language skills (TD-matched; Squires, Lugo-Neris, Peña, Bedore, & Gillam, 2014) within the larger sample. Matching was based on sex, age in months at time of final testing (within 5 months; *M* difference = 1.86 months), month of birth (within 4 months; *M* difference = 1.31 months), semester of testing, nonverbal IQ score (within 1 *SD*; *M* difference = 9.05) on the Universal Nonverbal Intelligence Test (Bracken & McCallum, 1998), and language exposure. Matches were selected to be within 20% English and Spanish input and output (on average, matches were within 12.6%). We also matched on parental report of the age of first exposure to English within 2 years, and, on average, matches were within 0.85 years. There were no significant differences across groups on any of the matching measures.

Participants were 18 girls and 24 boys between the ages of 4;6 (years;months) and 6;2. Of the 90.5% of the sample who reported, 78.6% received free or reduced-price lunch at school. Information collected from detailed parent interviews using the Bilingual Input Output Survey (Peña et al., 2014) showed that participants' input and output in English ranged from 22% to 74% of the time (see Table 2).

Measures

The BESOS consists of two subtests (semantics and morphosyntax) in each language (Spanish and English) that were drawn from larger item sets of the 2008 experimental version of the BESA (Peña et al., 2014), and items were selected on the basis of item analyses and how well they discriminated children with and without impairment between the ages of 4;6 and 5;6. The BESOS semantics subtest measures vocabulary knowledge through both receptive and expressive items (e.g., "Show me the dog that is different"; "Tell me all the foods you eat for lunch"). Each subtest is administered in the target language, and following a conceptual scoring system (Bedore, Peña, García, & Cortez, 2005; Pearson, Fernandez, & Oller, 1993), responses in either language on the semantics subtests are accepted as correct. The BESOS morphosyntax subtest includes cloze and sentence repetition items targeting grammatical forms that are typically more difficult for children with LI. The subtests in each language are not direct translations of one another. The semantics subtests contain 10–12 items, and the morphosyntax subtests contain 16–17 items, depending on the test language and the child's age. There are two separate versions for each age group (4- and 5-year-olds). Correlations between the full experimental BESA semantics subtests and the BESOS semantics subtests have been reported as .855 for Spanish and .887 for English (Summers, Bohman, Gillam, Peña, & Bedore, 2010). Alpha coefficients for the screening subtests ranged from .883 to .894 for English morphosyntax, from .880 to .899 for Spanish morphosyntax, from .579 to .725 for English semantics, and from .664 to .705 for Spanish semantics. These indicate acceptable to good internal consistency. Ranges represent coefficients for the different age versions of the BESOS.

Diagnostic Reference Standard

Children were determined to have LI on the basis of the expertise of three bilingual SLPs with more than 10 years of experience diagnosing and treating LI in bilingual children (Gillam et al., 2013). These bilingual SLPs reviewed children's parent and teacher questionnaires on the history of language use, exposure and concerns about language development, transcripts of narrative samples, and transcribed responses to items from standardized tests in English and Spanish—including the Test of Language Development–Primary: Third Edition (Newcomer & Hammill, 1997), the Test of Narrative Language (Gillam & Pearson, 2004), and the experimental version of the BESA (Peña et al., 2014). On the basis of the framework established by Records and Tomblin (1994) and Tomblin et al. (1996), they independently rated each child's performance in vocabulary, grammar, and narration in both English and Spanish using a 6-point scale (ranging from 0 = *severe/profound impairment* to 5 = *typical performance*). After assigning scores by each domain and in each language, they were asked to make an overall judgment of language ability using the same 6-point scale. Participants were identified as having LI if they were assigned a rating of 2 (*mild impairment*) or below by at least two of the SLP raters. The overall point-to-point agreement among the raters was 90%. This diagnostic referent was unique in that it was based on a rating system of multiple sources of information that included direct testing and narrative samples in both languages, parent or teacher concerns, and SLP clinical judgment (Dollaghan & Horner, 2011). For more detail about the procedure for diagnosing impairment in this sample, see Gillam et al. (2013).

Table 2. Descriptive information by matched language ability groups.

	TD-matched (n = 21)		LI (n = 21)	
Subtest	*M*	*SD*	*M*	*SD*
Average English input and output (%)	48	12	45	13
English morphosyntax	78.90	17.10	66.33	8.03
Spanish morphosyntax	88.12	20.55	68.74	9.28
English semantics	81.70	16.07	66.80	20.17
Spanish semantics	88.24	13.43	71.32	17.01
Best morphosyntax[a]	95.10	18.36	71.46	9.07
Best semantics[a]	91.53	12.34	79.25	14.90

Note. Average English input and output scores were computed on the basis of parents' responses to the Bilingual Input Output Survey questionnaire; numbers represent each group's mean percentage of combined English input and output in a typical week. Bilingual English Spanish Oral Screener scores are standard scores. TD = typically developing; LI = language impairment.

