AJSLP

Research Article

The Utility of an English Semantics Measure for Identifying Developmental Language Disorder in Spanish–English Bilinguals

Javier Jasso,[a,b] Stephanie McMillen,[c] Jissel B. Anaya,[a] Lisa M. Bedore,[b] and Elizabeth D. Peña[d]

Purpose: We examined the English semantic performance of three hundred twenty-seven 7- to 10-year-old Spanish–English bilinguals with (n = 66) and without (n = 261) developmental language disorder (DLD) with varying levels of English experience to classify groups.
Method: English semantic performance on the Bilingual English–Spanish Assessment—Middle Extension Experimental Test Version (Peña et al., 2008) was evaluated by language experience, language ability, and task type. Items that best identified DLD for children with balanced and high English experience were selected. Separately, items that best identified children with high Spanish experience were selected.
Results: Typically developing bilingual children performed significantly higher than their peers with DLD across semantic tasks, with differences associated with task type. Classification accuracy was fair when item selection corresponded to balanced or high level of experience in English, but poor for children with high Spanish experience. Selecting items specifically for children with high Spanish experience improved classification accuracy.
Conclusions: Tailoring semantic items based on children's experience is a promising direction toward organizing items on a continuum of exposure. Here, classification effectively ruled in impairment. Future work to refine semantic items that more accurately represent the continuum of exposure may help rule out language impairment.

Language experience, culture, and sociolinguistic attitudes shape bilingual children's performance. Efforts to minimize language assessment bias in bilingual/bidialectal speakers in light of these factors have been promising. While assessing both languages is considered best practice (Peña et al., 2017), few speech-language pathologists (SLPs) have the language proficiency and clinical preparation to assess the child's first language (American Speech-Language-Hearing Association, 2016). Furthermore, bilingual acquisition data and clinical markers of impairment are unknown for many minority languages. A possible solution is to explore whether English testing can inform clinical decisions about the language ability (or disorder status) of children in the process of learning English as a second language (L2). This line of inquiry may aid SLPs who do not speak the child's first language in determining whether low performance on English language measures is consistent with the variability associated with L2 acquisition or with a profile of developmental language disorder (DLD).

The term "DLD" is the current consensus term for children with language learning impairments, referring both to children classified as having a specific language impairment (SLI), a narrow subset of children with exceptional language difficulties whose diagnosis is not attributable to any frank biomedical condition and who display average nonverbal intelligence (Leonard, 2014; Tomblin et al., 1997), as well as children with language disorders based on a broader designation that includes comorbid conditions and low nonverbal intelligence that does not meet the criteria for intellectual disability (Bishop et al., 2017). For this study, we adopt the term "DLD" while recognizing that

[a]Department of Communication Sciences and Disorders, The University of Texas at Austin
[b]Department of Communication Sciences and Disorders, Temple University, Philadelphia, PA
[c]Department of Communication Sciences and Disorders, Syracuse University, NY
[d]School of Education, University of California, Irvine
Correspondence to Javier Jasso: javierjasso@utexas.edu

Editor-in-Chief: Julie Barkmeier-Kraemer
Editor: Anny Castilla-Earls
Received December 19, 2018
Revision received April 30, 2019
Accepted January 6, 2020
https://doi.org/10.1044/2020_AJSLP-19-00202

Disclosure: *Elizabeth D. Peña and Lisa M. Bedore are co-authors of the BESA and receive royalties from its sale. All other authors have declared that no competing interests existed at the time of publication.*

our own participants and the majority of the participants in the work we review would meet the stricter SLI criteria.

In the lexical–semantic domain, bilingual children with DLD and their typically developing (TD) peers often have overlapping performance profiles on standardized measures of single-word vocabulary (Anaya et al., 2018) and, when testing in only one language, in semantics (Peña et al., 2015). This study addresses these challenges by evaluating the role of language experience on English semantic performance in Latino Spanish–English bilingual children.

Language Experience in Bilingual Children

Children's language performance is highly associated with their language experience (Bohman et al., 2010). Accurate diagnosis of language disorders in bilingual children should consider language experiences specific to the child, particularly because of bilinguals' increased risk of misclassification (Grimm & Schulz, 2014; Kraemer & Fabiano-Smith, 2017; Sullivan & Bal, 2013). Language experience, as indexed by current use or age of acquisition, accounts for up to 61.9% of the variance in semantics and 64.4% in morphosyntactic performance in bilingual children (Bedore et al., 2012). On measures of semantics, English exposure correlates with same-language performance on receptive–expressive vocabulary or "lexical breadth" (Hoff et al., 2012, 2014), as well as indices of semantic depth (Bohman et al., 2010). Here, "lexical breadth" refers to the stored lexical entries that a speaker has; this roughly corresponds to words and includes nouns like *árbol* (tree), phrasal verbs like "turn on," and quantifiers like *cada* (every). This construct is typically assessed using single-word vocabulary tests. "Semantic depth" comprises interconnected information within a network that pertains to word meanings (i.e., definitions) and associations such as synonyms and attributes.

Divided language exposure across cultural and linguistic contexts results in distributed lexical knowledge for bilinguals (Pearson et al., 1993; Peña et al., 2002). Differences in performance reflect documented relationships between exposure in a particular language and vocabulary knowledge in that language (Hoff et al., 2012; Pearson et al., 1997; Thordardottir, 2011). This variation in language input reduces exposures to words in a given language, such that bilinguals have comparatively less breadth in each of their languages than their monolingual peers (Hoff et al., 2012; Pearson et al., 1993; Peña et al., 2002). This experience–knowledge relationship holds true for semantic depth. For example, Spanish–English bilingual children have higher repeated association scores in the language in which they have greater experience (Sheng et al., 2013), suggesting that these stronger connections between information within semantic networks depend upon child-specific sociolinguistic experiences (Sheng et al., 2013).

Semantic Difficulties in Children With DLD

Although children with DLD are typically characterized by their morphosyntactic deficits (Rice & Wexler, 1996), weaknesses in semantics are also part of the DLD profile (Leonard, 2014; Schulz & Roeper, 2011). Behavioral differences in children with DLD suggest weaker semantic representations that impact sentence- and discourse-level comprehension. Crucially, difficulty comprehending semantic relationships is associated with poorer academic outcomes and persists into adulthood (McGregor et al., 2013). Monolingual English-speaking children with DLD show difficulties on vocabulary and semantic tasks including word learning (fast mapping; Alt et al., 2004), lexical retrieval (naming speed and accuracy; Lahey & Edwards, 1996, 1999), and semantic depth (generating definitions; McGregor et al., 2002) and repeated word associations (McGregor et al., 2002; Sheng & McGregor, 2010). Poor performance on tasks assessing depth and longitudinal evidence of slower growth in depth compared to growth in breadth suggest that children with DLD have less robust semantic networks (McGregor et al., 2002, 2013; Sheng & McGregor, 2010).

Across languages, bilingual children with DLD show similar patterns of difficulty on semantic depth tasks such as repeated word associations (Sheng et al., 2012) and generating definitions (Gutiérrez-Clellen & DeCurtis, 1999). These differences extend to narratives, with bilingual children with DLD producing significantly fewer of the most frequently occurring words in a given language compared to their TD peers (Shivabasappa et al., 2018). These difficulties cannot be attributed to a lack of language experience (i.e., a language difference), and these findings underscore the need to assess a broad range of semantic tasks beyond lexical breadth.

Limitations of Using Vocabulary Measures to Identify DLD

Monolingual–bilingual differences in exposure have implications for test performance. Bilingual children tested on single-word vocabulary in English often have lower scores than their monolingual peers (Paradis, 2016); this apparent difference disappears when conceptual scoring is used to account for knowledge in both languages. For diagnostic decision making, however, these measures do not meet minimum standards (i.e., classification accuracy > 80%) in monolingual (Gray et al., 1999) or bilingual populations (Anaya et al., 2018), presumably because children with DLD may show age-appropriate performance, thereby obscuring group differences (Spaulding et al., 2013). As a result, measures of single-word vocabulary are minimally informative for diagnosing DLD. Instead, assessing semantic depth classifies more accurately, as children with DLD often lack the depth needed to support their breadth of knowledge (e.g., Peña et al., 2015; Peña, Bedore, & Kester, 2016). Identifying DLD requires tasks that go beyond breadth and target depth, consistent with their observed semantic deficits (Brackenbury & Pye, 2005; McGregor et al., 2013).

