Research Article

Performance of Low-Income Dual Language Learners Attending English-Only Schools on the Clinical Evaluation of Language Fundamentals–Fourth Edition, Spanish

Beatriz Barragan,[a] Anny Castilla-Earls,[b] Lourdes Martinez-Nieto,[a] M. Adelaida Restrepo,[a] and Shelley Gray[a]

Purpose: The aim of this study was to examine the performance of a group of Spanish-speaking, dual language learners (DLLs) who were attending English-only schools and came from low-income and low-parental education backgrounds on the Clinical Evaluation of Language Fundamentals–Fourth Edition, Spanish (CELF-4S; Semel, Wiig, & Secord, 2006).

Method: Spanish-speaking DLLs (*N* = 656), ages 5;0 (years; months) to 7;11, were tested for language impairment (LI) using the core language score of the CELF-4S and the English Structured Photographic Expressive Language Test (Dawson, Stout, & Eyer, 2003). A subsample (*n* = 299) was additionally tested using a Spanish language sample analysis and a newly developed Spanish morphosyntactic measure, for identification of children with LI and to conduct a receiver operating characteristics curve analysis.

Results: Over 50% of the sample scored more than 1 *SD* below the mean on the core language score. In our subsample, the sensitivity of the CELF-4S was 94%, and specificity was 65%, using a cutoff score of 85 as suggested in the manual. Using an empirically derived cutoff score of 78, the sensitivity was 86%, and the specificity was 80%.

Conclusions: Results suggest that the CELF-4S overidentifies low-income Spanish–English DLLs attending English-only schools as presenting with LI. For this sample, 1 in every 3 Latino children from low socioeconomic status was incorrectly identified with LI. Clinicians should be cautious when using the CELF-4S to evaluate low-income Spanish–English DLLs and ensure that they have converging evidence before making diagnostic decisions.

I n the current study, we examined the performance of a group of low-income, low-parental education, Spanish-speaking children attending English-only school programs on the Clinical Evaluation of Language Fundamentals–Fourth Edition, Spanish (CELF-4S; Semel, Wiig, & Secord, 2006). Although significant advances have been made in the last 20 years to establish appropriate assessment practices for Spanish-speaking dual language learners (DLLs; Bedore & Leonard, 2001, 2005; Bedore & Peña, 2008; Bedore et al., 2012; Gutiérrez-Clellen, Restrepo, & Simón-Cereijido, 2006; Gutiérrez-Clellen & Simon-Cereijido, 2007; Restrepo, 1998), we still have limited information regarding the validity of available standardized language tests designed to identify Spanish-speaking children with language disorders (Dollaghan & Horner, 2011). The CELF-4S is the most widely used standardized test among clinicians and researchers who work with school-age Spanish–English DLLs in the United States (Arias & Friberg, 2016; Crowley, 2010). However, independent researchers have not yet examined the validity of this measure with populations with multiple risk factors, such as low socioeconomic status (SES), low parental education, and subtractive language environments (i.e., social environment that favors the acquisition of the dominant language, while slowing or reversing the development of the native language [Wright, Taylor, & Macarthur, 2000]). Because speech-language pathologists rely on accurate standardized measures to identify children with language

[a]Department of Speech and Hearing Science, Arizona State University, Tempe
[b]Communication Sciences and Disorders, University of Houston, TX
Correspondence to Beatriz Barragan: beatriz.barragan@asu.edu

impairment (LI; Caesar & Kohler, 2007; Crowley, 2010; Huang, Hopkins, & Nippold, 1997), it is important to investigate the performance of Spanish-speaking DLLs on the CELF-4S.

Hispanic children are more likely to live in poverty than non-Hispanic mainstream children; one in every five Hispanic children in the United States lives in poverty in comparison to one in every 10 non-Hispanic mainstream children (DeNavas-Walt & Proctor, 2015). This is particularly important for this investigation given that SES is associated with low language skills due to reduced quantity and quality of input (e.g., Chodrogianni & Marinis, 2011; Hart & Risley, 1992; Hoff, 2003). Therefore, children from low SES backgrounds typically perform worse than expected on standardized language measures (e.g., Gilliam & de Mesquita, 2000; Qi, Kaiser, Milan, Yzquierdo, & Hancock, 2003; Washington & Craig, 1999), which confounds language interpretations for children with LI. Low scores on standardized language tests could be the result of an LI profile, low parental education, or a combination of risk factors.

Spanish-speaking DLLs with LI are often at risk of low performance in reading comprehension and academic achievement and are disproportionally represented in special education (e.g., Morgan et al., 2015; Samson & Lesaux, 2009). For clinicians, identifying the presence of LI in DLLs is a challenge; overidentification and underidentification of LI are frequent. Overidentification is often due to the presence of language characteristics in typically developing bilingual children that overlap with those frequently observed in monolingual children with LI (Anderson & Souto, 2005; Castilla-Earls et al., 2016; Morgan, Restrepo, & Auza, 2013; Paradis, 2010a). Underidentification is often due to concerns that language difficulties are secondary to second language (L2) acquisition and not LI (Morgan et al., 2015; Paradis, 2010a; Samson & Lesaux, 2009). These patterns of misdiagnosis stem from variability in language characteristics in the bilingual population, similarities between L2 acquisition skills and LI, and bilingual effects on grammatical abilities as children learn an L2 (Castilla-Earls et al., 2016; Morgan et al., 2013). DLLs in subtractive language environments typically demonstrate use of ungrammatical structures in their first language (L1) for a longer period of time when compared with monolinguals and show cross-linguistic influences for extended periods during development (Montrul, 2008; Morgan et al., 2013; Restrepo & Gutiérrez-Clellen, 2001). For example, in Spanish, they may present more gender agreement errors in articles and clitic pronouns (e.g., Anderson, 2001; Castilla-Earls et al., 2016; Morgan et al., 2013; Restrepo, 1998). Children under this environmental circumstance can present plateaus in the development of the L1 or first-language loss (Kohnert, 2010; Restrepo, 2003; Restrepo et al., 2010).

Language Assessment in DLLs

To evaluate children's language development, clinicians and researchers rely on two important tools: standardized language measures and language sample analyses (Peña, Bedore, & Kester, 2016). Although other tools and methods, such as dynamic assessment, are highly recommended, they are still not widely used or validated for the population at large. Standardized measures and language sample analyses allow clinicians to assess several language skills like lexical knowledge, morphosyntactic structures, semantic knowledge, and pragmatics.

Language Sample Analyses

Language sample analyses provide indexes of vocabulary, fluency, and grammatical skills in children's language abilities in everyday informal settings. Some of these measures have been identified as sensitive to LI in monolingual English (Dunn, Flax, Sliwinski, & Aram, 1996; Heilmann, Miller, Nockerts, & Dunaway, 2010), monolingual Spanish (Anderson & Souto, 2005), and bilingual Spanish–English-speaking children (Bedore, Peña, Gillam, & Ho, 2010; Muñoz, Gillam, Peña, & Gulley-Faehnle, 2003; Restrepo, 1998; Simon-Cereijido & Gutiérrez-Clellen, 2007). The most sensitive measures within language sample analyses to identify Spanish–English bilingual children with LI are mean length of utterance (MLU; Restrepo, 1998), number of different words (NDW; Bedore & Leonard, 1998; Uccelli & Páez, 2007), and Ungrammaticality Index (UGI; Macswan & Rolstad, 2006; Restrepo, 1998; Simon-Cereijido & Gutiérrez-Clellen, 2007).

Researchers have used language sample analyses to reduce bias in language assessment in bilingual children (Bedore et al., 2010; Gutiérrez-Clellen, Restrepo, Bedore, Peña, & Anderson, 2000; Gutiérrez-Clellen & Simon-Cereijido, 2009; Restrepo, 1998), and it is still considered the gold standard for evaluating productive language use and identifying LI in children who speak one or more languages (Heilmann et al., 2010). For example, Restrepo (1998) found that the UGI from a Spanish language sample analysis in combination with parent report was more efficient than commonly used standardized measures to identify LI in Spanish-speaking DLLs. Simon-Cereijido and Gutiérrez-Clellen (2007) also examined language samples in Spanish–English bilingual children, finding that a combination of measures, including MLU and ungrammaticality, have fair sensitivity (79%) and very good specificity (100%) in the identification of preschool children with LI.