[a]Best morphosyntax and semantics scores represent participants' higher score in each domain.

Procedure

Procedures for the screener administration were the same as those followed in Peña et al. (2011). Screening occurred prior to the beginning of kindergarten. All children were individually administered sections of the BESOS in a quiet area at their schools. Administration of all four subtests of the BESOS took approximately 20 min, and testing on a given subtest was discontinued when a child completed all items or if a child did not respond to five items in a row. Each language was tested separately within the 20-min session. Administration of the subtests was completed in random order, varying both the first language and first domain of testing.

Examiners included bilingual SLPs, trained bilingual graduate and undergraduate students in communication sciences and disorders, and research associates with undergraduate degrees in related fields. Because classification of LI was not made until all children had completed the first grade, testers were blind to impairment or risk status. Examiners recorded individual item responses and followed specific scoring guidelines to judge correct/incorrect responses. A TeleForm system was utilized to scan individual scores into a computerized database for analysis. Standard scores were subsequently computed on the basis of a larger norming sample.

Analysis

Discriminant analysis offers a unique way to both empirically derive cutoffs and identify a linear combination of scores that would result in the best classification accuracy for multiple measures (Bedore & Leonard, 1998; Skarakis-Doyle, Dempsey, & Lee, 2008). Using SPSS, we conducted a series of discriminant analyses to determine the linear combination of BESOS scores (semantics or morphosyntax in English and/or Spanish) that best predicted group membership (LI vs. TD-matched) and maximized the sensitivity and specificity for the matched groups (21 TD and 21 LI). This also allowed us to empirically derive cut-point scores for the combination of measures that yielded the highest accuracy.

First, to determine how well each individual BESOS subtest predicted risk for language ability for the matched groups, we calculated the predictive classification accuracy for each subtest. Then, an exploratory discriminant analysis was conducted to determine how well the linear combination of all four subtest scores—English and Spanish semantics and English and Spanish morphosyntax—predicted risk for language ability (LI vs. TD).

In an effort to improve both predictive sensitivity and specificity, we then selected participants' best language score in semantics and morphosyntax following the approach of Peña and Bedore (2011) and that of the BESA (Peña et al., 2014). We conducted additional exploratory discriminant functions of the best score for each individual domain (semantics and morphosyntax) and their combination.

Bossuyt et al. (2003) recommended interpreting sensitivity and specificity values in light of likelihood ratios (LRs). These measure the likelihood of a positive or negative screening result or, in other words, the odds that a child at a given cutoff will be correctly identified. For diagnostic tests, a positive likelihood ratio (LR+) ≥ 10 and a negative likelihood ratio (LR−) ≤ 0.10 are highly informative, from 5 to 10 and from 0.1 to 0.2 are moderately informative, between 2 and 5 and between 0.2 and 0.5 are modestly informative, and < 2 or > 0.5 are uninformative (Dollaghan, 2007; Dollaghan & Horner, 2011; Hanley & McNeil, 1982; Jaeschke, Guyatt, & Sackett, 1994; Sackett, Haynes, Guyatt, & Tugwell, 1991).

Results

Table 2 lists the means and standard deviations for participants in each group for each subtest of the BESOS. All participants had data for all four subtests, and no outliers were identified. Scores for the LI group were significantly lower than the TD-matched group on all four subtests ($p < .05$).

Predictive Classification Accuracy

Table 3 lists the overall predictive classification accuracy, sensitivity, specificity, LRs, and relevant statistics of

Table 3. Discriminant functions.

Discriminant function	Cutoff score	Overall classification	Sensitivity	Specificity	LR+ [95% CI]	LR− [95% CI]	Wilks's λ	χ^2	Canonical correlation	*p*
English morphosyntax[a]	72.60	71.4	95.2	47.6	1.82 [1.20, 2.76]	0.10 [0.01, 0.71]	.811	8.25	.434	.004
Spanish morphosyntax[a]	78.43	71.4	90.5	52.4	1.90 [1.19, 3.04]	0.18 [0.05, 0.72]	.720	12.96	.529	<.001
English semantics	74.25	69.0	66.7	71.4	2.33 [1.11, 4.89]	0.47 [0.24, 0.91]	.851	6.38	.386	.012
Spanish semantics	79.78	73.8	66.7	81.0	3.50 [1.38, 8.89]	0.41 [0.22, 0.78]	.758	10.97	.492	.001
All four[a]		78.6	85.7	71.4	3.00 [1.49, 6.03]	0.20 [0.07, 0.59]	.522	24.73	.692	<.001
Best morphosyntax[a]	83.28	76.2	85.7	66.7	2.57 [1.37, 4.83]	0.21 [0.07, 0.64]	.588	20.95	.642	<.001
Best semantics[a]	85.39	71.4	66.7	76.2	2.80 [1.23, 6.37]	0.44 [0.23, 0.84]	.825	7.58	.418	.006
Best morphosyntax + best semantics[a]		83.3	95.2	71.4	3.33 [1.68, 6.60]	0.07 [0.01, 0.46]	.561	22.56	.663	<.001

Note. LR = likelihood ratio; CI = confidence interval.