Assessing Semantic Performance in Bilingual Children

The observed semantic difficulties in DLD are reflected in TD–DLD group differences on semantic depth tasks in both monolingual (McGregor et al., 2013; Sheng & McGregor, 2010) and bilingual children (Peña et al., 2015; Peña, Bedore, & Kester, 2016; Peña et al., 2003; Sheng et al., 2013, 2012, 2006). Peña et al. (2015) have examined Spanish–English bilingual children's performance on various semantic tasks. These tasks tap into semantic depth by requiring children to retrieve semantic information to make semantic associations, state the functions of objects, and define age-appropriate vocabulary. On these types of tasks, Peña et al. (2003) found task type and language experience to moderate Spanish and English performance across both bilingual and functionally monolingual (English- or Spanish-speaking) groups.

The diagnostic accuracy of this and related semantic depth measures proves clinically informative. Peña et al. (2001) found good classification accuracy for 4- to 7-year-old functionally monolingual children, with 81% classification. For balanced bilinguals, classification accuracy ranged from fair to good, with 76% and 90% classification accuracy using single-language scores in English and Spanish, respectively (Peña et al., 2015). A follow-up study (Peña, Bedore, & Kester, 2016) demonstrated improved classification accuracy (above 85%) when scores from both languages were used together.

Using L2 Testing to Inform Bilingual Assessment

A possible alternative to testing bilingual children in both languages is testing in English while accounting for level of language experience (Bedore et al., 2018; Gillam et al., 2013; Paradis et al., 2013). Several general approaches have addressed this issue by considering the norming population in question. For example, local norms have been developed for community-based bilingual groups by determining appropriate cutoffs (e.g., Junker & Stockman, 2002). Alternatively, monolingual tests have been renormed and adapted for bilingual populations at large (e.g., Expressive One-Word Picture Vocabulary Test–Spanish-Bilingual Edition; Brownell, 2001). While these approaches compare bilingual children against a more representative norm, yielding more valid results, neither the effects of experience nor the items tested are considered. In this vein, Bedore et al. (2018) explored whether the same morphosyntactic items in English discriminate 7- to 10-year-old Spanish–English bilingual children depending on their English (and Spanish) experience. They showed that, for children who used English more or had balanced language experience, many of the same clinical markers used for monolingual speakers (e.g., plural markers, third-person singular) classified children accurately. For children who used more Spanish, complex forms that placed a high demand on memory (e.g., question inversion, relative clauses) were more informative. The current study extends this work to the semantic domain and explores the use of item selection from an English semantics measure to classify children.

Summary and Questions

Accurately identifying DLD in bilingual learners is complicated by overlapping profiles, as low semantic performance in English can be due to limited L2 experience or, instead, to a true semantic deficit. Accounting for bilingual language experience may help disambiguate these patterns of performance. The current study explores the extent to which an index of English semantics from the Bilingual English–Spanish Assessment—Middle Extension (BESA-ME) Experimental Test Version (Peña et al., 2008) can classify DLD in children with varying degrees of English experience. We asked the following questions:

1. Do language ability (i.e., DLD and TD) and experience (i.e., high Spanish experience [HSE], balanced English–Spanish experience [BESE], and high English experience [HEE]) predict children's English semantic knowledge, as measured by the English Semantics subtest of the BESA-ME Experimental Test Version scored in English only?
2. Are there differences in performance across semantic task types as a function of children's language ability and experience, as described above?
3. What is the classification accuracy of a set of semantic items
   a. by language experience group using the same item set and cut-points?
   b. by language experience groups using an item set tailored to each group?

We predicted that TD children would outperform their peers with DLD across language experience groups and that children with more English experience would outperform those with less experience (e.g., Peña et al., 2011; Sheng et al., 2013). For semantic task types, we predicted that some task types would prove more challenging for children at different levels of English experience, consistent with Peña et al. (2003). For classification, we predicted the best classification in children with HEE, and a tailored item approach would improve classification relative to a single set with cutoffs for each experience group.

Method

Participants

We report retrospective data on 327 Latino Spanish–English bilingual children (DLD: $n = 66$) between ages 7;0 and 10;11 (years;months) who were pooled from three studies and are described in the study of Bedore et al. (2018). Participants included 163 children (DLD: $n = 36$) from *Phenotype Assessment Tools for Bilingual (Spanish–English) Children* (Peña & Bedore, 2006), 30 children (DLD: $n = 2$) from *Diagnostic Markers of Language Impairment* (Peña, Bedore, & Gillam, 2006), and 134 children (DLD: $n = 28$)

from *Cross-Language Outcomes of Typical and Atypical Development in Bilinguals* (Peña et al., 2010). Children were recruited through a single-gate design consisting of a screening and confirmatory testing phase, which is considered best practice for minimizing spectrum bias (e.g., Dollaghan & Horner, 2011). DLD was oversampled in each of the parent projects to reach adequate statistical power for planned analyses, yielding a DLD sample that exceeds the 7% prevalence rate reported by Tomblin et al. (1997) for the SLI subtype. For the participant database, selection criteria included use of English and Spanish at least 20% of the time, age between 7 and 10 years, completion of the English Semantics subtest of the BESA-ME Experimental Test Version (Peña et al., 2008), and sufficient data to determine language experience and language ability status. Seventy-seven cases were excluded due to participants' monolingual status. Maternal education, an index of socioeconomic status, had a mean Hollingshead score of 2.73 (SD = 1.60), which corresponds to a partial high school education (Hollingshead, 1975; Scarr & Weinberg, 1978).

Given pooling of children across studies, potential by-study differences and Study × Ability interactions on demographic variables and the outcome variable were tested in a series of two-way analyses of variance (ANOVAs; see Table 1). Cross-study participants had similar proportions of females, and results were nonsignificant for age and maternal education (ps > .05). Study was significant for age of acquisition (p < .001, partial η^2 = .060), and a small but significant effect of study on the semantics index measure emerged (p < .001, partial η^2 = .095). A follow-up sensitivity analysis indicated that running the analyses with and without this subset of children resulted in comparable findings; thus, nested models were not adopted in the analysis, and we report results with all children. No significant Study × Ability interactions emerged.

Reference Standard

The composite reference standard used to determine language ability was based on clinical measures consistent with best practice identification of DLD (e.g., Gladfelter & Leonard, 2013; Gray, 2003; Grinstead et al., 2013; Spaulding et al., 2013). Across studies, children were identified through converging evidence of best language performance: parent/teacher language ratings, semantic performance, morphosyntactic performance, narratives, screening data, and/or clinical judgment by an experienced SLP. While decision rules and measures differed slightly across studies, the same constructs were used, as detailed below.

Parent/teacher language ratings used the Instrument to Assess Language Knowledge questionnaire (ITALK; Peña et al., 2018). Semantic and morphosyntactic performance was measured by (a) the BESA (Peña et al., 2018) or the BESA-ME Field Test Version (Peña, Bedore, Gutiérrez-Clellen, et al., 2016) and (b) the Test of Language Development–Primary: Third Edition (Newcomer & Hammill, 1997). Narrative language was measured by the Test of Narrative Language (Gillam & Pearson, 2004) for English, the Test of Narrative Language–Spanish Adaptation Experimental Version (Gillam et al., 2006) for Spanish, or narratives using wordless picture books. Screening, conducted 1 year prior, used the Bilingual English–Spanish Oral Screener (Peña, Bedore, Gutiérrez-Clellen, et al., 2006). SLP ratings included either DLD referrals or clinical judgment of DLD by an SLP with expertise in bilingualism. These decision rules formed the reference standard for classification analysis (see Bedore et al., 2018, for more information). The single-gate design

Table 1. Cross-study differences by language experience and language ability.

Variable	Study			Main effects	
	Phenotypes (n = 163)	Diagnostic markers (n = 30)	Cross-language outcomes (n = 134)	Study	Ability
Sex[a]	73 F	17 F	70 F		
Age in months	100.78 (10.45)	103.33 (4.68)	99.57 (9.84)	*ns*	*ns*
Maternal education[b]	2.86 (1.60)	2.63 (1.56)	2.57 (1.60)	*ns*	*ns*
English experience	47.11 (14.32)	44.66 (17.22)	45.11 (12.22)	*ns*	.018*
Age of first exposure in years[c]	3.29 (2.08)	1.50 (1.85)	2.82 (1.79)	.060***	.013*
English semantics	23.48 (9.54)	33.30 (4.81)	25.91 (7.86)	.095***	.308***

Note. Effect sizes (partial η^2) are reported for all significant main effects. *ns* = nonsignificant effect; Maternal education = Hollingshead's Index; English semantics = Bilingual English–Spanish Assessment—Middle Extension Experimental Version total raw score, scored in English only.