Standardized Measures

Standardized measures are widely used for the identification of LI in children (Crowley, 2010; Huang et al., 1997). Speech-language pathologists use standardized language measures for a variety of reasons, including that they offer norm-referenced comparisons, they are easy to administer and interpret, they have less examiner's bias effects, and they are required, in some cases, for determination of qualification for speech and language services in the schools (Fulcher-Rood, Castilla-Earls, & Higginbotham, 2015). Huang et al. (1997), for example, reported that 81% of clinicians use standardized measures for identifying the presence of LI; however, the validity of existing standardized measures for Spanish–English bilingual children is a concern

because these measures do not facilitate the sampling of natural language nor do they provide an accurate representation of a bilingual child's language skills (Anderson, 1996; Restrepo, 1998; Restrepo & Silverman, 2001). Moreover, validation studies of these measures are limited.

Only a handful of standardized measures are available to identify language disorders in Spanish–English DLLs, and fewer are developed to target Spanish language structures that are sensitive to LI in Spanish. Independent studies examining the validity of measures for Spanish-speaking DLLs are difficult to find. For example, the Bilingual English-Spanish Assessment (Bedore et al., 2010; Peña, Gutiérrez-Clellen, Iglesias, Goldstein, & Bedore, 2014) is a measure developed for Spanish–English DLLs and is normed with DLLs in the United States. This measure is currently limited to 4- to 6-year-olds and, thus, not available for older children. The CELF-4S (Semel et al., 2006) is a frequently used measure (Arias & Friberg, 2016; Crowley, 2010) developed for school-age children; however, it is based on an English model of LI.

Normative samples in standardized measures often do not reflect differences in language proficiency, and thus, measures may over identify or underidentify. For example, Morgan et al. (2009, 2013) found that the CELF-4S underidentified monolingual Spanish speakers with LI. The cutoff score had to be adjusted up 1 SD in order to capture more accurately the difference between monolingual children with LI and those with typical development. On the other hand, Restrepo and Silverman (2001) found that the Preschool Language Scale–Third Edition (Zimmerman, Steiner, & Pond, 1992) overidentified bilingual children as presenting with LI. It is also possible that children with varying language proficiencies, language use contexts, risk factors, and dialects may score differently on standardized measures (Restrepo & Silverman, 2001).

Factors That Influence Performance on Standardized Language Testing

Some of the factors that might impact performance on standardized measures include the language of instruction and proficiency in the language being assessed. McCauley and Swisher (1984) recommended separate norms to be available when there are differences in performance according to variations in linguistic or demographic characteristics. Similarly, Restrepo and Silverman (2001) argued that separate norms may be needed when there are linguistic differences, such as monolingual versus bilingual groups in the normative sample. Given the possible overlap between bilingual children in the lower end of the normal distribution of proficiency and those monolinguals with language disorders, the interpretability of unified norms is questionable (Morgan et al., 2013; Paradis, 2005, 2010b).

Language Proficiency

Language proficiency levels in the L2 affect performance on L2 language tasks, such as standardized measures (Pearson, 2007; Restrepo, 1998; Restrepo & Silverman, 2001). Language proficiency involves acquisition of knowledge and efficient use of language (Kohnert & Bates, 2002) on a continuum from nonproficient to native-like proficiency. Assessment of language proficiency in young sequential bilinguals is challenging because these children are exposed to an L2 while they are still developing their home language. Therefore, it is crucial to distinguish between language difficulties caused by low proficiency levels due to insufficient exposure to a particular language and language difficulties due to LI. Low language proficiency levels are linked with longer language processing time and lower scores in grammar and vocabulary (Chee, Hon, Lee, & Soon, 2001). Therefore, the performance of typically developing bilingual children in a language in which they have not attained native-like proficiency may resemble the performance of monolingual children with LI (Paradis, 2010a; Paradis & Crago, 2000). For instance, Abutalebi (2008) found that bilinguals show less elaborated linguistic comprehension in the less proficient language during sentence processing tasks.

As children become more proficient in their L2, their home language skills may decrease, plateau, or develop more slowly if academic and/or social support for the home language is limited (Castilla-Earls et al., 2016; Montrul, 2011; Morgan et al., 2013; Restrepo et al., 2010). A decrease in the home language can explain differences between children who live in or attend schools in additive language contexts and those in subtractive language contexts where the home language is supported in a limited capacity. However, measures disaggregating performance on the basis of language proficiency are not available, as far as we know.

Language of Instruction and Language Performance

The type of education that children receive at school can affect the development of language skills, such as vocabulary and grammar. English-only education in the United States focuses on the development of English, and thus, all school activities are conducted in English, whereas the children's home language use is discouraged. For example, in Arizona, bilingual programs that support bilingual development are restricted to children who demonstrate high levels of English proficiency, and thus, children learning English as an L2 typically have no access to home language instruction at school (Arizona Department of Education, 2000).

Research on bilingual versus English-only education indicates that bilingual programs for Spanish–English speakers improve Spanish skills without detrimental or negative effects on English development (e.g., Mahoney, Thompson, & MacSwan, 2005; Restrepo, Morgan, & Thompson, 2013). Pearson (2007) found that bilingual children in bilingual programs scored similarly to those in English-only programs on English language grammar and vocabulary standardized measures. In contrast, those in English-only programs decreased significantly in Spanish grammar and vocabulary measures over time relative to age norms; children in bilingual programs gained significantly in Spanish standardized measures compared with those in English-only programs. These studies suggest that the language of education will

impact performance on standardized language measures of vocabulary, grammar, and fluency, making the assessment process more complex.

The CELF-4S

The CELF-4S is a frequently used language assessment measure (Arias & Friberg, 2016; Crowley, 2010) designed to identify LI in Spanish-speaking children. According to the authors, this test is not a translated version of the English Clinical Evaluation of Language Fundamentals–Fourth Edition (Semel, Wiig, & Secord, 2003) but rather a parallel version because it was developed to represent Spanish morphosyntactic rules (Semel et al., 2006). The normative sample included 1,019 Spanish-speaking students (both monolingual and bilingual) in the United States and Puerto Rico (Semel et al., 2006). Semel et al. (2006) found no differences on the performance on the CELF-4S between Spanish monolingual and Spanish–English bilingual children and determined that separate norms for bilingual and monolingual children were not necessary. This is important because their finding contrasts with many researchers who have advocated for separate norms for DLLs (e.g., Bedore & Peña, 2008; Restrepo & Silverman, 2001; Morgan et al., 2013).

The CELF-4S normative data are reported in 6-month intervals from age 5;0 (years;months) to 6;11 and in 1-year intervals from 7;0 to 16;11. The normative sample size for ages 5;0 to 6;11 was 70 per 6-month interval and, for ages 7;0 to 7;11, was 50, which is below the recommended sample size for test standardization. Norms derived from small sample sizes are likely to be less reliable and stable; therefore, data from a different group of children might result in different norms (McCauley & Swisher, 1984). Parental education was used as a proxy for SES in the standardization process. The normative sample was distributed as follows: 37.88% of the sample completed 11th grade or less; 31.51% completed 12 years of school; 19.01% completed 13 to 15 years of school; and 11.51% completed more than 16 years of school. Normative tables are not disaggregated by group differences (i.e., family income and language of instruction), and thus, the normative sample may not be representative of children from low-income families, attending English-only programs.