[a] Box's M assumption of homogeneity of covariance matrices was not met; separate covariance matrices were used whenever it improved classification by more than 2% (Burns & Burns, 2008).

the discriminant function for each individual test and combination of tests.

Individual Subtests

All four discriminant functions for each individual subtest were statistically significant ($p < .05$), meaning they uniquely explained some of the between-groups variability. The overall predictive classification was higher for Spanish subtests, but most values were less than 80%. Both morphosyntax subtests (Spanish and English) reached good predictive sensitivity (more than 90%); however, predictive specificity was near chance. Specificity was highest for Spanish semantics (81%), but sensitivity was poor (66.7%). No one subtest, however, had both sensitivity and specificity values that were more than 70%. For each individual subtest, LR+ was between 1.82 and 3.5, and LR− was between 0.10 and 0.47. These LRs are considered to be between modestly and moderately informative.

Combination of Scores

The exploratory discriminant function for the combination of all four subtests revealed a significant association between groups, explaining 47.89% of between-groups variability. The predictive classification showed that, overall, 78.6% were predicted correctly. Predictive sensitivity reached 85.7%, and predictive specificity was 71.4%. There were three LI cases and six TD cases in which risk for LI was predicted incorrectly. LRs showed slight improvement from those of each individual subtest.

Best Scores

The best morphosyntax score resulted in similar predictive sensitivity to the combination of all four subtests but also resulted in poorer predictive specificity. The best semantics score had better specificity to the combination of all four tests but also had poorer sensitivity. However, the linear combination of the best semantics and best morphosyntax scores together resulted in predictive sensitivity of 95.2%, resulted in specificity of 71.4%, and explained 44% of the between-groups variance, which is an improvement on the classification accuracy of the combination of all four tests. This discriminant function correctly classified all of the children with LI except for one. LR+ was higher and LR− was lower than the combination of all four subtests (see Table 3). The positive likelihood is modestly informative, and the negative likelihood range is highly informative. The resulting formula for the discriminant function obtained from combining the best semantics and morphosyntax scores for the matched groups was $D = 0.024(\text{semantics}) + 0.060(\text{morphosyntax}) - 7.117$. The mean discriminant score for the TD group was 0.864, and the LI group was −0.864.

Discussion

The primary purpose of this study was to determine the combination of scores on the BESOS that maximized its ability to accurately predict risk for LI in first grade when bilingual children were tested prior to entering kindergarten. We conducted exploratory discriminant analyses with a set of matched bilingual preschoolers with and without LI using all four subtest scores on the BESOS and explored different combinations of scores to identify the one with the highest accuracy. We found that the combination of the highest score in each domain (semantics and morphosyntax) resulted in the highest predictive accuracy. This yielded an empirically derived cutoff on the basis of the weighted linear combination of both subtest scores.

Predicting Risk for LI

Although the individual subtests of the BESOS had modestly and moderately informative predictive properties, when we selected each participant's best (Spanish vs. English) score on each domain (semantics and morphosyntax), the means and distribution of these scores appeared to show more separation between language ability groups. A combination of children's best score from each domain yielded the best overall prediction of risk. This discriminant analysis explained a significant proportion of the variance between groups (44%) and resulted in high predictive sensitivity (95.2%). The formula provided earlier can be used to derive a discriminant score, *D*. Each of the child's best scores in semantics and morphosyntax are multiplied by its corresponding function coefficient summed to a constant. The resulting *D* score can be compared with the mean discriminant scores for each group (TD = 0.864; LI = −0.864) to determine which group the child may belong in or to classify each child.

This combination of best scores was clinically significant because it correctly classified all but one of the 21 children with LI. This particular child with LI had an interesting profile, as he was a Spanish-dominant bilingual with a high Spanish semantics score (0.5 *SD*s above the mean) and a low Spanish morphosyntax score (1.5 *SD*s below the mean). The weighted combination of this child's scores on the discriminant function (*D* = 0.189) was very close to the midpoint between the means for both TD and LI groups, and the child ended up being misclassified as TD.

Sensitivity and specificity values can be influenced by the prevalence of impairment of the sample in which a test's classification accuracy is being tested. At first glance, our sample appears to have a higher rate of LI (12%) compared with that reported in the monolingual literature (7%; Tomblin et al., 1997). However, recall the inclusion criteria—our sample was derived from a larger group of 1,029 children, and only those who scored below the 30th percentile on two of the four subtests were invited to participate in a longitudinal study of diagnostic markers of LI. We presume that if we had followed those other 1,000 participants, we would have found a similar prevalence to that reported in the monolingual literature. Nonetheless, we report likelihood values in Table 3 to more carefully interpret our sensitivity and specificity values (Bossuyt et al., 2003). Our highest LR+ of 3.33 with the combination of best semantics and morphosyntax scores is considered modestly informative. For the same analysis, we obtained an LR− of 0.07, which is considered highly informative and demonstrates that we can confidently rule out impairment, which is an important function of a screener.