[a]One missing. [b]Three missing. [c]Four missing.
*p < .05. ***p < .001.

and use of better language performance minimized potential spectrum bias across experience groups. A significant but small negative biserial correlation emerged between ability status and language experience groups, $r(325) = -.129$, $p = .020$, which was interpreted as not clinically meaningful.

Children in the *Phenotypes* study were identified with DLD if they met two of the three indicators: (a) parent/teacher rating score below 4.2 (on a 6-point scale ranging from 0 to 5) on the ITALK in both languages, (b) score 1 *SD* below the mean on the Test of Narrative Language and Test of Narrative Language–Spanish Adaptation Experimental Version in both languages, and (c) DLD identification by a school-based SLP.

In the *Diagnostic Markers* study, DLD status was judged by three expert SLPs who rated transcribed responses on the semantics, morphosyntax, and narrative tasks in English and Spanish on a 6-point scale: 0 (*severe/profound*), 1 (*moderate*), 2 (*mild*), 3 (*low normal*), 4 (*normal*), and 5 (*above normal*). Children were identified with DLD if two of the three SLPs assigned a rating of 2 or lower. These 30 children were originally identified in first grade using indicators from semantics, morphosyntactic, and narrative performance (Gillam et al., 2013). They were later tested in third grade using the index measure.

For the purpose of this study, children in the *Cross-Language Outcomes* study were reclassified without the BESA/BESA-ME semantics indicator to avoid circularity in the analysis, given the overlap of test items in the index measure and original classification procedures. This procedure ensured greater interpretability of the models presented. Classification was determined by children's average performance on the following: ITALK parent/teacher language ratings, Bilingual English–Spanish Oral Screener composite (average of the better morphosyntax and semantics score), Test of Narrative Language narrative performance, and BESA/BESA-ME morphosyntax. For each measure, the higher language score was considered. ITALK parent/teacher raw scores were converted into standard scores ($M = 100$, $SD = 15$), using the entire data set. Morphosyntax was weighted twice. Classification according to this procedure produced results that were highly consistent with original classification using a semantics indicator, with 96% of children not changing classification.

Language Experience Grouping

The BIOS (Peña et al., 2018) measured parent- and teacher-reported information about children's hourly exposure to (i.e., input) and use of (i.e., output) Spanish and English to generate a language experience estimate in each language, where English and Spanish experience variables are inverses of each other. Participants were grouped according to their language experience: Children were considered HSE if they used 20%–39% English, BESE if they used 40%–59% English, and HEE if they used 60%–79% English. Table 2 presents participant characteristics based on these groupings.

Procedure

BESA-ME Index Item Set

The index measure was the English Semantics subtest of the BESA-ME Experimental Test Version (Peña et al., 2008), a 42-item measure that assesses children's receptive/expressive semantic knowledge in English through seven semantic tasks: analogies (three items), associations (eight items), categories (one item), characteristics (seven items), definitions (six items), functions (seven items), and similarities and differences (10 items). This test showed high internal consistency ($\alpha = .91$). Given our focus on English performance, English-only responses were used for expressive items in the reanalysis. Of the seven semantics task types, the categories task was excluded from analysis because there was only one item, yielding a total of six task types and 41 items.

Planned Analyses

Ability- and experience-related differences in bilingual children's English-only semantic performance (Research Question 1) were explored in a two-way ANOVA with language ability and experience as between-subjects factors. Possible task type–related contributions (Research Question 2) were explored in a mixed-model ANOVA, with the percent correct as the dependent variable, language ability and language experience as between-subjects factors, and task type as a within-subject factor. Finally, discriminant function analysis was used for the classification analysis (Research Question 3), and all approaches were analyzed in IBM SPSS 25 or R Studio.

Both language ability and language experience were entered as categorical variables. Language experience cutoffs are consistent with differences in bilingual performance on measures of morphosyntax and semantics in English and Spanish (Peña et al., 2011). Empirical support for our approach comes from research reporting certain thresholds in bilingual children's language learning, and there is a precedence in the literature for categorizing bilingual participants into two or greater dominance/experience groups (e.g., Birdsong, 2014; Yip & Matthews, 2005). Our rationale for dichotomizing to answer our research questions was a clinical one with a theoretical basis.

Results

Performance Across Language Ability and Language Experience

Results showed significant main effects for language ability, $F(1, 321) = 87.28$, $p < .001$, partial $\eta^2 = .21$, and language experience, $F(2, 321) = 6.69$, $p = .001$, partial $\eta^2 = .04$. Effect sizes for partial η^2 are heuristically interpreted as "small" (.01), "medium" (.06), and "large" (.14; Richardson, 2011). Applying these cutoffs, language ability showed a large effect size, and language experience showed a small effect. Bonferroni post hoc pairwise comparisons showed that TD children scored significantly higher ($M = 28.33$) than their peers with DLD ($M = 16.84$; $M_\Delta =$

Table 2. Mean demographic information and language performance by experience and ability.

Measure	HSE (n = 122)		BESE (n = 154)		HEE (n = 51)		All	
Ability status	TD	DLD	TD	DLD	TD	DLD	TD	DLD
n	88	34	129	25	44	7	261	66
Sex	48 F	11 F	64 F	10 F	25 F	2 F	137 F	23 F
Age (months)	98.84 (8.54)	97.79 (10.41)	101.11 (10.15)	98.56 (9.78)	104.34 (9.65)	107.00 (9.87)	100.89 (9.70)	99.06 (10.35)
Maternal education[a]	2.47 (1.45)	2.97 (1.82)	2.63 (1.51)	2.50 (1.47)	3.48 (1.82)	2.57 (1.72)	2.72 (1.58)	2.75 (1.68)
AoA (years)[b]	3.52 (1.75)	3.85 (1.60)	2.44 (1.99)	2.91 (2.33)	2.39 (2.18)	3.71 (1.60)	2.80 (2.01)	3.49 (1.92)
English semantics	89.70 (17.55)	54.30 (20.24)	92.30 (16.16)	63.60 (22.53)	95.70 (12.74)	57.90 (19.14)	92.48 (16.21)	58.78 (21.18)
Spanish semantics[c]	97.97 (12.35)	62.71 (20.70)	96.19 (13.91)	61.80 (24.26)	82.86 (21.12)	70.29 (16.70)	95.05 (15.76)	63.69 (21.60)
English morphosyntax	78.5 (24.52)	34.20 (21.40)	85.70 (22.00)	41.20 (21.80)	98.90 (13.20)	59.90 (23.80)	85.50 (22.70)	39.60 (22.90)
Spanish morphosyntax[d]	98.00 (15.40)	49.60 (18.10)	94.60 (14.90)	47.60 (20.10)	68.32 (37.06)	53.00 (18.39)	91.40 (22.60)	49.20 (18.70)
English vocabulary[e]	3.04 (1.37)	1.73 (1.17)	3.61 (1.24)	2.90 (1.25)	3.98 (1.01)	3.67 (1.03)	3.48 (1.29)	2.37 (1.37)
Spanish vocabulary[f]	4.64 (0.65)	3.94 (1.06)	4.61 (0.72)	3.90 (0.89)	3.98 (1.10)	3.57 (1.51)	4.52 (0.81)	3.88 (1.05)
English sentence length[g]	3.34 (1.39)	2.14 (1.38)	4.38 (0.96)	2.90 (1.33)	4.43 (0.89)	4.33 (0.82)	4.05 (1.21)	2.57 (1.47)
Spanish sentence length[h]	4.84 (0.45)	3.87 (1.18)	4.83 (0.46)	4.19 (0.98)	4.14 (1.10)	4.43 (0.98)	4.72 (0.66)	4.05 (1.09)
English grammar[i]	3.16 (1.21)	2.20 (1.22)	3.84 (0.92)	3.11 (0.66)	4.34 (0.76)	3.40 (0.89)	3.72 (1.08)	2.67 (1.11)
Spanish grammar[j]	4.46 (0.63)	3.71 (1.10)	4.48 (0.75)	3.86 (0.83)	3.71 (1.10)	3.93 (1.07)	4.38 (0.81)	3.70 (1.01)

Note. HSE = high Spanish experience; BESE = balanced English–Spanish experience; HEE = high English experience; TD = typically developing; DLD = developmental language disorder; Maternal education = Hollingshead index; AoA = age of first English acquisition; English semantics = Bilingual English–Spanish Assessment—Middle Extension (BESA-ME) Field Test Version standard score; Spanish semantics = BESA-ME Field Test Version standard score; English morphosyntax = BESA-ME Field Test Version standard score; Spanish morphosyntax = BESA-ME Field Test Version standard score; English vocabulary = Instrument to Assess Language Knowledge (ITALK) parent rating; Spanish vocabulary = ITALK parent rating; English sentence length = ITALK parent rating; Spanish sentence length = ITALK parent rating; English grammar = ITALK parent rating; Spanish grammar = ITALK parent rating.