The sensitivity reported in the manuals is 0.96, 0.86, and 0.52 for −1.0, −1.5, and −2.0 SDs below the mean, respectively, which are considered *good* to *unacceptable* according to standards in the field (Plante & Vance, 1994). The specificity is 0.87, 0.95, and 1.00 for −1.0, −1.5, and −2.0 SDs below the mean, respectively, which are considered *fair* to *good*. The test–retest reliability coefficients ranged from .52 to .93 across ages (Semel et al., 2006). These results were based on an analysis of children with language disorders, who comprised 6.2% of the normative sample (Semel et al., 2006, p. 87). These children were tested by speech-language pathologists and scored 1.5 SD or more below the mean on an unidentified standardized test of language ability (Semel et al., 2006, p. 91). It is important to note that the validity of the specificity and sensitivity estimates are questionable because the reference measure used is not reported. High-quality studies to establish the diagnostic accuracy of a test include a reference measure that is considered to be the "gold standard" for identification (Dollaghan, 2007).

The purpose of the current study was to examine performance on the CELF-4S in a group of DLLs who were attending English-only schools and came from low-income and low-parental education backgrounds. Language sample analyses and performance on two additional grammar measures were used to examine diagnostic accuracy. We evaluated the following questions:

1.   How does a low-income, low-parental education Spanish–English DLL sample attending English-only education in the United States perform on the CELF-4S?

2.   Does the CELF-4S accurately differentiate between low-income, Spanish-dominant DLLs attending English-only education with and without LI?

Method

Participants

The participants in this study were part of a larger project focused on the development of a language screener for Spanish–English bilingual children at risk for LI (Spanish Screener for Language Impairment in Children [SSLIC]; Restrepo, Gorin, & Gray, 2013). A group of 656 Latino children between age 5;0 and 7;11, attending kindergarten through second grade, were included in the study. Children were attending English-only education programs at public and charter schools in the greater Phoenix metropolitan area in Arizona. The children were not preselected on the basis of risk, special education, or individualized education program criteria. Income was determined by the qualification for free and reduced lunch program at school. We reported parental education separately. Table 1 describes the participants' demographic information by age, parental education, and SES in comparison to the CELF-4S normative sample.

To participate in this study all children met the following criteria: (a) they spoke Spanish at home at least 50% of the time on the basis of parent report; (b) their teacher identified them as English language learners on a teacher questionnaire; (c) they passed a hearing screening; (d) they did not demonstrate significant sensory or cognitive disabilities per parent or teacher report; and (e) they scored as Spanish proficient or Spanish dominant compared with English, on the Spanish–English Language Proficiency Scale (SELPS; Smyk, Restrepo, Gorin, & Gray, 2013). Children who scored higher in English than in Spanish on the SELPS were excluded from the sample to ensure that only Spanish-dominant children were included in the study.

All children previously described were included in the analysis to examine the performance of the CELF-4S in a low-income, low-parental education Spanish–English

Table 1. Participants' demographic information compared with CELF-4S normative sample demographics.

Demographics	Study sample		CELF-4S normative sample	
	n			*n*
Age (years;months)				
5;0–5;5	73			70
5;6–5;11	115			70
6;0–6;5	162			70
6;6–6;11	94			70
7;0–7;11	212			50
Total	656			330
Mother education level				
Elementary school or less	260	11th grade or less		125
12 years of school	320	12 years of school		104
College degree	50	13 + years of school		101
NR	26			NR
Lunch program				
Free lunch	592			NR
Reduced price lunch	13			NR
Full price lunch	6			NR
NR	45			NR

Note. CELF-4S = Clinical Evaluation of Language Fundamentals–Fourth Edition, Spanish; NR = not reported.

DLLs attending English-only education. A convenience subsample of 299 DLLs was selected from the participant pool to further examine the accuracy of the CELF-4S for the identification of low-income DLLs with and without LI. The convenience subsample was selected based on the availability of language transcriptions for this large group of children and the selection criteria for determination of language ability described below.

Determination of Language Ability

The current best practice recommendation to estimate language ability in bilingual children is to conduct assessment in both languages and to use converging evidence for the determination of bilingual LI (Kohnert, 2010; Peña et al., 2016; Restrepo, 1998). A subsample of 299 children were classified in two language ability groups, using converging evidence that included language sample analysis in Spanish and standardized test performance in both Spanish and English. All assessment tools used in this study are further described in the Measures section. Children with typical language development met the following criteria: (a) scored higher than −1 *SD* from the mean on the SSLIC morphosyntactic task and (b) met two out of three of the following criteria on the Spanish language sample analysis: (i) scored above −1 *SD* from the mean on MLU per T-units on the basis of the Systematic Analysis of Language Transcripts (SALT; Miller & Iglesias, 2008) database, (ii) scored above −1 *SD* from the mean on NDW in the SALT database, and (iii) scored less than 20% on the UGI per total number of utterances (Restrepo, 1998).

Children with LI met the following criteria: (a) scored below −1 *SD* from the mean in the SSLIC Morphosyntactic

task; (b) met two out of three of the following criteria on the Spanish language sample analysis: (i) scored below −1 *SD* from the mean on MLU per T-units on the basis of the SALT database, (ii) scored below −1 *SD* from the mean on NDW on the SALT database, and (iii) scored more than 20% on the UGI per total number of utterances (Restrepo, 1998); and (c) scored below 75 on the Structured Photographic Expressive Language Test–Third Edition (SPELT-3) in English (Dawson et al., 2003). Using this criteria, 28 children were found to present LI (9.6%), whereas 265 children presented with typically developing language (90.4%). Six children were eliminated from the subsample because they met the LI criteria for the Spanish Morphosyntactic task and the Spanish language sample but scored above 75 on the SPELT-3, which may indicate English-language dominance and not LI (Table 2 for details).

Measures

Core Language Score–CELF-4S

The core language score is a measure of general language ability used to make clinical decisions about the presence or absence of language disorders and to establish the need for special education services (Rhein, 2013). The mean-scaled core language score is 100, and the *SD* is 15. The CELF-4S manual suggests that the core language be used to identify children as LI using a cut score of 85 (1 *SD* below the mean). The core language score is derived from the scale scores of four subtests: Concepts and Following Directions, Word Structure, Recalling Sentences, and Formulating Sentences (Semel et al., 2006). The means and *SD*s for the subtests are 10 and 3, respectively.

In the Concepts and Following Directions subtest, children point to pictured objects following the examiner's oral directions. The test is designed to evaluate children's ability to understand spoken directions of increasing complexity and the ability to remember names of objects. In the Word Structure subtest, children complete sentences probing a grammatical structure. This subtest evaluates children's morphological knowledge by examining children's ability to mark inflections, derivations, and comparisons and the ability to use appropriate pronouns to refer to people, objects, and possessive relations. In the Recalling Sentences subtest, children repeat sentences. The subtest evaluates children's ability to listen to spoken sentences of different length and complexity and repeat the sentences without changing word meanings, inflections, derivations or comparisons (morphology), or sentence structure (syntax). In the Formulated Sentences subtest, children formulate a sentence on the basis of a visual stimulus and a target word or phrase that the research assistant (RA) gives to the child. The subtest evaluates children's ability to create sentences using prescribed words and a visual stimulus.

SSLIC Morphosyntactic Task

The SSLIC morphosyntactic task is a measure designed to evaluate morphological skills found to be deficient in Spanish-speaking children with LI. Target structures include

Table 2. Performance of subsample of children with TD and LI on standardized measures and language sample analyses.

	TD				LI			
	n = 265				n = 28			
			Range				Range	
Measures	*M*	*SD*	Min	Max	*M*	*SD*	Min	Max
SSLIC-Morph.	15.15	3.32	9	23	6.57	3.58	1	12
MLU	6.62	0.86	4.81	9.49	5.51	1.43	3.68	9.39
NDW	75.57	17.23	33	145	51.86	11.72	23	69
UGI	12.3	11.8	0	81.5	29.34	16.32	4.3	85.7
SPELT-3	66.59	19.87	7	112	48.89	16.41	19	71
CELF-4S	89.49	13.19	59	130	67.21	11.94	47	95

Note. TD = typical development; LI = language impairment; SSLIC-Morph = Spanish Screener for Language Impairment in Children morphosyntactic task; MLU = mean length of utterances; NDW = number of different words; UGI = Ungrammaticality Index; SPELT-3 = Structured Photographic Expressive Language Test–Third Edition; CELF-4S = Clinical Evaluation of Language Fundamentals–Fourth Edition, Spanish.

articles, direct object clitic pronouns, prepositions, subjunctive, and derivational morphemes (e.g., Morgan et al., 2009, 2013). Children look at colored pictures and complete a sentence or respond to questions that elicit the target grammatical structures. For example, for clitic pronouns: "¿Qué hace el perro con los regalos? Los lame" (*What does the dog do with the presents*? *It licks them*). Each item was scored as correct or incorrect. The total score on this task is the sum of correct items. This measure differentiates children with typical language and LI with a specificity of 74% and a sensitivity of 98% for age 5 years; 85% and 98% for age 6 years; and 75% and 96% for age 7 years (Restrepo, Gorin, & Gray, 2013).