The obtained specificity (71.4%) is below Plante and Vance's (1994) recommendation for diagnostic tests. However, acceptable levels of specificity for screeners should be less stringent than those for diagnostic tests (Barnes, 1982; Bright Futures Steering Committee and Medical Home Initiatives for Children with Special Needs Project Advisory Committee, 2006). In a typical educational screening context, the BESOS would likely not be used in isolation (Dollaghan & Horner, 2011; Warner, 2004). Prior to a clinical diagnosis, logical follow-ups would include teacher or parent interviews or a formal referral for additional testing. In some instances, screening results may help identify children who could benefit from watchful maintenance or progress-monitoring programs, which may provide a long-term cost benefit against the potential loss of educational opportunity (Black, 2010; Dollaghan, 2007; Glascoe, 2001). We do not know whether any of our participants received language intervention between the time the BESOS screener was administered and the clinical diagnosis at first grade. We suppose that if some children did receive services, some of their difficulties may have resolved by first grade, which strengthens the predictive validity of the BESOS because we were still able to capture 95% of the participants who truly had impairments prior to entering kindergarten.

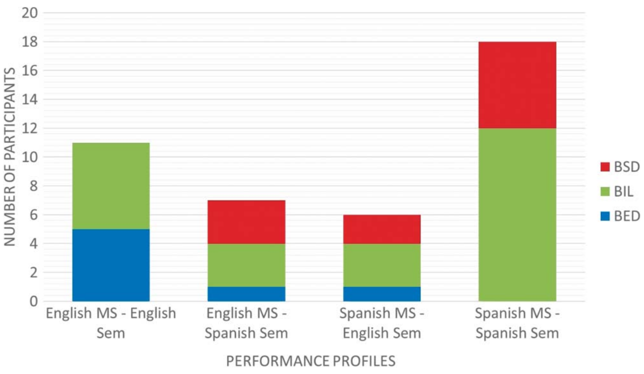
If directly comparing our current results with the only commercially available Spanish screener, the sensitivity of the BESOS (71.4%) appears to be below that reported for the PLSSST-5 (79%). These specificity values are not directly comparable because the present study assessed predictive accuracy, and the PLSSST-5 manual reports concurrent accuracy. The preliminary concurrent sensitivity for the BESOS is 91% (Peña et al., 2008), which is above that reported for the PLSSST-5. Also, the diagnostic referents used to measure the accuracy of each test were different. The PLSSST-5 reports solely using the parent Preschool Language Scales–Fifth Edition Spanish test, and for the current study, we used a rating system of multiple sources of information that included direct testing and narrative samples in both languages, parent or teacher concerns, and SLP clinical judgment. An additional reason why scores on these two screeners are not directly comparable is that the PLSSST-5 was normed on Spanish speakers or Spanish-dominant bilinguals, and the BESOS norming group includes children across the bilingual continuum. To make direct comparisons across these two measures, the instruments would have to be tested across the same sample of bilingual children, using an identical reference standard and deriving scores from a similar norming sample.

Why Is a Cross-Language Combination of Scores Useful?

One of the challenges clinicians face in bilingual language screening is the lack of guidance on how to clinically interpret scores from tests in multiple languages or domains. Our results showed that the highest predictive classification accuracy was obtained from the linear combination of participants' scores across languages and domains. An interesting finding was that, for part of our sample, this meant a cross-linguistic combination of best scores. Four best language score profiles were possible: (1) best score on English morphosyntax and English semantics, (2) best score on English morphosyntax and Spanish semantics, (3) best score on Spanish morphosyntax and English semantics, and (4) best score on Spanish morphosyntax and Spanish semantics. About 30% of our sample had Profiles 2 and 3 above. Recall that children's scores on each domain and language may be influenced by their cumulative knowledge, input, and current use of that language (Bedore et al., 2012). The range of English input/output of our sample was 22%–74%, indicating a wide range of linguistic backgrounds. One way to conceptualize the effect of experience on these mixed dominance profiles is to classify students by language experience groups on the basis of parental report of exposure and use of each language. A child was considered bilingual English dominant if he or she heard and used English between 80% and 60% of the time; a child was considered bilingual Spanish dominant if exposure and use of Spanish was between 80% and 60%; and a child was considered balanced bilingual if he or she heard and used English and Spanish between 40% and 60% of the time (Bohman et al., 2010). Figure 1 illustrates the number of participants in Profiles 2 and 3 by language experience groups. When we divide out the participants into these groups, we can see that participants in all three categories (balanced bilingual, bilingual English dominant, bilingual Spanish dominant) benefited from using a "best score" approach. By selecting their best score in each domain, we maximized the likelihood of accurately capturing their ability and attempted to control for some of these differential experiences.