[a]Three missing. [b]Four missing. [c]Ten missing. [d]Twelve missing. [e]Sixteen missing. [f]Eleven missing. [g]Twenty-nine missing. [h]Eleven missing. [i]Forty-nine missing. [j]Ten missing.

11.49, SE = 1.23, $p < .001$). Language experience also predicted semantic performance: Both the HEE (M = 25.26; M_Δ = 5.41, SE = 1.66, p = .004) and BESE (M = 22.65; M_Δ = 2.80, SE = 1.09, p = .032) groups scored significantly higher than the HSE group (M = 19.85). The HEE and BESE groups, however, did not differ significantly. The Ability × Experience interaction was not significant, $F(2, 321) = 0.82$, p = .444, partial η^2 = .005.

Semantic Task Type

Differences in performance by task type, language ability, and experience were examined in a mixed-model ANOVA. Table 3 shows task type accuracy across experience groups. As before, there were significant main effects for language ability, $F(1, 320) = 81.37$, $p < .001$, partial η^2 = .20, and experience, $F(2, 320) = 5.91$, p = .003, partial η^2 = .04. Mauchly's test of sphericity was significant for task type, indicating that group sizes were not equal. Greenhouse–Geisser corrected results showed a significant main effect for task type, $F(3.92, 1256.20) = 47.15$, $p < .001$, partial η^2 = .133, and a small but significant Task Type × Ability interaction, $F(3.92, 1256.20) = 4.22$, p = .002, partial η^2 = .013. The Ability × Experience and Task Type × Experience interactions were not significant, and experience was not explored further.

The effects of language ability on semantic task types were explored in a series of univariate ANOVAs, with the percent correct on each task type as the dependent variable. For functions and analogies, Levene's statistic was significant, indicating that the assumption of homogeneity of variance was not met. To account for this violation, a stricter significance level of $p < .01$ was used for these two task types (Stevens, 2012); the original results are reported for values that remained significant using this criterion. There was a significant main effect for language ability. The TD group outperformed children with DLD on each task type: characteristics, $F(1, 320) = 84.34$, $p < .001$, partial η^2 = .209;

Table 3. Semantic task type percent accuracy (standard deviation) across experience groups.

Task type	HSE		BESE		HEE		All	
	TD	DLD	TD	DLD	TD	DLD	TD	DLD
Characteristics	75.97 (19.68)	39.39 (24.23)	77.08 (20.16)	46.29 (20.73)	78.57 (16.45)	53.06 (17.91)	76.96 (19.37)	43.52 (22.51)
Definitions	67.61 (27.84)	39.39 (24.59)	70.28 (25.34)	50.67 (28.25)	74.24 (20.79)	61.91 (20.89)	70.05 (25.53)	46.15 (26.48)
Functions	59.74 (26.66)	29.00 (22.72)	69.88 (22.75)	44.00 (28.25)	80.84 (18.21)	44.89 (26.64)	68.30 (24.50)	36.48 (26.13)
Associations	63.78 (21.70)	34.09 (19.08)	65.60 (21.93)	41.00 (22.97)	72.16 (19.23)	50.00 (27.00)	66.09 (21.53)	38.46 (21.80)
Similarities and differences	63.86 (19.56)	34.55 (19.22)	65.66 (19.72)	38.00 (21.21)	72.50 (18.44)	32.86 (31.47)	66.21 (19.61)	35.69 (21.21)
Analogies	39.02 (33.61)	11.11 (18.00)	40.83 (29.24)	26.67 (31.92)	43.94 (33.54)	33.33 (27.22)	40.74 (31.43)	19.49 (26.28)
Total	63.85 (17.94)	32.13 (17.44)	67.06 (17.00)	41.52 (19.90)	73.06 (12.41)	45.93 (17.10)	66.99 (16.88)	37.15 (18.88)

Note. HSE = high Spanish experience; BESE = bilingual English–Spanish experience; HEE = high English experience; TD = typically developing; DLD = developmental language disorder.

definitions, $F(1, 320) = 27.71$, $p < .001$, partial $\eta^2 = .064$; functions, $F(1, 320) = 58.98$, $p < .001$, partial $\eta^2 = .156$; associations, $F(1, 320) = 49.86$, $p < .001$, partial $\eta^2 = .135$; similarities and differences, $F(1, 320) = 97.80$, $p < .001$, partial $\eta^2 = .225$; and analogies, $F(1, 320) = 11.79$, $p < .001$, partial $\eta^2 = .036$.

Classification Analyses

English semantics items were analyzed to determine the combination that best classified TD and DLD groups. Item-level metrics were calculated separately for the HSE, BESE, and HEE groups. Item difficulty, the percentage of accurate responses for each test item, was calculated for TD children and children with DLD, and a discrimination index was calculated as the difference in item difficulty between the ability groups. Items with a difficulty difference of ≥ .30 met the criterion for further consideration (Allen & Yen, 2002; Friedenberg, 1995). Table 4 presents the 32 semantic items that met this set of criteria for at least one experience group. Of the 41 total items, the following numbers met the criteria: 21 for HSE children, 13 for BESE children, and 18 for the HEE group (see Table 5 for item counts by task type). Items answered correctly by fewer than 60% of the TD children were deemed too difficult and were discarded. The final items retained were 15 for HSE, 11 for BESE, and 13 for HEE. Items showed acceptable internal consistency for the HSE ($\alpha = .87$), BESE ($\alpha = .78$), and HEE ($\alpha = .82$) groups. Three approaches were used to optimally classify cases (see Table 6). In all approaches, leave-one-out cross-validation was the resampling technique used. If group covariances could be assumed equal, as indexed by Box's M and evaluated using a significance level of .01, analyses were run using a within-groups covariance matrix. When this assumption was violated, analyses were run using separate-groups covariance matrices (Burns & Burns, 2008).

For each approach, measures of sensitivity, specificity, positive likelihood ratio (LR+), and negative likelihood ratio (LR−) are reported. Sensitivity and specificity quantify the correct classification of DLD and TD cases, respectively. In contrast, "likelihood ratios" predict the probability of a test result indicating impairment (LR+) or typical language (LR−). LR+/LR− values can aid in clinical decision making. LR+ values of > 10 indicate that a score in the affected range is almost certain to be true (i.e., true DLD). Similarly, LR− values of < 0.1 indicate that a score in the typical range is almost certain to be true (Dollaghan & Horner, 2011). These values are used to evaluate the relative diagnostic utility of each approach for each language experience group.

Classification Using a Single Set of Items

The procedure for maximizing ability-related differences and minimizing experiential factors for bilingual children with 40%–79% English experience (i.e., HEE and BESE) was based on the study of Bedore et al. (2018). For these two groups, nine items met the discriminant criteria of > .30, had a HEE–BESE difficulty difference of ≤ .20, and had a TD difficulty index of > .60 (Friedenberg, 1995). Total raw scores for these nine items were entered into the discriminant analysis. Sensitivity (se) and specificity (sp) quantify the correct classification.

Classification for pooled HEE–BESE groups is reported. TD children and those with DLD had an average score of 6.84 and 3.31, respectively. The assumption of equality for the group covariance matrices was not met ($p = .016$), and the log determinants were dissimilar (TD = 1.15, DLD = 1.78). The resulting classification using a separate-groups covariance matrix improved, and we report this solution. The chi-square test was significant (Wilk's $\lambda = .68$, $\chi^2 = 77.02$, $df = 1$, canonical correlation = .562, $p < .001$). For the HEE–BESE group, 86.8% of cases were correctly classified (se = 71.9%, sp = 89.6%, LR+ = 6.91, LR− = 0.31),

Table 4. Semantic item difficulty across language experience.