Language Sample Analyses

A story-retelling task was used to collect a language sample. An RA read a Spanish script of the wordless book *Frog on His Own* (Mayer, 1973) to each child. The script assured a consistent narrative across all participants. The RA asked the children to retell the story back. Language samples were recorded and later transcribed and coded using SALT (Miller & Iglesias, 2008). The following measures were obtained from the language samples:

MLU

Following Gutiérrez-Clellen and Hofstetter's (1994) adaptation to Spanish of Hunt's (1965) terminable unit (TU) procedure, the RA segmented the language samples into TUs. A TU consists of a main clause and all its subordinated clauses, for example, "El niño lloró cuando la rana saltó" (*The boy cried when the frog jumped*) represents one TU, whereas "La rana saltó y cayó en el barco" (*The frog jumped and fell into the boat*) represents two TUs (Spanish is a pro-drop language, and thus, sentences may not contain explicit subjects). The definition of a *clause* was taken from Berman and Slobin (1986) who stated: "A clause is a unit with a unified predicate, which expresses a unique situation"

(p. 37). We used SALT (Miller & Iglesias, 2008) to calculate the MLU in words.

NDW

Spanish word forms were linked to their morphological roots to avoid overestimating the NDW, for example, "llevó" and "llevaron" (*[he] took, [they] took*) were linked to "llevar" (*take*) and, therefore, were considered as one. SALT (Miller & Iglesias, 2008) estimates the total number of different root words per sample.

UGI

Each sentence was reviewed for grammatical errors. Any sentence with grammatical errors was coded as ungrammatical. UGI was computed by dividing the total number of ungrammatical TUs by the total number of TUs. Restrepo (1998) found that a grammaticality index combined with parent report provided high sensitivity and specificity in identifying Spanish–English bilingual children with LI. In addition, Bedore et al. (2010) found that grammaticality is a good predictor of language status in kindergarten bilinguals. Grammaticality is scored based on errors that impact morphology and syntax. Lexical choice, semantic, and phonological errors were not included (Restrepo, 1998). Few instances of code switching were found in the language samples analyzed for this study, probably due to the fact that these children were Spanish dominant. Most of the code switching that occurred in the samples took the form of borrowed words from English. Utterances were not marked as ungrammatical due to the presence of code switching. The few utterances that were completely in English were excluded from the analysis.

One investigator double-checked 100% of the transcripts and coding. Then, different raters transcribed and coded 12% of the samples independently. Interrater reliability was estimated in 97% for TUs, 86% for grammatical errors, and 93% for NDW. In addition, results from the analyses

were compared with SALT database. The SALT software (Miller & Iglesias, 2008) manages the process of eliciting, transcribing, and analyzing language samples. In order to help clinicians and researchers, the authors generated several databases. For the purpose of this study, we selected the Bilingual Spanish/English Story Retell Database. We chose this database because narratives were obtained using the same elicitation procedure and the same wordless picture books (Mayer, 1973). Participants for the database were 4,667 typically developing Spanish–English bilingual children ages 5 to 9;9 who attended public schools in Texas and California. Each narrative is associated with the age, grade, and gender of the participant. We narrowed the database to children 5 to 7 years old to match the age of our participants.

SELPS

The SELPS measures the level of oral language proficiency in Spanish and English in bilingual children 5 to 8 years old. The test uses a story retell task to elicit a language sample, which is scored on four domains: syntactic complexity, grammatical accuracy, verbal fluency, and lexical diversity. Each domain is scored between 1 and 5 according to how well the child can speak the target language. The SELPS total score is the sum of scores obtained in the four domains. The English measure has strong correlations with story retells for syntax (.53, $p < .001$), grammar (−.63, $p < .001$), lexical diversity (.50, $p < .001$), and fluency (−.36, $p < .002$; Smyk et al., 2013). Similarly, the Spanish version was found to have strong correlations between the SELPS total scores and syntax ($r = .695$, $p < .001$), grammar ($r = −.705$, $p < .001$), and lexical diversity ($r = −.504$, $p < .014$), indicating that the SELPS rating measure is valid in comparison to objective language sample measures (Tavizón, 2014).

SPELT-3

The SPELT-3 is a norm-referenced measure used to examine morphological and syntactic structures in children's expressive English. It includes 54 photographs of everyday situations and objects, paired with elicitation questions that allow the analysis of a variety of morphosyntactic structures. Although this test is not intended to diagnose LI in DLLs, a standard score of 75 was used as cutoff score in this study to verify that children with LI showed language deficits in both Spanish and English (e.g., children with low Spanish language skills and standard scores below 75 on the SPELT-3) and to rule out the possibility of typical developing children undergoing a language dominance switch from Spanish to English (e.g., children with low Spanish language skills but standard scores above 75 on the SPELT-3). Because the SPELT-3 does not have sensitivity and specificity data available for DLLs, we chose a score of 75 on the basis of visual inspection of the data: Those children with 75 or greater on the SPELT and LI criteria in Spanish were eliminated because their qualification on either group was not clear. This approach eliminated six children from the group with LI.

Procedure

Teachers distributed parent consent forms among Latino children in their classrooms. Parents who authorized their child's participation filled out a questionnaire related to language use at home (percentage of Spanish and English language use during a regular day) and characteristics of the child's language performance (concerns related to the child's comprehension and production in both languages). Teachers also completed a questionnaire about language performance in the classroom for each participant. Examiners were native Spanish-speaking RAs, specifically trained for testing the children individually in a separate classroom at the school. Three testing sessions lasting up to 45 min each were required for each participant. RAs tested participants using the core language section of the CELF-4S, the SSLIC morphosyntactic measure, language sample analysis, the SELPS, and the SPELT. RAs administered the tests in a fixed sequence to distribute evenly task demands and effort through the sessions. The first session was in English to screen children for hearing or cognitive deficits or English-only skills, and then, two Spanish sessions followed. Language samples were recorded and later transcribed and analyzed using SALT (Miller & Iglesias, 2008).

Results

General Performance on the CELF-4S

The first question we addressed was how a low-income, low-parental education Spanish–English DLL sample attending English-only education in the United States performed on the CELF-4S. We compared our sample's scaled scores by age to the normative sample means of the CELF-4S (see Table 3 for the study's descriptive data compared to the national norms). Our results suggest that the average core language score for this study's total sample ($N = 656$) was 83.57 ($SD = 14.97$), which is more than 1 SD below the mean of the CELF-4S normative sample ($M = 100$, $SD = 15$). Furthermore, 53.5% of the children in this study scored below 85 (−1 SD) on the CELF-4S core language scale. These findings suggest that more than half of the low-income Spanish-speaking DLLs attending English-only schools included in this study could be misidentified as presenting with LI if the results of the CELF-4S core language score were used as the main classification criteria using −1.0 SD. When we use −1.5 SD, 33% of the children scored as LI.

To further explore the discrepancy in our sample against the normative sample of the CELF-4S, we compared the results by subtests and by age group. We found that the average scores for our sample on all four subtests (Concepts and Following Directions, Word Structure, Recalling Sentences, and Formulated Sentences) were between 6.5 and 9.1, below the normative sample of the CELF-4S (SDs between 2.6 and 3; see Table 3). Concepts and Following Directions and Word Structure subtest mean scores were below 1 SD from the normative mean. However, Recalling

Table 3. Study sample and CELF-4S normative sample performance on CELF-4S subtests and core language scores.