Another reason why a cross-language combination of best scores is useful relates to the design of the BESOS as well as factors related to English language development of ELLs. The BESOS was created using items from the experimental version of the BESA that best identified impairment. However, some of these same items are also developmentally challenging for ELLs, particularly the English morphosyntax items that coincide with linguistic features that are learned with more experience in the language (Gutiérrez-Clellen et al., 2008; Paradis, 2005). Errors in children's English responses may be due to both the process of acquiring a second language, impairment, or both. By using their better language score in each domain, we maximize the likelihood that children's most representative performance relative to their ability is captured during a brief screening.

Figure 1. Best language score performance profiles on an experimental version of the Bilingual English Spanish Oral Screener. This figure represents four performance profiles that are based on the number of participants with higher scores on each subtest: (1) best score on English morphosyntax (MS) and English semantics (Sem), (2) best score on English MS and Spanish Sem, (3) best score on Spanish MS and English Sem, and (4) best scores on Spanish MS and Spanish Sem. BSD = bilingual Spanish dominant; BIL = balanced bilingual; BED = bilingual English dominant.



Selecting participants' best language score in each domain is different than testing in only one language or making an arbitrary decision about dominance prior to testing. Testing in two domains across both languages reveals strengths and weaknesses in specific areas that testing in only one language would not reveal. As expected, children who scored high in both semantics and morphosyntax (regardless of language) were correctly classified as TD. Some might have projected that children with LI would score poorly in both domains or in both languages. However, participants in both the TD-matched and LI groups were equally represented in the mixed dominance profiles mentioned above. Some scored poorly in one language and domain, and some scored within 1 *SD* (but still low average) on one or both subtests in each language. To identify risk for impairment, selecting their higher score on each subtest ensured that we represented their highest ability in each domain for the clinical interpretation of the BESOS screener scores.

It may appear that screening in English would be faster and more cost-effective, particularly because of the lack of bilingual personnel in many school districts across the nation. However, the cost–benefit ratio of obtaining greater accuracy by combining children's best scores outweighs the administrative and personnel costs of an additional 10 min of screening in Spanish. The combination of the best language scores also yielded the highest predictive classification accuracy, thus reducing the potential loss of educational opportunity related to underidentification.

Clinical Implications and Future Directions

The BESOS seems to be an efficient way to predict risk for LI in bilingual preschoolers because it is quick to administer, samples both languages, is psychometrically sound, and has good predictive accuracy. The BESOS takes 20 min to administer, which is no longer than it takes to administer some single-language screeners. We were able to derive an empirically based cutoff score that combined a child's best scores in each domain across languages, thereby maximizing each child's opportunity to demonstrate his or her linguistic abilities on a brief test. Our team of examiners included trained individuals with bachelor's degrees from related fields who were bilingual, and they were able to both administer and score the BESOS. Because there is a documented shortage of bilingual personnel in schools (American Speech-Language-Hearing Association, 2012), future studies could explore the viability of training assistants or paraprofessionals, both bilingual and nonnative speakers, to reliably administer and interpret the BESOS.

Future studies should also continue to validate the use of the BESOS for independent samples of bilingual children, both with and without impairment, by comparing its validity with that of other commercially available screeners and by cross-validating the empirically derived cutpoint scores from this study (Law et al., 1998; Sturner et al., 1994). Further item analyses could also be informative in refining the test and could increase its efficiency by reducing its administration time. In addition, future studies could focus on empirically combining screener scores with additional information, such as parent or educator reports of concern or dynamic assessment, before resulting in referral for speech and language evaluation (see Dockrell & Marshall, 2015; Dollaghan & Horner, 2011), which could potentially improve specificity.

Acknowledgments

This project was supported by National Institute on Deafness and Other Communication Disorders Grant R01DC007439 and Department of Education Training Grant H325D090068 (both grants awarded to Elizabeth D. Peña). We thank all of the interviewers and examiners for their assistance with collecting the data and the school districts for allowing us access to the participants.

References

Allen, D. V., & Bliss, L. S. (1987). Concurrent validity of two language screening tests. *Journal of Communication Disorders, 20,* 305–317.

American Speech-Language-Hearing Association. (2004). *Preferred practice patterns for the profession of speech-language pathology* [Preferred practice patterns]. Available from http://www.asha.org/policy

American Speech-Language-Hearing Association. (2012). *2012 Schools Survey report: SLP caseload characteristics*. Available from http://www.asha.org/research/memberdata/schoolssurvey/

Aud, S., Wilkinson-Flicker, S., Kristapovich, P., Rathbun, A., Wang, X., & Zhang, J. (2013). *The condition of education 2013* (Report No. NCES 2013-037). Retrieved from http://nces.ed.gov/pubs2013/2013037.pdf

Barnes, K. E. (1982). *Preschool screening: The measurement and prediction of children at-risk*. Springfield, IL: Charles C Thomas.