Item	Task type	HSE		BESE		HEE	
		TD	DLD	TD	DLD	TD	DLD
27	Associations	**.38**	**.03**	.42	.16	.41	.14
41	Analogies	**.40**	**.03**	.40	.36	.34	.29
31	Definitions	**.51**	**.21**	.57	.40	.66	.43
24	Associations	**.63**	**.15**	.53	.36	.55	.71
25	Associations	**.64**	**.30**	.61	.36	.73	.57
29	Definitions	**.69**	**.36**	.73	.64	.73	.86
19	Functions	**.69**	**.21**	.79	.60	.86	.57
7	Characteristics	**.83**	**.52**	.88	.76	.98	.86
36	Similarities and differences	**.44**	**.06**	**.46**	**.12**	.59	.43
20	Functions	**.68**	**.27**	**.71**	**.36**	.80	.86
37	Characteristics	**.70**	**.36**	**.69**	**.16**	.57	.43
34	Associations	**.70**	**.36**	**.74**	**.36**	.84	.57
9	Characteristics	**.81**	**.27**	**.78**	**.36**	.80	.71
35	Similarities and differences	**.44**	**.03**	.43	.20	**.59**	**.14**
30	Definitions	**.60**	**.27**	.68	.48	**.89**	**.43**
18	Functions	**.85**	**.39**	.93	.68	**.89**	**.43**
2	Similarities and differences	**.95**	**.62**	.93	.80	**1**	**.43**
32	Similarities and differences	**.59**	**.15**	**.62**	**.16**	**.73**	**.29**
16	Characteristics	**.61**	**.21**	**.67**	**.32**	**.73**	**.43**
13	Characteristics	**.63**	**.24**	**.65**	**.28**	**.73**	**.29**
8	Similarities and differences	**.83**	**.39**	**.87**	**.40**	**.89**	**.29**
11	Definitions	.69	.52	**.74**	**.44**	.75	.57
22	Functions	.32	.03	**.41**	**.08**	**.66**	**0**
23	Associations	.65	.42	**.81**	**.44**	**.89**	**.57**
4	Functions	.82	.56	**.91**	**.60**	**1**	**.57**
42	Analogies	.24	0	.22	.04	**.32**	**0**
26	Associations	.26	0	.32	.12	**.50**	**.14**
39	Similarities and differences	.33	.06	.32	.20	**.50**	**.14**
38	Functions	.40	.21	.52	.32	**.50**	**0**
33	Similarities and differences	.56	.39	.52	.24	**.64**	**.29**
17	Similarities and differences	.67	.48	.68	.56	**.68**	**.14**
6	Associations	.97	.76	.92	.72	**.91**	**.57**

Note. Bilingual English–Spanish Assessment—Middle Extension Semantics items in bold met the discrimination index of ≥ .30 for that experience group. For each combination of experience groupings, items are ordered by difficulty for the TD group. HSE = high Spanish experience; BESE = balanced English–Spanish experience; HEE = high English experience; TD = typically developing; DLD = developmental language disorder.

indicating unacceptably low sensitivity and high specificity. The LR+ indicates that a score in the impaired range is suggestive of true DLD, while the LR− is uninformative for a score in the normal range.

Cross-validation was used to assess the classification of HSE children using the same cut score of 5.08 as the HEE–BESE pooled group. The resulting classification accuracy was 71.3% (se = 88.2%, sp = 64.8%, LR+ = 2.50, LR− = 0.18). Given the low LR+ for the cross-validated HSE group, a score in the impaired range is uninformative. On the other hand, the LR– value indicates that a score in the typical range is suggestive of TD. Overall, this procedure overidentified HSE children as having DLD. We then evaluated whether revising cut-points by language experience group would improve classification.

Classification Using Revised Cut-Points

HSE performance was evaluated on the same nine items from before, with the following group scores: TD = 6.34, DLD = 2.91. As indexed by Box's M, group covariance matrices were equal (p = .553), and the log determinants were similar (TD = 1.30, DLD = 1.48). The chi-square test was significant (Wilk's λ = .62, χ^2 = 57.52, df = 1, canonical correlation = .618, p < .001). The total raw score classified 79.5% of the cases (se = 76.5%, sp = 80.7%, LR+ = 3.96, LR− = 0.29). This revised cut score for HSE children decreased sensitivity and increased specificity, with overall improved classification. The LR+ and LR– indicate that scores of TD or DLD are suggestive, with an overall bias toward underidentification.

For the BESE group, the average raw scores were calculated (TD = 6.74, DLD = 3.08). The assumption of equality of the group covariance matrices was met (p = .182), and the log determinants were similar across groups (TD = 1.24, DLD = 1.64). The chi-square test was significant (Wilk's λ = .67, χ^2 = 60.99, df = 1, canonical correlation = .576, p < .001). The cut score classified 84.4% of cases (se = 72.0%, sp = 86.8%, LR+ = 5.46, LR− = 0.32). BESE classification was largely unchanged using this approach.

For the HEE group, average raw scores were 7.16 and 4.14 for TD children and children with DLD, respectively. The assumption of equality of group covariance matrices

Table 5. Number of items (%) that accurately classified each experience group.

Group	Characteristics (7)	Definitions (6)	Functions (7)	Associations (8)	Similarities and differences (10)	Analogies (3)	Total (41)
HSE	5 (33.3)	2 (13.3)	3 (20.0)	3 (20.0)	2 (13.3)	0 (0.0)	15 (100.0)
BESE	4 (36.4)	1 (9.1)	2 (18.2)	2 (18.2)	2 (18.2)	0 (0.0)	11 (100.0)
HEE	2 (15.4)	1 (7.7)	3 (23.1)	2 (15.4)	5 (38.5)	0 (0.0)	13 (100.0)

Note. HSE = high Spanish experience; BESE = balanced English–Spanish experience; HEE = high English experience.

was not met ($p = .012$), and the log determinants for the two groups were dissimilar (TD = .82, DLD = 2.18). To correct for this, analyses were rerun using a separate-groups covariance matrix; we report the updated results. The chi-square test was significant (Wilk's $\lambda = .73$, $\chi^2 = 15.07$, $df = 1$, canonical correlation = .517, $p < .001$). The total raw score classified with 84.3% accuracy (se = 71.4%, sp = 86.4%, LR+ = 5.24, LR− = 0.33). As with the BESE group, revising the cut score resulted in little change to classification. For both groups, positive test results are suggestive of a DLD diagnosis. The high LR−, however, is uninformative, making it difficult to rule out impairment with a score in the TD range.

Classification Using Tailored Item Sets

A limited number of items met the differentiation criteria across levels of English and Spanish experience. Given the differences in performance on item types across the HSE and BESE groups, we ran a third set of analyses to determine if using the items that met the criteria within each language experience group would improve classification. This time, raw scores from the experience-tailored items (i.e., 15 items for HSE, 11 for BESE, and 13 for HEE) were entered into the discriminant analysis separately for each group.

For the HSE group, the average raw scores of the selected 15 items for TD and DLD groups were 10.77 and 4.33, respectively. Group covariance matrices could be assumed equal ($p = .476$), and log determinants were similar (TD = 2.31, DLD = 2.53). The chi-square test was significant (Wilk's $\lambda = .58$, $\chi^2 = 65.57$, $df = 1$, canonical correlation = .650, $p < .001$). Tailored items classified 83.6% of the cases (se = 76.7%, sp = 85.9% , LR+ = 5.43, LR− = 0.27). Relative to a revised cut score for the HEE–BESE items, here, sensitivity did not improve, while specificity was highest using this approach. Both LR+ and LR− values remain suggestive. Therefore, using a tailored approach, HSE children's scores above or below the cut had suggestive likelihood to reflect their true ability status.

For the BESE group, average raw scores were 8.19 and 3.88 for the TD and DLD groups, respectively, for the 11 items selected. Box's M was nonsignificant ($p = .279$), and log determinants were similar across groups (TD = 1.63, DLD = 1.96). The chi-square test was significant (Wilk's $\lambda = .68$, $\chi^2 = 58.62$, $df = 1$, canonical correlation = .566, $p < .001$). The total raw score accurately classified 77.9% of cases (se = 76.0%, sp = 78.3%, LR+ = 3.50, LR− = 0.31). Compared to the revised cut score approach above, cut scores derived from a tailored item set slightly improved

Table 6. Comparison of three approaches for classifying experience groups.