	Study sample								CELF-4S normative sample	
	5-year-olds		6-year-olds		7-year-olds		Total sample		Total sample	
N	188		256		212		656		330	
Subtest	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
C&FD	6.66*	2.6	6.75*	2.8	6.69*	2.8	6.71*	2.7	10	3
RSc	11.23	5.6	17.34	7.9	22.54	8.4	17.29	8.7		
WS	6.50*	2.9	6.93*	2.8	6.62*	3.0	6.71*	2.9	10	3
RSc	11.71	5.1	14.50	5.2	16.77	5.7	14.43	5.7		
RS	8.16	2.7	7.46	2.8	7.10	2.7	7.54	2.8	10	3
RSc	15.91	13.1	23.67	15.9	29.32	18.6	23.28	16.9		
FS	9.10	2.7	8.53	2.8	8.40	2.7	8.60	2.8	10	3
RSc	9.12	6.2	14.07	7.9	19.58	8.3	14.44	8.6		
CLS	84.96*	13.8	83.61*	15.2	82.29*	15.6	83.57*	15.0	100	15

Note. CELF-4S = The Clinical Evaluation of Language Fundamentals–Fourth Edition, Spanish; C&FD = Concepts and Following Directions; RSc = raw score; WS = Word Structure; RS = Recalling Sentences; FS = Formulated Sentences; CLS = core language score.

*Scores below 1 *SD* of the normative sample mean.

Sentences and Formulated Sentences were within 1 *SD* of the normative mean.

A visual analysis of the data revealed a downward trend between age groups on the average core language scores. To further explore this trend, one-way analyses of variance for each subtest were conducted, showing significant differences on the performance between age groups for the Recalling Sentences, $F(2, 653) = 7.79$, $p < .01$, $\eta^2 = .02$, and Formulated Sentences, $F(2, 653) = 3.62$, $p = .027$, $\eta^2 = .01$, subtests. Results of Fisher least significant difference post hoc test revealed significant differences between 5- and 6-year-olds and 5- and 7-year-olds with means and *SD*s of 8.17 (2.7), 7.46 (2.8), and 7.10 (2.7), respectively, for Recalling Sentences and 9.10 (2.7), 8.52 (2.8), and 8.40 (2.7) for Formulated Sentences. No significant differences were found for the Concepts and Following Directions, $F(2, 653) = 0.076$, $p \geq .05$, and Word Structure, $F(2, 653) = 1.384$, $p \geq .05$, subtests (see Figure 1).

Discriminatory Accuracy of the CELF-4S

The second question examined whether the CELF-4S accurately discriminated between low-income Spanish–English DLLs attending English-only education with typical development from those with LI. The scores of a subsample of 299 children were analyzed using a receiver operating characteristics (ROC) curve analysis with the IBM SPSS statistics 23 software, to calculate estimates for sensitivity and specificity. An area under the ROC curve of 1.0 represents a perfect test, whereas a 0.5 area represents a clinically uninformative measure. The clinical accuracy of a test with an area under the ROC curve above 0.7 is considered fair, and above 0.8 is considered good (Tape, 2003). The area under the ROC curve for the CELF-4S was 0.89, 95% CI [.832, .955]; $p < .01$. Using the standard score of 85 as cutoff criterion, as suggested in the CELF-4S manual, the

sensitivity was estimated to be 93% and the specificity 65%. The positive likelihood ratio using the 85 cutoff score was estimated to be 2.39, and the negative likelihood ratio was 0.12. The guideline for likelihood ratios is that measures with a positive likelihood ratio equal or greater than 10 and a negative likelihood ratio equal or less than 0.1 are considered clinically informative for identification of LI (Dollaghan, 2007; Dollaghan & Horner, 2011; Sackett, Straus, Richardson, Rosenberg, & Haynes, 2000). In this sample, the positive likelihood ratio of the Spanish CELF-4S is considerably lower than 10 and, therefore, not clinically informative. The negative likelihood ratio is above 0.1, indicating that a negative result (i.e., a score above the cutoff criterion) is very likely to rule out LI.

Figure 1. Performance by age group on the Clinical Evaluation of Language Fundamentals–Fourth Edition, Spanish (CELF-4S) subtests. FD = Concepts and Following Directions; WS = Word Structure; RS = Recalling Sentences; FS = Formulated Sentences. *Significant differences at $p \leq .05$ level.
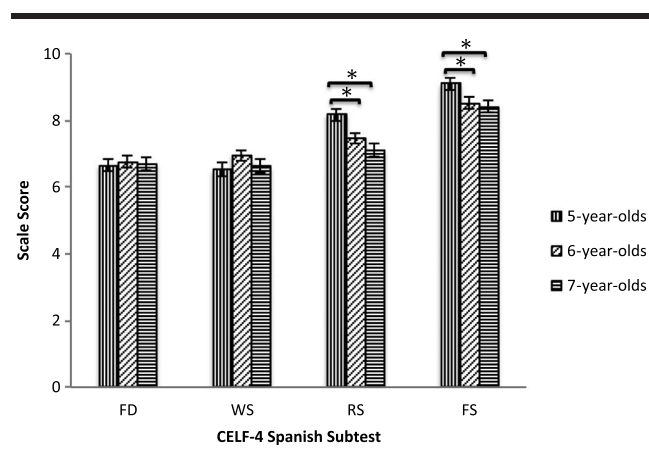
Further exploration of sensitivity and specificity using various cutoff scores revealed that it was possible to adjust the cutoff score to reflect an adequate balance between sensitivity and specificity. The best possible cutoff score to establish a balance between sensitivity and specificity was 78, with 9.6% of the children classified as LI. With this cutoff score, the CELF-4S had a sensitivity of 86% and a specificity of 80%. This sensitivity and specificity could be considered acceptable (Plante & Vance, 1994). The positive likelihood ratio using a score of 78 was estimated to be 4.37, which is considered mildly suggestive of a language disorder, and the negative likelihood ratio of 0.18 suggests that a score in the typical range is very likely to indicate no impairment.

Discussion

The purpose of the current study was to examine the performance on the CELF-4S of low-income, low-parental education Spanish–English DLL children attending English-only education and determine if the CELF-4S accurately differentiates between children with and without LI in this population. We assessed Spanish language ability in 656 children using the Spanish version of the CELF-4. Our results showed that these children from low-income homes, attending English-only education programs scored on average one standard deviation below the mean on the core language standard score. Using the CELF-4S manual recommended cut score of 85, 53.5% of our total sample scored within the LI range.

There are several differences between our sample and the CELF-4S normative sample that could potentially explain the lower results for the participants in the current study. The sample size in this study is larger than the sample used in the standardization of the CELF-4S at this age range. The CELF-4S sample included 1,019 Spanish-speaking students in total, with 330 of the participants corresponding to children between 5;0 and 7;11 of age. Our sample was considerably larger, including 656 participants in total with more children per age group (ranging from 73 in the 5;0 to 5;5 group, to 212 in the 7;0 to 7;11 group (see Table 1 for details). It is possible that the larger sample size for children between 5;0 and 7;11 of age in this study represented better the low-income, English-only education group compared with the CELF-4S normative sample.

The majority of children in our sample (97%) came from low-income and low-parental education homes, according to eligibility for free and reduced lunch program and parental report of their education attainment. In comparison, 37.8% of the CELF-4S sample included children from homes where parents did not complete high school (attended 11 years of school or less) and 31.5% from homes with parents who completed high school. In contrast with this study, the CELF-4S used parental education as their only measure of SES. The differences in standardized scores seen in this study are, in principle, not surprising given that this sample comes from a lower SES than the CELF-4S normative sample, and low SES has been associated with lower language

performance (e.g., Bohman, Bedore, Peña, Mendez-Perez, & Guillam, 2010; Hegde & Pomaville, 2013; Qi et al., 2003). However, the magnitude of the difference is of concern because the CELF-4S misidentified approximately half of the DLLs from low SES backgrounds as having LI if only one measure is used.