Batalova, J., & McHugh, M. (2010). *Top languages spoken by English language learners nationally and by state*. Washington, DC: Migration Policy Institute.

Bedore, L. M., & Leonard, L. B. (1998). Specific language impairment and grammatical morphology: A discriminant function analysis. *Journal of Speech, Language, and Hearing Research, 41,* 1185–1192.

Bedore, L. M., & Leonard, L. B. (2001). Grammatical morphology deficits in Spanish-speaking children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 44,* 905–924.

Bedore, L. M., & Leonard, L. B. (2005). Verb inflections and noun phrase morphology in the spontaneous speech of Spanish-speaking children with specific language impairment. *Applied Psycholinguistics, 26,* 195–225.

Bedore, L. M., & Peña, E. D. (2008). Assessment of bilingual children for identification of language impairment: Current findings and implications for practice. *International Journal of Bilingual Education and Bilinguals, 11,* 1–29. doi:10.2167/beb392.0

Bedore, L. M., Peña, E. D., García, M., & Cortez, C. (2005). Conceptual versus monolingual scoring: When does it make a difference? *Language, Speech, and Hearing Services in Schools, 36,* 188–200. doi:10.1044/0161-1461(2005/020)

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. doi:10.1017/S1366728912000090

Black, R. S. (2010). Can underidentification affect exceptional learners? In F. E. Obiakor, J. P. Bakken, & A. F. Rotatori (Eds.), *Advances in special education: Vol. 19. Current issues and trends in special education: Identification, assessment, and instruction* (pp. 37–51). Bingley, England: Emerald Group Publishing Limited.

Blaxley, L., Clinker, M., & Warr-Leeper, G. (1983). Two language screening tests compared with developmental sentence scoring. *Language, Speech, and Hearing Services in Schools, 14,* 38–46.

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 early sequential Spanish–English bilinguals. *International Journal of Bilingual Education and Bilingualism, 13,* 325–344. doi:10.1080/13670050903342019

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., . . . Lijmer, J. G. (2003). The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. *Clinical Chemistry, 49,* 7–18.

Bracken, B., & McCallum, S. (1998). *Universal Nonverbal Intelligence Test*. Itasca, IL: Riverside.

Bright Futures Steering Committee and Medical Home Initiatives for Children with Special Needs Project Advisory Committee. (2006). Identifying infants and young children with developmental disorders in the medical home: An algorithm for developmental surveillance and screening. *Pediatrics, 118,* 405–420.

Burns, R. P., & Burns, R. (2008). *Business research methods and statistics using SPSS*. Thousand Oaks, CA: Sage.

Core, C., Hoff, E., Rumiche, R., & Señor, M. (2013). Total and conceptual vocabulary in Spanish–English bilinguals from 22 to 30 months: Implications for assessment. *Journal of Speech, Language, and Hearing Research, 56,* 1637–1649. doi:10.1044/1092-4388(2013/11-0044)

Dockrell, J. E., & Marshall, C. R. (2015). Measurement issues: Assessing language skills in young children. *Child and Adolescent Mental Health, 20,* 116–125. doi:10.1111/camh.12072

Dollaghan, C. A. (2007). *The handbook of evidence-based practice in communication disorders*. Baltimore, MD: Brookes.

Dollaghan, C. A., & Horner, E. A. (2011). Bilingual language assessment: A meta-analysis of diagnostic accuracy. *Journal of Speech, Language, and Hearing Research, 54,* 1077–1088. doi:10.1044/1092-4388(2010/10-0093)

Fluharty, N. B. (2001). *Fluharty 2: Fluharty Preschool Speech and Language Screening Test*. Austin, TX: Pro-Ed.

Friberg, J. C. (2010). Considerations for test selection: How do validity and reliability impact diagnostic decisions? *Child Language Teaching and Therapy, 26*(1), 77–92. doi:10.1177/0265659009349972

Gillam, R. B., & Pearson, N. (2004). *Test of Narrative Language*. Austin, TX: Pro-Ed.

Gillam, R. B., Peña, E. D., Bedore, L. M., Bohman, T. M., & Mendez-Perez, A. (2013). Identification of specific language impairment in bilingual children, Part 1: Assessment in English. *Journal of Speech, Language, and Hearing Research, 56,* 1813–1823. doi:10.1044/1092-4388(2013/12-0056)

Glascoe, F. P. (2001). Are over referrals on developmental screening tests really a problem? *Archives of Pediatric and Adolescent Medicine, 155,* 54–59. doi:10.1001/archpedi.155.1.54

Greene, K. J. (2012). *Cognitive based intervention for bilingual preschoolers* (Unpublished doctoral dissertation). The University of Texas at Austin.