Approach	Sensitivity (95% CI)	Specificity (95% CI)	Likelihood of DLD (95% CI)	Likelihood of TD (95% CI)
Single set				
HSE	**88.2% [72.5, 96.7]**	64.8% [53.9, 74.7]	2.50 [1.84, 3.41]	**0.18 [0.07, 0.46]**
HEE–BESE	71.9% [53.3, 86.3]	**89.6% [84.1, 93.7]**	**6.91 [4.24, 11.25]**	0.31 [0.18, 0.55]
Revised cut-point				
HSE	76.5% [58.8, 89.3]	80.7% [70.9, 88.3]	3.96 [2.48, 6.31]	0.29 [0.16, 0.54]
BESE	72.0% [50.6, 87.9]	86.8% [79.7, 92.1]	5.46 [3.29, 9.06]	0.32 [0.17, 0.61]
HEE	71.4% [29.0, 96.3]	86.4% [72.6, 94.8]	5.24 [2.18, 12.61]	0.33 [0.10, 1.07]
Tailored items				
HSE	76.7% [57.7, 90.1]	**85.9% [77.0, 92.3]**	**5.43 [3.16, 9.32]**	0.27 [0.14, 0.52]
BESE	**76.0% [54.9, 90.6]**	78.3% [70.2, 85.1]	3.50 [2.36, 5.20]	**0.31 [0.15, 0.62]**
HEE	**85.7% [42.1, 99.6]**	**95.5% [84.5, 99.4]**	**18.86 [4.71, 75.51]**	**0.15 [0.02, 0.92]**

Note. The highest diagnostic measures for each experience groups are bolded. CI = confidence interval; DLD = developmental language disorder; TD = typically developing; HSE = high Spanish experience; HEE = high English experience; BESE = balanced English–Spanish experience.

sensitivity and decreased specificity, resulting in unacceptable accuracy for the BESE group.

For the HEE group, the average raw score was 10.61 for the TD children and 4.71 for the children with DLD. Box's M was statistically significant ($p = .013$), violating the assumption of homogeneity of covariance matrices, and the log determinants were dissimilar (TD = 1.21, DLD = 2.56). A discriminant analysis using separate-groups covariance matrix was run, but the classification did not improve. Thus, we report the original results. The chi-square test was significant (Wilk's $\lambda = .51$, $\chi^2 = 32.32$, $df = 1$, canonical correlation = .697, $p < .001$). The total raw score accurately classified 94.1% of cases (se = 85.7%, sp = 95.5%, LR+ = 18.86, LR− = 0.15). This tailored approach resulted in the highest sensitivity/specificity for this group and overall good classification accuracy. The high LR+ indicates that scores in the impaired range are informative for diagnosing DLD, and the acceptably low LR– indicates that scores in the normal range are suggestive for TD. Table 6 compares the diagnostic accuracy across the three approaches.

Discussion

This study explored how well an English semantics measure could classify school-age bilingual children with and without DLD who had varying levels of English experience. Significant group differences in performance emerged as a function of language ability and language experience. English semantics testing proves to be diagnostically useful for this population, with tailored item sets yielding acceptable levels of classification accuracy across language experience groups.

Language ability also predicted task type performance. The significant task type by language ability interaction (i.e., differential performance depending on ability status) is consistent with previous findings of TD–DLD group differences (e.g., McGregor et al., 2002, 2013). While the precise nature of these task-modulated differences is beyond the scope of this article, semantic skills that are especially challenging for children with DLD may be informative indicators of impairment. Our predictions of experience-related contributions to task type were not supported in this older group of children.

Of the three approaches used to classify children, using items selected according to language experience yielded the highest sensitivity and specificity for the HEE group, with good classification. These tailored item sets yielded acceptable levels for HSE group, but not for the balanced group (i.e., BESE). For the HSE and HEE experience groups, a score below the cut on their respective combination of items was indicative of DLD, but a score in the normal range did not necessarily indicate TD status. Even for children with less L2 experience (i.e., HSE), English semantic performance effectively ruled in impairment. The high specificity came at the cost of underidentifying DLD. One reason that tailoring items to language experience improves classification is that items are sensitive to the amount of linguistic knowledge. This aligns with findings that show observed semantic difficulties across various behavioral measures and the effect of bilingual experience on semantic performance (e.g., Brackenbury & Pye, 2005; McGregor et al., 2013; Peña et al., 2015; Peña, Bedore, & Kester, 2016; Sheng et al., 2013, 2012, 2006). Tailoring item sets according to language experience maximized observed TD–DLD differences. It is possible that further research may reveal by-experience sequences.

The BESE group had greater variability than the other two experience groups. This resulted in similar performance to the HEE group but the least optimal classification of the three experience groups. Although the first analysis did not show a significant group difference between the HEE and BESE children, the classification analyses demonstrated that the power of the same items to detect impairment was compromised for the BESE children. Previous studies (e.g., Bedore et al., 2012; Bohman et al., 2010; Hammer et al., 2012) report similar patterns, where balanced bilingual groups perform similarly to children with greater English experience. In contrast, Peña et al. (2011) found significant group differences for younger HEE and BESE groups. It may be that additional English language experience affords older children the opportunity to "catch up" to their English-dominant counterparts. Bedore et al. (2012) found that performance on a semantics screener became more similar above 75% of English use, suggesting a minimum language experience threshold is needed to perform well on tests in the L2.

For the HSE group, we anticipated poorer classification given their limited experience in English, as was shown for morphosyntax in the study of Bedore et al. (2018). However, when using the same items across groups, our results showed marginally acceptable classification accuracy of 79.5% (se = 76.5%, sp = 80.7%) with a revised cut. Classification improved to 83.6% (se = 76.7%, sp = 85.9%) when tailored sets were used. Previous investigations testing English morphosyntactic productions have found that bilingual children with less than 30% experience in English are likely to be misclassified (Bedore et al., 2018; Gutiérrez-Clellen et al., 2008). However, L2 skills across language domains emerge at distinct rates. Previous work (Bedore et al., 2012; Peña et al., 2014) has shown that bilingual children require less language experience to approach TD norms on semantics tasks compared to morphosyntactic tasks. It is possible that the differences in demands of the language tasks employed across these studies explain some of the differences in classification outcomes.

The children in this study were older school-age children with at least 20% current English experience; thus, findings may not be applicable to younger bilingual children, who would have less cumulative experience. For children with less experience, using English-only testing to diagnose impairment is not recommended. Instead, best practices for bilingual assessment—for example, testing in both languages (Bedore et al., 2012) and using nonstandardized methods to evaluate language learning potential (Peña et al., 2001)—would better capture children's language ability. For example, dynamic assessment (e.g., Kapantzoglou

et al., 2012; Peña et al., 2014, 2001) could improve diagnostic accuracy by minimizing experiential factors.

Examining English-only performance in bilingual children has implications for practitioners and for innovations in assessment. In the absence of bilingual SLPs or interpreters, available resources are used to assess bilingual children's language abilities. To this end, English language testing can highlight areas of relative strength and areas that would benefit most from increased English exposure. Mapping out group patterns by both ability and experience situates assessment findings within the appropriate bilingual context and is especially critical when assessing children of different language backgrounds. For example, computerized adaptive testing, an assessment method that allows the difficulty level and number of questions to be individualized, might be developed to select items tailored to a particular language experience level in real time to avoid floor or ceiling effects and optimize classification (e.g., Bachman, 2000).

Limitations and Future Research

These findings should be considered in light of several methodological limitations: experience groupings and item analysis approaches. Our decision to use experience groupings over a continuous approach was clinically informed (Peña et al., 2011), with the purpose of guiding clinical decision making in bilingual assessment. Because SLPs are likely to rely on discrete levels of experience or dominance levels to make clinical decisions, as suggested in Kritikos (2003), findings aligning with these procedures may be more interpretable. However, this approach may not fully capture subtle effects. Moreover, specific cutoffs need to be further validated across age groups and with other samples before concrete cutoff recommendations can be made. A final point relates to the item analysis procedure used here. A procedure incorporating item response theory might yield greater information about item-level statistics (Fan, 1998). Future work using bilingual language experience as a continuous variable could be used in general questions of performance to better understand experiential effects on language ability and English performance in the school-age years.

Conclusions

This study found English semantic assessment that accounts for bilingual experience can be clinically informative, with significant differences in ability across language experience. Compared to a local norm approach, this method better maximized TD–DLD differences by considering the child's specific language experience and using challenging items.

Tailoring semantic items to children's language experience resulted in acceptable classification in the HSE and HEE groups. It should be noted, however, that while the English semantics measure was a useful indicator of DLD, classification was poorer than what was found in the morphosyntactic domain in the same group of children (Bedore et al., 2018); therefore, we recommend that children's semantic performance be complemented with other evidence-based practices, including use of morphosyntactic scores (e.g., Bedore et al., 2018) and parent/teacher report (e.g., Paradis, 2017). Future research should continue to explore how a tailored-items approach of English measures, as well as further validation of experience cutoffs, can be leveraged to diagnose DLD in bilingual children.