The results of this study indicate the need for separate norms for DLLs from low SES families who are receiving English-only education. Children from low SES families have lower vocabulary abilities when compared with children from higher SES background (Chondrogianni & Marinis, 2011; Dixon, Wu, & Daraghmeh, 2012; Golberg, Paradis, & Crago, 2008), and mothers with higher education contribute to a higher quality language input that fosters a child's language level (Bohman et al., 2010; Chondrogianni & Marinis, 2011). Considering that 5.7 million Latino children in the United States live in poverty, more than any other racial or ethnic group (Gamboa, 2015), norms that capture the performance of these children with multiple risk factors should be more accurate in identifying the language abilities of this population. Our results are not sufficient to generate a new set of norms because a local sample from Phoenix, AZ, metropolitan area does not represent Latino children in other geographical regions of the United States. Additionally, the range of ages in our sample was limited, and the linguistic context of the participants is specific to Arizona with English-only instruction in public schools. However, we provide guidance (see Table 3) to clinicians that are using the CELF-4S as a tool to identify Latino children with LI from low-income and low-parental education homes.

The CELF-4S reported that 12% of their samples were receiving bilingual education, but no further information is provided about educational setting. One hundred percent of our participants were attending English-only education programs, and in many of these schools, their native language is not allowed even for social purposes. These programs have a negative impact on Spanish language skills because U.S. English immersion programs do not support the use of the children's home language, and as a consequence, Spanish vocabulary and language development are protracted or undergo loss in comparison to children attending dual language programs (Barnett, Yarosz, Thomas, Jung, & Blanco, 2007; Morgan et al., 2013; Restrepo et al., 2010). These results indicate that the educational context may partially account for the significantly lower scores in our sample compared with the CELF-4S normative sample. However, it is not possible to establish a comparison with the normative sample because no further information about the type of bilingual educational context was provided and there is significant variability in bilingual instruction programs.

Interestingly, we found a decrease on the standard scores in two of the subtests with increasing age. Six- and 7-year-olds scored significantly lower than 5-year-olds on the scaled scores in the CELF-4S Recalling Sentences and Formulated Sentences subtests, which may suggest L1 loss or protracted development (Montrul, 2011). It is likely that the majority of the sample started preschool or kindergarten

as primarily Spanish speakers and have since being in English-only education. Therefore, 6- and 7-year-olds would have greater exposure to English than 5-year-olds. L1 loss is common in children who experience a shift in language exposure from the home language to the school language, in particular for English-only programs, and could explain the decrease in scores with age in Spanish. Further, the children in English-only education may demonstrate protracted development, especially in Spanish for academic purposes, and thus, gains seen in other groups may develop at a slower rate (Restrepo et al., 2010; Restrepo, Morgan, & Thompson, 2013).

The combination of risk factors in this study's sample (SES, parental education, and English-only education) place these children at risk for low academic achievement and low performance on the CELF-4S. It is possible that a single factor would not account for such significant difference in performance compared with the national sample, but the interaction of these risk factors may place these children at higher risk of academic difficulties that are not due to LI but rather due to the lack of exposure to language-rich environments for the home language. However, qualifying these children as presenting LI would have detrimental effects due to poor fit of services and students needs and the potential adverse effects of identification, such as lowered student expectations and stigma associated with a diagnosis (e.g., Alvarez-McHatton & Correra, 2005).

The purpose of the CELF-4S is to identify Spanish-speaking children with LI; however, this study found a general pattern of overidentification of LI with 53.5% of the total sample scoring below 1 *SD* from the mean, the CELF-4S author's recommended cut score. Further, we examined the accuracy of the CELF-4S to differentiate children with typical development and LI in a subsample of 299 children previously classified into ability groups. The CELF-4S ability to detect the disorder in those children, who have it, seems to be good for this sample on the basis of a sensitivity estimation of 93%, which is a potential strength of the test. However, our results indicated that the specificity (the proportion of children without the disorder that are correctly identified as such by the measure; Thordardottir et al., 2011) of the CELF-4S was 65%, which is considered unacceptable (Plante & Vance, 1994), suggesting that at least a third of the children identified by the CELF-4S as LI were classified incorrectly. This means that one in every three Spanish–English bilingual children similar to the children in our sample would be misidentified when using the manual suggested cutoff score. Additionally, a positive likelihood ratio (how much the score is associated with LI if the test result is positive for the disorder; Thordardottir et al., 2011) of 2.39 found in this study is considered insufficient to be clinically informative to determine a child presenting with LI. A positive likelihood ratio around 3.0 is interpreted only as suggestive, meaning that additional testing is necessary to diagnose the disorder with confidence (Dollaghan & Horner, 2011). An improvement in diagnostic accuracy was evident in this sample

with an adjusted cutoff score of 78, that is, 1.5 *SD* below the suggested cutoff score for a clinically identified sample. For both cutoff scores, 85 (the manual suggested) and 78 (empirically derived), a negative result can be used to reliably rule out LI, but a positive result is only suggestive of impairment when the empirically derived cutoff score is used. The standard of best practice is to use a cutoff score that has been empirically derived in order to improve a test's ability to accurately classify children. This empirical approach maximizes sensitivity and specificity, as well as positive and negative likelihood ratios (Greenslade, Plante, & Vance, 2009; Peña, Spaulding, & Plante, 2006). However, this study did not include a confirmatory group in its design, limiting further application of this adjusted cutoff score to the population. Clinicians should interpret standard scores between 78 and 85 in the CELF-4S with caution. It is important to note that the empirically derived cutoff of 78 is closely aligned with the cutoff score of −1.5 *SD*s below the mean that is often used in school settings. This study provides initial support for the use of this cutoff score on the CELF-4S with low-income DLLs attending English-only schools. However, this suggested score does not apply universally to other assessments because diagnostic accuracy measures are test specific, and the CELF-4S may still over identify at this cut-point when we examine the larger sample results.

These results emphasize the need for using converging evidence in the identification of LI in DLLs. The use of one single measure is problematic, given that there is no single highly accurate and valid measure to identify LI in this population. For instance, Restrepo (1998) found that no single measure was sufficient for accurate identification; however, a combination of language sample measures and parent report increased the accuracy in identification. In our sample, we used three measures derived from language samples, a Spanish morphological measure, and an English measure to identify children with and without LI. New measures, such as the Bilingual English-Spanish Assessment (Peña et al., 2014), are promising, but independent studies validating it are needed. Nevertheless, given the variability in bilingual language development at this point, the use of converging evidence is important.

There are important consequences of overidentification of language disorders for this population. First, there is some evidence suggesting that children who are indentified with special needs receive less attention and are less encouraged to succeed than typically developing children (Rosenthal effect; Rosenthal & Jacobson, 1968). Second, children are typically pulled out of classroom time to receive special education services to address their speech and language needs; however, a child with typical development would miss important academic content that is key for academic success and would receive unnecessary language intervention if misdiagnosed with LI. Third, there are economic consequences of receiving an incorrect diagnosis. Parents and school systems use economic and logistic resources to provide children with language intervention. These include time invested by the children, families, therapist, and teachers

and the economic resources involved in paying salaries and resources for special education services. Lastly, there are socioemotional consequences for a child that may impact their development. Thus, the negative effects of over-identification of LI for children, families, and school systems are important, although these effects must be differently understood for the distinct groups (children vs. families) and these subject groups versus the school systems (Conti-Ramsden & Botting, 2004; Jerome, Fujiki, Brinton, & James, 2002; Sciberras et al., 2015; Skeat et al., 2014).

Clinicians should interpret the results of the CELF-4S with caution. Although the use of spontaneous language samples continues to be the gold standard for the identification of children with LI (Heilmann et al., 2010), clinicians who work in a school system are often required to used standardized assessments to qualify children in special education programs (Fulcher-Rood et al., 2015), even when these assessments are not valid. The availability of language assessment tests normed with Spanish–English bilingual children is limited; consequently, finding alternative tests to comply with school policies is not always feasible. However, we strongly recommend that speech-language pathologists providing diagnostic services for this population use language measures that target the true language characteristics of Spanish, instead of using translated and or adapted measures from English-based tests.