Grimes, D. A., & Schulz, K. F. (2002). Uses and abuses of screening tests. *The Lancet, 359,* 881–884. doi:10.1016/S0140-6736(02)07948-5

Gutiérrez-Clellen, V. F., Restrepo, M. A., & Simón-Cereijido, G. (2006). Evaluating the discriminant accuracy of a grammatical measure with Spanish-speaking children. *Journal of Speech, Language, and Hearing Research, 49,* 1209–1223.

Gutiérrez-Clellen, V. F., & Simón-Cereijido, G. (2007). The discriminant accuracy of a grammatical measure with Latino English-speaking children. *Journal of Speech, Language, and Hearing Research, 50,* 968–981.

Gutiérrez-Clellen, V. F., Simón-Cereijido, G., & Wagner, C. (2008). Bilingual children with language impairment: A comparison with monolinguals and second language learners. *Applied Psycholinguistics, 29*(1), 3–19.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29–36.

Illerbrun, D., Haines, L., & Greenough, P. (1985). Language identification screening test for kindergarten: A comparison with four screening and three diagnostic language tests. *Language, Speech, and Hearing Services in Schools, 16,* 280–291.

Jacobson, P. F., & Schwartz, R. G. (2005). English past tense use in bilingual children with language impairment. *American Journal of Speech-Language Pathology, 14,* 313–323.

Jaeschke, R., Guyatt, G. H., & Sackett, D. L. (1994). Users' guides to the medical literature. *Journal of the American Medical Association, 271,* 703–707.

Kohnert, K. (2010). Bilingual children with primary language impairment: Issues, evidence and implications for clinical actions. *Journal of Communication Disorders, 43,* 465–473. doi:10.1016/j.jcomdis.2010.02.002

Law, J., Boyle, J., Harris, F., Harkness, A., & Nye, C. (1998). Screening for speech and language delay: A systematic review of the literature. *Health Technology Assessment, 2,* 37–48.

Law, J., Rush, R., Anandan, C., Cox, M., & Wood, R. (2012). Predicting language change between 3 and 5 years and its implications for early identification. *Pediatrics, 130,* e132–e137.

McCauley, R. J., & Swisher, L. (1984). Psychometric review of language and articulation tests for preschool children. *Journal of Speech and Hearing Disorders, 49,* 34–42.

Merrell, A. W., & Plante, E. (1997). Norm-referenced test interpretation in the diagnostic process. *Language, Speech, and Hearing Services in Schools, 28,* 50–58.

Nelson, H. D., Nygren, P., Walker, M., & Panoscha, R. (2006). Screening for speech and language delay in preschool children: Systematic evidence review for the US Preventive Services Task Force. *Pediatrics, 117,* 298–319. doi:10.1542/peds.2005-1467

Newcomer, P. L., & Hammill, D. D. (1997). *Test of Language Development–Primary* (3rd ed.). Austin, TX: Pro-Ed.

Oetting, J. B., Cleveland, L. H., & Cope, R. F., III. (2008). Empirically derived combinations of tools and clinical cutoffs: An illustrative case with a sample of culturally/linguistically diverse children. *Language, Speech, and Hearing Services in Schools, 39,* 44–53. doi:10.1044/0161-1461(2008/005)

Paradis, J. (2005). Grammatical morphology in children learning English as a second language: Implications of similarities with specific language impairment. *Language, Speech, and Hearing Services in Schools, 36,* 172–187.

Pearson, B. Z., Fernandez, S. C., & Oller, D. K. (1993). Lexical development in bilingual infants and toddlers: Comparison to monolingual norms. *Language Learning, 43*(1), 93–120.

Peña, E. D., & Bedore, L. M. (2011, November 1). It takes two: Improving assessment accuracy in bilingual children. *The ASHA Leader*. Retrieved from http://www.asha.org/Publications/leader/2011/111101/It-Takes-Two–Improving-Assessment-Accuracy-in-Bilingual-Children/

Peña, E. D., Bedore, L. M., Iglesias, A., Gutiérrez-Clellen, V. F., & Goldstein, B. A. (2008). *Bilingual English Spanish Oral Screener–Experimental Version (BESOS)*. Unpublished instrument.

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

Peña, E. D., Gillam, R. B., Bedore, L. M., & Bohman, T. M. (2011). Risk for poor performance on a language screening measure for bilingual preschoolers and kindergarteners. *American Journal of Speech-Language Pathology, 20,* 302–314. doi:10.1044/1058-0360(2011/10-0020)

Peña, E. D., Gutiérrez-Clellen, V. F., Iglesias, A., Goldstein, B. A., & Bedore, L. M. (2014). *Bilingual English Spanish Assessment (BESA)*. San Rafael, CA: AR Clinical Publications.

Perona, K., Plante, E., & Vance, R. (2005). Diagnostic accuracy of the Structured Photographic Expressive Language Test: Third Edition (SPELT-3). *Language, Speech, and Hearing Services in Schools, 36,* 103–115.

Plante, E., & Vance, R. (1994). Selection of preschool language tests: A data-based approach. *Language, Speech, and Hearing Services in Schools, 25,* 15–24.