Acknowledgments

This research was supported in part by National Institute on Deafness and Other Communication Disorders Grants R21 HD053223, R01 DC007439, and R01 DC010366 (Principal Investigator: Elizabeth D. Peña).

References

Allen, M. J., & Yen, W. (2002). *Introduction to measurement theory*. Waveland Press.

Alt, M., Plante, E., & Creusere, M. (2004). Semantic features in fast-mapping: Performance of preschoolers with specific language impairment versus preschoolers with normal language. *Journal of Speech, Language, and Hearing Research, 47*(2), 407–420. https://doi.org/10.1044/1092-4388(2004/033)

American Speech-Language-Hearing Association. (2016). *2016 Schools survey report: SLP caseload characteristics*. www.asha.org/research/memberdata/schoolssurvey/

Anaya, J. B., Peña, E. D., & Bedore, L. M. (2018). Conceptual scoring and classification accuracy of vocabulary testing in bilingual children. *Language, Speech, and Hearing Services in Schools, 49*(1), 85–97. https://doi.org/10.1044/2017_LSHSS-16-0081

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing, 17*(1), 1–42. https://doi.org/10.1177/026553220001700101

Bedore, L. M., Peña, E. D., Anaya, J. B., Nieto, R., Lugo-Neris, M. J., & Baron, A. (2018). Understanding disorder within variation: Production of English grammatical forms by English language learners. *Language, Speech, and Hearing Services in Schools, 49*(2), 277–291. https://doi.org/10.1044/2017_LSHSS-17-0027

Bedore, L. M., Peña, E. D., Summers, C. L., Boerger, K. M., Resendiz, M. D., Greene, K., Bohman, T. M., & Gillam, R. B. (2012). The measure matters: Language dominance profiles across measures in Spanish–English bilingual children. *Bilingualism: Language and Cognition, 15*(3), 616–629. https://doi.org/10.1017/S1366728912000090

Birdsong, D. (2014). Dominance and age in bilingualism. *Applied Linguistics, 35*(4), 374–392. https://doi.org/10.1093/applin/amu031

Bishop, D. V., Snowling, M. J., Thompson, P. A., Greenhalgh, T., & CATALISE-2 Consortium. (2017). Phase 2 of CATALISE: A multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology. *Journal of Child Psychology and Psychiatry, 58*(10), 1068–1080. https://doi.org/10.1111/jcpp.12721

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*(3), 325–344. https://doi.org/10.1080/13670050903342019

Brackenbury, T., & Pye, C. (2005). Semantic deficits in children with language impairments: Issues for clinical assessment. *Language, Speech, and Hearing Services in Schools, 36*(1), 5–16. https://doi.org/10.1044/0161-1461(2005/002)

Brownell, R. (2001). *Expressive One-Word Picture Vocabulary Test–Spanish-Bilingual Edition*. Academic Therapy Publications.

Burns, R. P., & Burns, R. (2008). Discriminant analysis. In R. P. Burns & R. Burns (Eds.), *Business research methods and statistics using SPSS* (pp. 589–608). Sage.

Dollaghan, C. A., & Horner, E. A. (2011). Bilingual language assessment: A meta-analysis of diagnostic accuracy. *Journal of Speech, Language, and Hearing Research, 54*(4), 1077–1088. https://doi.org/10.1044/1092-4388(2010/10-0093)

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357–381. https://doi.org/10.1177/0013164498058003001

Friedenberg, L. (1995). *Psychological testing: Design, analysis, and use*. Allyn & Bacon.

Gillam, R. B., & Pearson, N. A. (2004). *TNL: Test of Narrative Language*. 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: I. Assessment in English. *Journal of Speech, Language, and Hearing Research, 56*(6), 1813–1823. https://doi.org/10.1044/1092-4388(2013/12-0056)

Gillam, R. B., Peña, E. D., Bedore, L. M., & Pearson, N. (2006). *Test of Narrative Language–Spanish Adaptation Experimental Version (TNL-S)*. Pro-Ed.

Gladfelter, A., & Leonard, L. B. (2013). Alternative tense and agreement morpheme measures for assessing grammatical deficits during the preschool period. *Journal of Speech, Language, and Hearing Research, 56*(2), 542–552. https://doi.org/10.1044/1092-4388(2012/12-0100)

Gray, S. (2003). Diagnostic accuracy and test–retest reliability of nonword repetition and digit span tasks administered to preschool children with specific language impairment. *Journal of Communication Disorders, 36*(2), 129–151. https://doi.org/10.1016/S0021-9924(03)00003-0

Gray, S., Plante, E., Vance, R., & Henrichsen, M. (1999). The diagnostic accuracy of four vocabulary tests administered to preschool-age children. *Language, Speech, and Hearing Services in Schools, 30*(2), 196–206. https://doi.org/10.1044/0161-1461.3002.196

Grimm, A., & Schulz, P. (2014). Specific language impairment and early second language acquisition: The risk of over- and underdiagnosis. *Child Indicators Research, 7*(4), 821–841. https://doi.org/10.1007/s12187-013-9230-6

Grinstead, J., Baron, A., Vega-Mendoza, M., De la Mora, J., Cantú-Sánchez, M., & Flores, B. (2013). Tense marking and spontaneous speech measures in Spanish specific language impairment: A discriminant function analysis. *Journal of Speech, Language, and Hearing Research, 56*(1), 352–363. https://doi.org/10.1044/1092-4388(2012/11-0289)

Gutiérrez-Clellen, V. F., & DeCurtis, L. (1999). Word definition skills in Spanish-speaking children with language impairment. *Communication Disorders Quarterly, 21*(1), 23–31. https://doi.org/10.1177/152574019902100104

Gutiérrez-Clellen, V. F., Simon-Cereijido, G., & Wagner, C. (2008). Bilingual children with language impairment: A comparison with monolinguals and second language learners. *Applied Psycholinguistics, 29*(1), 3–19. https://doi.org/10.1017/S0142716408080016

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*(5), 1251–1264. https://doi.org/10.1044/1092-4388(2012/11-0016)

Hoff, E., Core, C., Place, S., Rumiche, R., Señor, M., & Parra, M. (2012). Dual language exposure and early bilingual development. *Journal of Child Language, 39*(1), 1–27. https://doi.org/10.1017/S0305000910000759

Hoff, E., Rumiche, R., Burridge, A., Ribot, K. M., & Welsh, S. N. (2014). Expressive vocabulary development in children from bilingual and monolingual homes: A longitudinal study from two to four years. *Early Childhood Research Quarterly, 29*(4), 433–444. https://doi.org/10.1016/j.ecresq.2014.04.012

Hollingshead, A. B. (1975). Four factor index of social status. *Yale Journal of Sociology, 8,* 21–52.

Junker, D. A., & Stockman, I. J. (2002). Expressive vocabulary of German–English bilingual toddlers. *American Journal of Speech-Language Pathology, 11*(4), 381–394. https://doi.org/10.1044/1058-0360(2002/042)

Kapantzoglou, M., Restrepo, M. A., & Thompson, M. S. (2012). Dynamic assessment of word learning skills: Identifying language impairment in bilingual children. *Language, Speech, and Hearing Services in Schools, 43*(1), 81–96. https://doi.org/10.1044/0161-1461(2011/10-0095)

Kraemer, R., & Fabiano-Smith, L. (2017). Language assessment of Latino English learning children: A records abstraction study. *Journal of Latinos and Education, 16*(4), 349–358. https://doi.org/10.1080/15348431.2016.1257429

Kritikos, E. P. (2003). Speech-language pathologists' beliefs about language assessment of bilingual/bicultural individuals. *American Journal of Speech-Language Pathology, 12*(1), 73–91. https://doi.org/10.1044/1058-0360(2003/054)

Lahey, M., & Edwards, J. (1996). Why do children with specific language impairment name pictures more slowly than their peers? *Journal of Speech and Hearing Research, 39*(5), 1081–1098. https://doi.org/10.1044/jshr.3905.1081

Lahey, M., & Edwards, J. (1999). Naming errors of children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 42*(1), 195–205. https://doi.org/10.1044/jslhr.4201.195

Leonard, L. B. (2014). *Children with specific language impairment*. MIT Press.