In summary, the results of the current study found that the CELF-4S, when used alone, overidentifies Spanish–English bilingual children as presenting with LI when they come from low-income and low-parental education homes and attend English-only programs. These results do not apply to bilingual children with no other risk factors. Two pieces of evidence support the overidentification pattern observed in this study. We found below average performance on a nonselected random group of children attending multiple school programs in a metropolitan city, scoring on average 1 *SD* below the mean. Further, we found that the CELF-4S functioned with low accuracy in identifying children with LI. Clinicians should be cautious when using the CELF-4S to identify bilingual children with LI who come from disadvantageous social conditions, such as low income and low parental education and who are enrolled in English-only programs.

Limitations of the Study and Future Research

This study included a sample of children attending English-only education programs at public and charter schools in the Phoenix, AZ, metropolitan area. This is group with geographical and specific language instruction that may not represent the Latino population in the United States. A study including children from diverse geographical areas, attending public schools with bilingual and English-only programs could provide alternative norms for the CELF-4S in this population. This would also allow for a direct comparison on the CELF-4S performance between children attending bilingual schools versus children in English-only immersion programs.

A second limitation of the study was the use of a complex language sample analysis as one of the criteria to identify children with LI. This type of analysis is unlikely to occur in the field because clinicians working in the school system are often required to used standardized assessments to qualify children in special education programs (Fulcher-Rood et al., 2015) or do not have the time to do such analysis. Additional validation studies of the standardized test most commonly used by speech-language pathologist at the schools are necessary to provide helpful tools to improve diagnostic accuracy of language ability in DLLs.

Lastly, although this investigation provides further guidance for the interpretation of the scores of the CELF-4 for children from low-income English-only education backgrounds, there is still a scarcity of research and training available for clinicians to accurately identify LI in DLLs. As suggested by one of the reviewers, an important issue in our field is that bilingual assessment across the United States is often conducted by interpreters given that only a minority of speech-language pathologists speak Spanish in the United States. Training and support will be needed to implement best practices for assessment in bilingual children.

References

Abutalebi, J. (2008). Neural aspects of second language representation and language control. *Acta Psychologica, 128*(3), 466–478. https://doi.org/10.1016/j.actpsy.2008.03.014

Alvarez-McHatton, P., & Correra, V. (2005). Stigma and discrimination: Perspectives from Mexican and Puerto Rican mothers of children with special needs. *Topics in Early Childhood Special Education, 25,* 131–142.

Anderson, R. T. (1996). Assessing the grammar of Spanish-speaking children: A comparison of two procedures. *Language, Speech, and Hearing Services in Schools, 27*(4), 333–344.

Anderson, R. T. (2001). Lexical morphology and verb use in child first language loss: A preliminary case study investigation. *International Journal of Bilingualism, 5*(4), 377–401.

Anderson, R. T., & Souto, S. M. (2005). The use of articles by monolingual Puerto Rican Spanish-speaking children with specific language impairment. *Applied Psycholinguistics, 26*(4), 621.

Arias, G., & Friberg, J. (2016). Bilingual language assessment: Contemporary versus recommended practice in American schools. *Language, Speech, and Hearing Services in Schools, 48,* 1–15. https://doi.org/10.1044/2016_LSHSS-15-0090

Arizona Department of Education. (2000). *Proposition 203 English language education for children in public schools.* Retrieved from https://www.azed.gov/wp-content/uploads/PDF/PROPOSITION203.pdf

Barnett, W. S., Yarosz, D. J., Thomas, J., Jung, K., & Blanco, D. (2007). Two-way and monolingual English immersion in preschool education: An experimental comparison. *Early Childhood Research Quarterly, 22*(3), 277–293.

Bedore, L. M., & Leonard, L. B. (1998). Specific language impairment and grammatical morphologya discriminant function analysis. *Journal of Speech, Language, and Hearing Research, 41*(5), 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*(4), 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*(02), 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 Bilingualism, 11*(1), 1–29. https://doi.org/10.2167/beb392.0

Bedore, L. M., Peña, E. D., Gillam, R. B., & Ho, T. H. (2010). Language sample measures and language ability in Spanish–English bilingual kindergarteners. *Journal of Communication Disorders, 43*(6), 498–510.

Bedore, L. M., Peña, E. D., Summers, C. L., Boerger, K. M., Resendiz, M. D., Greene, K., ... Guillam, R. B. (2012). The measure matters: language dominance profiles across measures in Spanish–English bilingual children. *Bilingualism: Language and Cognition, 15*(03), 616–629.

Berman, R. A., & Slobin, D. I. (1986). *Frog story procedures in coding manual: Temporality in discourse.* Berkeley, CA: Institute of Human Development, University of California at Berkeley.

Bohman, T. M., Bedore, L., Peña, E. D., Mendez-Perez, A., & Guillam, 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

Caesar, L. G., & Kohler, P. D. (2007). The state of school-based bilingual assessment: Actual practice versus recommended guidelines. *Language, Speech, and Hearing Services in Schools, 38*(3), 190–200.

Castilla-Earls, A., Restrepo, M. A., Perez-Leroux, A. T., Gray, S., Holmes, P., Gail, D., & Chen, Z. (2016). Interactions between bilingual effects and language impairment: Exploring grammatical markers in Spanish-speaking bilingual children. *Applied Psycholinguistics, 37,* 1147–1173. https://doi.org/10.1017/S0142716415000521

Chee, M. W., Hon, N., Lee, H. L., & Soon, C. S. (2001). Relative language proficiency modulates BOLD signal change when bilinguals perform semantic judgments. Blood oxygen level dependent. *NeuroImage, 13*(6, Pt. 1), 1155–1163. https://doi.org/10.1006/nimg.2001.0781

Chondrogianni, V., & Marinis, T. (2011). Differential effects of internal and external factors on the development of vocabulary, tense morphology and morpho-syntax in successive bilingual children. *Linguistic Approaches to Bilingualism, 3,* 318–345. https://doi.org/10.1075/lab.1.3.05cho

Conti-Ramsden, G., & Botting, N. (2004). Social difficulties and victimization in children with SLI at 11 years of age. *Journal of Speech, Language, and Hearing Research, 47*(1), 145–161.

Crowley, C. J. (2010). *A critical analysis of the CELF-4: The responsible clinician's guide to the CELF-4.* Ann Arbor, MI: ProQuest LLC.

Dawson, J. I., Stout, C. E., & Eyer, J. A. (2003). *SPELT-3: Structured Photographic Expressive Language Test.* DeKalb, IL: Janelle Publications.

DeNavas-Walt, C., & Proctor, B. D. (2015). *Income and poverty in the United States: 2014.* Washington, D.C.: U.S. Census Bureau.

Dixon, L. Q., Wu, S., & Daraghmeh, A. (2012). Profiles in bilingualism: Factors influencing kindergartners' language proficiency. *Early Childhood Education Journal, 40*(1), 25–34. https://doi.org/10.1007/s10643-011-0491-8

Dollaghan, C. A. (2007). *The handbook for 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.

Dunn, M., Flax, J., Sliwinski, M., & Aram, D. (1996). The use of spontaneous language measures as criteria for identifying children with specific language impairment: An attempt to reconcile clinical and research incongruence. *Journal of Speech, Language, and Hearing Research, 39*(3), 643–654.

Fulcher-Rood, K., Castilla-Earls, A., & Higginbotham, J. (2015). *Reframing clinical expertise: Understanding the how & why behind clinical decisions.* Poster presentation at the American Speech-Language-Hearing Association, Denver, Colorado.

Gamboa, S. (2015, December 8). More Latino-kids in low income but more financially stable households. *NBC News.* Retrieved from http://www.nbcnews.com/news/latino/more-latino-kids-financially-stable-low-income-households-n476146

Gilliam, W. S., & de Mesquita, P. B. (2000). The relationship between language and cognitive development and emotional-behavioral problems in financially-disadvantaged preschoolers: A longitudinal investigation. *Early Child Development and Care, 162*(1), 9–24.