Records, N. L., & Tomblin, J. B. (1994). Clinical decision making: Describing the decision rules of practicing speech-language pathologists. *Journal of Speech and Hearing Research, 37,* 144–156.

Restrepo, M. A. (1998). Identifiers of predominantly Spanish-speaking children with language impairment. *Journal of Speech, Language, and Hearing Research, 41,* 1398–1411.

Sackett, D. L., Haynes, R. B., Guyatt, G. H., & Tugwell, P. (1991). The interpretation of diagnostic data. In D. L. Sackett, R. B. Haynes, G. H. Guyatt, & P. Tugwell (Eds.), *Clinical epidemiology: A basic science for clinical medicine* (2nd ed., pp. 69–152). Boston, MA: Lippincott, Williams & Wilkins.

Samson, J. F., & Lesaux, N. K. (2009). Language-minority learners in special education: Rates and predictors of identification for services. *Journal of Learning Disabilities, 42,* 148–162. doi:10.1177/0022219408326221

Shipley, K. G., Stone, T. A., & Sue, M. B. (1983). *Test for Examining Expressive Morphology (TEEM)*. Tucson, AZ: Communication Skill Builders.

Skarakis-Doyle, E., Dempsey, L., & Lee, C. (2008). Identifying language comprehension impairment in preschool children. *Language, Speech, and Hearing Services in Schools, 39,* 54–65. doi:10.1044/0161-1461(2008/006)

Spaulding, T. J., Plante, E., & Farinella, K. A. (2006). Eligibility criteria for language impairment: Is the low end of normal always appropriate? *Language, Speech, and Hearing Services in Schools, 37,* 61–72. doi:10.1044/0161-1461(2006/007)

Squires, K., Lugo-Neris, M. J., Peña, E. D., Bedore, L. M., & Gillam, R. (2014). Story retelling by bilingual children with language impairments and typically developing controls. *Journal of Language and Communication Disorders, 49,* 60–74. doi:10.1111/1460-6984.12044

Sturner, R. A., Funk, S. G., & Green, J. A. (1996). Preschool speech and language screening: Further validation of the Sentence

Repetition Screening Test. *Journal of Developmental Behavior Pediatrics, 17,* 405–413.

Sturner, R. A., Layton, T. L., Evans, A. W., Funk, S. G., & Machon, M. W. (1994). Preschool speech and language screening: A review of currently available tests. *American Journal of Speech-Language Pathology, 3,* 25–36. doi:10.1044/1058-0360.0301.25

Summers, C., Bohman, T. M., Gillam, R. B., Peña, E. D., & Bedore, L. M. (2010). Bilingual performance on nonword repetition in Spanish and English. *International Journal of Language & Communication Disorders, 45,* 480–493.

Thordardottir, E., Rothenberg, A., Rivard, M., & Naves, R. (2006). Bilingual assessment: Can overall proficiency be estimated from separate measurement of two languages? *Journal of Multilingual Communication Disorders, 4,* 1–21. doi:10.1080/14769670500215647

Tomblin, J., Records, N., Buckwalter, P., Zhang, X., Smith, E., & O'Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research, 40,* 1245–1260.

Tomblin, J. B., Records, N. L., & Zhang, X. (1996). A system for the diagnosis of specific language impairment in kindergarten children. *Journal of Speech and Hearing Research, 39,* 1284–1294. doi:10.1044/jshr.3906.1284

Valdés, G., & Figueroa, R. A. (2004). *Bilingualism and testing: A special case of bias*. Norwood, NJ: Ablex.

Warner, J. (2004). Clinicians' guide to evaluating diagnostic and screening tests in psychiatry. *Advances in Psychiatric Treatment, 10,* 446–454. doi:10.1192/apt.10.6.446

Washington, J. A., & Craig, H. K. (2004). A language screening protocol for use with young African American children in urban settings. *American Journal of Speech-Language Pathology, 13,* 329–340.

Wiig, E. H., Secord, W., & Semel, E. M. (2004). *CELF Preschool 2: Clinical Evaluation of Language Fundamentals Preschool*. San Antonio, TX: Pearson/PsychCorp.

Wilson, J. M. G., & Jungner, G. (1968). *Principles and practice of screening for disease* (Public Health Papers, No. 34). Retrieved from http://whqlibdoc.who.int/php/who_php_34.pdf

Wilson, P., McQuaige, F., Thompson, L., & McConnachie, A. (2013). Language delay is not predictable from available risk factors. *The Scientific World Journal, 2013,* 947018. doi:10.1155/2013/947018

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2012a). *Preschool Language Scales–Fifth Edition Spanish*. Bloomington, MN: NCS Pearson.

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2012b). *Preschool Language Scales Screening Test–Fifth Edition*. Bloomington, MN: NCS Pearson.

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2012c). *Preschool Language Scales Spanish Screening Test–Fifth Edition*. Bloomington, MN: NCS Pearson.