McGregor, K. K., Newman, R. M., Reilly, R. M., & Capone, N. C. (2002). Semantic representation and naming in children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 45*(5), 998–1014. https://doi.org/10.1044/1092-4388(2002/081)

McGregor, K. K., Oleson, J., Bahnsen, A., & Duff, D. (2013). Children with developmental language impairment have vocabulary deficits characterized by limited breadth and depth. *International Journal of Language & Communication Disorders, 48*(3), 307–319. https://doi.org/10.1111/1460-6984.12008

Newcomer, P. L., & Hammill, D. D. (1997). *Test of Language Development–Primary: Third Edition*. Pro-Ed.

Paradis, J. (2016). The development of English as a second language with and without specific language impairment: Clinical implications. *Journal of Speech, Language, and Hearing Research, 59*(1), 171–182. https://doi.org/10.1044/2015_JSLHR-L-15-0008

Paradis, J. (2017). Parent report data on input and experience reliably predict bilingual development and this is not trivial. *Bilingualism: Language and Cognition, 20*(1), 27–28. https://doi.org/10.1017/S136672891600033X

Paradis, J., Schneider, P., & Duncan, T. S. (2013). Discriminating children with language impairment among English-language learners from diverse first-language backgrounds. *Journal of Speech, Language, and Hearing Research, 56*(3), 971–981. https://doi.org/10.1044/1092-4388(2012/12-0050)

Pearson, B. Z., Fernández, S. C., Lewedeg, V., & Oller, D. K. (1997). The relation of input factors to lexical learning by bilingual infants. *Applied Psycholinguistics, 18*(1), 41–58. https://doi.org/10.1017/S0142716400009863

Pearson, B. Z., Fernández, S. C., & Oller, D. K. (1993). Lexical development in bilingual infants and toddlers: Comparison to monolingual norms. *Language Learning, 43*(1), 93–120. https://doi.org/10.1111/j.1467-1770.1993.tb00174.x

Peña, E. D., Iglesias, A., & Lidz, C. S. (2001). Reducing test bias through dynamic assessment of children's word learning ability. *American Journal of Speech-Language Pathology, 10*(2), 138–154. https://doi.org/10.1044/1058-0360(2001/014)

Peña, E. D., & Bedore, L. M. (2006). *Phenotype assessment tools for bilingual (Spanish–English) children*. National Institute on Deafness and Other Communication Disorders.

Peña, E. D., Bedore, L. M., & Baron, A. (2017). Bilingualism and child language disorders. In R. G. Schwartz (Ed.), *Handbook of child language disorders* (2nd ed., pp. 297–327). Psychology Press.

Peña, E. D., Bedore, L. M., & Gillam, R. B. (2006). *Diagnostic markers of language impairment in Spanish–English bilinguals*. National Institute on Deafness and Other Communication Disorders.

Peña, E. D., Bedore, L. M., & Griffin, Z. (2010). *Cross-language outcomes of typical and atypical development in bilinguals*. National Institute on Deafness and Other Communication Disorders.

Peña, E. D., Bedore, L. M., Gutiérrez-Clellen, V. F., Iglesias, A., & Goldstein, B. A. (2006). *Bilingual English–Spanish Oral Screener (BESOS)*. Unpublished research version.

Peña, E. D., Bedore, L. M., Gutiérrez-Clellen, V. F., Iglesias, A., & Goldstein, B. A. (2008). *Bilingual English–Spanish Assessment—Middle Extension Experimental Test Version (BESA-ME)*. Unpublished research version.

Peña, E. D., Bedore, L. M., Gutiérrez-Clellen, V. F., Iglesias, A., & Goldstein, B. A. (2016). *Bilingual English–Spanish Assessment—Middle Extension Field Test Version (BESA-ME)*. Unpublished research version.

Peña, E. D., Bedore, L. M., & Kester, E. S. (2015). Discriminant accuracy of a semantics measure with Latino English-speaking, Spanish-speaking, and English–Spanish bilingual children. *Journal of Communication Disorders, 53,* 30–41. https://doi.org/10.1016/j.jcomdis.2014.11.001

Peña, E. D., Bedore, L. M., & Kester, E. S. (2016). Assessment of language impairment in bilingual children using semantic tasks: Two languages classify better than one. *International Journal of Language & Communication Disorders, 51*(2), 192–202. https://doi.org/10.1111/1460-6984.12199

Peña, E. D., 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*(1), 5–16. https://doi.org/10.1044/0161-1461(2003/001)

Peña, E. D., Bedore, L. M., & Zlatic-Giunta, R. (2002). Category-generation performance of bilingual children: The influence of condition, category, and language. *Journal of Speech, Language, and Hearing Research, 45*(5), 938–947. https://doi.org/10.1044/1092-4388(2002/076)

Peña, E. D., Gillam, R. B., & Bedore, L. M. (2014). Dynamic assessment of narrative ability in English accurately identifies language impairment in English language learners. *Journal of Speech, Language, and Hearing Research, 57*(6), 2208–2220. https://doi.org/10.1044/2014_JSLHR-L-13-0151

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*(4), 302–314. https://doi.org/10.1044/1058-0360(2011/10-0020)

Peña, E. D., Gutiérrez-Clellen, V., Iglesias, A., Goldstein, B., & Bedore, L. M. (2018). *BESA: Bilingual English–Spanish Assessment manual*. Brookes.

Rice, M. L., & Wexler, K. (1996). Toward tense as a clinical marker of specific language impairment in English-speaking children. *Journal of Speech and Hearing Research, 39*(6), 1239–1257. https://doi.org/10.1044/jshr.3906.1239

Richardson, J. T. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review, 6*(2), 135–147. https://doi.org/10.1016/j.edurev.2010.12.001

Scarr, S., & Weinberg, R. A. (1978). The influence of "family background" on intellectual attainment. *American Sociological Review, 43*(5), 674–692. https://doi.org/10.2307/2094543

Schulz, P., & Roeper, T. (2011). Acquisition of exhaustivity in *wh*-questions: A semantic dimension of SLI? *Lingua, 121*(3), 383–407. https://doi.org/10.1016/j.lingua.2010.10.005

Sheng, L., Bedore, L. M., Peña, E. D., & Fiestas, C. (2013). Semantic development in Spanish–English bilingual children: Effects of age and language experience. *Child Development, 84*(3), 1034–1045. https://doi.org/10.1111/cdev.12015

Sheng, L., & McGregor, K. K. (2010). Lexical–semantic organization in children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 53*(1), 146–159. https://doi.org/10.1044/1092-4388(2009/08-0160)

Sheng, L., McGregor, K. K., & Marian, V. (2006). Lexical–semantic organization in bilingual children: Evidence from a repeated word association task. *Journal of Speech, Language, and Hearing Research, 49*(3), 572–587. https://doi.org/10.1044/1092-4388(2006/041)

Sheng, L., Peña, E. D., Bedore, L. M., & Fiestas, C. E. (2012). Semantic deficits in Spanish–English bilingual children with language impairment. *Journal of Speech, Language, and Hearing Research, 55*(1), 1–15. https://doi.org/10.1044/1092-4388(2011/10-0254)

Shivabasappa, P., Peña, E. D., & Bedore, L. M. (2018). Core vocabulary in the narratives of bilingual children with and without language impairment. *International Journal of Speech-Language Pathology, 20*(7), 790–801. https://doi.org/10.1080/17549507.2017.1374462

Spaulding, T. J., Hosmer, S., & Schechtman, C. (2013). Investigating the interchangeability and diagnostic utility of the PPVT-III and PPVT-IV for children with and without SLI. *International Journal of Speech-Language Pathology, 15*(5), 453–462. https://doi.org/10.3109/17549507.2012.762042

Stevens, J. P. (2012). *Applied multivariate statistics for the social sciences*. Routledge.

Sullivan, A., & Bal, A. (2013). Disproportionality in special education: Effects of individual and school variables on disability risk. *Exceptional Children, 79*(4), 475–494. https://doi.org/10.1177/001440291307900406

Thordardottir, E. (2011). The relationship between bilingual exposure and vocabulary development. *International Journal of Bilingualism, 15*(4), 426–445. https://doi.org/10.1177/1367006911403202

Tomblin, J. B., Records, N. L., 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*(6), 1245–1260. https://doi.org/10.1044/jslhr.4006.1245

Yip, V., & Matthews, S. (2005). Dual input and learnability: Null objects in Cantonese–English bilingual children. In J. Cohen, K. T. McAlister, K. Rolstad, & J. MacSwan (Eds.), *Proceedings of the 4th International Symposium on Bilingualism* [Symposium] (pp. 2421–2431). Cascadilla Press.