Golberg, H., Paradis, J., & Crago, M. (2008). Lexical acquisition over time in minority first language children learning English as a second language. *Applied Psycholinguistics, 29,* 41–65.

Greenslade, K. J., Plante, E., & Vance, R. (2009). The diagnostic accuracy and construct validity of the structured photographic expressive language test—preschool: Second edition. *Language, Speech, and Hearing Services in Schools, 40,* 150–160.

Gutiérrez-Clellen, V. F., & Hofstetter, R. (1994). Syntactic complexity in narratives. *Journal of Speech and Hearing Research, 37,* 645–654.

Gutiérrez-Clellen, V. F., Restrepo, M. A., Bedore, L., Peña, E., & Anderson, R. (2000). Language sample analysis in Spanish-speaking children: Methodological considerations. *Language, Speech, and Hearing Services in Schools, 31*(1), 88–98.

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*(6), 1209–1223.

Gutiérrez-Clellen, V. F., & Simon-Cereijido, G. (2007). The discriminant accuracy of a grammatical measure with Latino English-speaking children. *Journal of Speech, Language, and Hearing Research, 50*(4), 968–981.

Gutiérrez-Clellen, V. F., & Simon-Cereijido, G. (2009). A cross-linguistic and bilingual evaluation of interdependence between lexical and grammatical domain. *Applied Psycholinguistics, 30*(2), 315–337.

Hart, B., & Risley, T. R. (1992). American parenting of language-learning children: Persisting differences in family–child interactions observed in natural home environments. *Developmental Psychology, 28*(6), 1096.

Hegde, M. N., & Pomaville, F. (2013). *Assessment of communication disorders in children: Resources and protocols* (2nd ed.). San Diego, CA: Plural.

Heilmann, J., Miller, J. F., Nockerts, A., & Dunaway, C. (2010). Properties of the narrative scoring scheme using narrative retells in young school-age children. *American Journal of Speech-Language Pathology, 19*(2), 154–166.

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development, 74*(5), 1368–1378.

Huang, R. J., Hopkins, J., & Nippold, M. A. (1997). Satisfaction with standardized language testing: A survey of speech-language pathologists. *Language, Speech, and Hearing Services in Schools, 28*(1), 12–29.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels. NCTE Research, Report No. 3.* Washington, D.C.: Office of Education.

Jerome, A. C., Fujiki, M., Brinton, B., & James, S. L. (2002). Self-esteem in children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 45*(4), 700–714.

Kohnert, K. (2010). Bilingual children with primary language impairment: Issues, evidence and implications for clinical actions. *Journal of Communication Disorders, 43*(6), 456–473.

Kohnert, K. J., & Bates, E. (2002). Balancing bilinguals II: Lexical comprehension and cognitive processing in children learning Spanish and English. *Journal of Speech, Language, and Hearing Research, 45,* 347–359.

MacSwan, J., & Rolstad, K. (2006). How language proficiency tests mislead us about ability: Implications for English language learner placement in special education. *The Teachers College Record, 108*(11), 2304–2328.

Mahoney, K., Thompson, M., & MacSwan, J. (2005). The condition of English language learners in Arizona: 2005. In D. Garcia & A. Molnar (Eds.), *The condition of pre-K–12 education in Arizona, 2005* (pp. 1–24). Tempe, AZ: Education Policy Research Laboratory, Arizona State University.

Mayer, M. (1973). *Frog on his own*. New York, NY: Dial Press.

McCauley, R. J., & Swisher, L. (1984). Psychometric review of language and articulation tests for preschool children. *Journal of Speech and Hearing Disorders, 49*(1), 34–42.

Miller, J., & Iglesias, A. (2008). *Systematic Analysis of Language Transcripts (SALT), research version 2008* [Computer software]. Madison, WI: SALT Software.

Montrul, S. (2011). Multiple interfaces and incomplete acquisition. *Lingua, 121*(4), 591–604.

Montrul, S. A. (2008). *Incomplete acquisition in bilingualism: Re-examining the age factor* (Vol. 39). Amsterdam, the Netherlands: John Benjamins.

Morgan, G., Restrepo, M. A., & Auza, A. (2009). Variability in the grammatical profiles of Spanish-speaking children with specific language impairment. *Hispanic child languages: Typical and impaired development, 50,* 283.

Morgan, G., Restrepo, M. A., & Auza, A. (2013). Comparison of Spanish morphology in monolingual and Spanish–English bilingual children with and without language impairment. *Bilingualism: Language and Cognition, 16*(3), 578–596.

Morgan, P., Farkas, G., Hillemeier, M., Matisson, R., Maczuga, S., Li, H., & Cook, M. (2015). Minorities are disproportionately underrepresented in special education: Longitudinal evidence across five disabilities conditions. *Educational Researcher, 20,* 1–15.

Muñoz, M. L., Gillam, R. B., Peña, E. D., & Gulley-Faehnle, A. (2003). Measures of language development in fictional narratives of Latino children. *Language, Speech, and Hearing Services in Schools, 34*(4), 332–342.

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*(3), 172–187.

Paradis, J. (2010a). Comparing typically-developing children and children with specific language impairment. *Experimental Methods in Language Acquisition Research, 27,* 223.

Paradis, J. (2010b). The interface between bilingual development and specific language impairment. *Applied Psycholinguistics, 31*(02), 227–252.

Paradis, J., & Crago, M. (2000). Tense and temporality: A comparison between children learning a second language and children with SLI. *Journal of Speech, Language, and Hearing Research, 43*(4), 834–847.

Pearson, B. Z. (2007). Social factors in childhood bilingualism in the United States. *Applied Psycholinguistics, 28*(03), 399–410. https://doi.org/10.1017/S014271640707021X

Peña, E., Spaulding, T. J., & Plante, E. (2006). The composition of normative groups and diagnostic decision making: Shooting ourselves in the foot. *American Journal of Speech-Language Pathology, 15,* 247–254.

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,* 192–202.

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

Plante, E., & Vance, R. (1994). Selection of preschool language tests: A data-based approach. *Language, Speech, and Hearing Services in Schools, 25*(1), 15–24.

Qi, C. H., Kaiser, A. P., Milan, S. E., Yzquierdo, Z., & Hancock, T. B. (2003). The performance of low-income, African American children on the Preschool Language Scale–3. *Journal of Speech, Language, and Hearing Research, 46*(3), 576–590.

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

Restrepo, M. A. (2003). Spanish language skills in bilingual children with specific language impairment. In S. Montrul & F. Ordoñez (Eds.), *Linguistic Theory and Language Development in Hispanic Languages. Papers from the 5th Hispanic Linguistics Symposium and the 2001 Acquisition of Spanish and Portuguese Conference* (pp. 365–374). Somerville, MA: Cascadilla Press.

Restrepo, M. A., Castilla, A. P., Schwanenflugel, P., Neuharth-Pritchett, S., Hamilton, C. E., & Arboleda, A. (2010). Effects of a supplemental Spanish oral language program on sentence length, complexity, and grammaticality in Spanish-speaking children attending English-only preschools. *Language, Speech, and Hearing Services in Schools, 41*(January), 3–13.

Restrepo, M. A., Gorin, J., & Gray, S. (2013). *Screening Spanish-speaking children for language impairment: Results From a Scale Development Grant*. Inaugural Bilingual Research Conference, University of Houston, Houston, TX.

Restrepo, M. A., & Gutiérrez-Clellen, V. F. (2001). Article use in Spanish-speaking children with specific language impairment. *Journal of Child Language, 28*(02), 433–452.

Restrepo, M. A., Morgan, G. P., & Thompson, M. S. (2013). The efficacy of a vocabulary intervention for dual-language

learners with language impairment. *Journal of Speech, Language, and Hearing Research, 56*(2), 748–765.

Restrepo, M. A., & Silverman, S. (2001). Validity of the Spanish Preschool Language Scale–3 for use with bilingual children. *American Journal of Speech-Language Pathology, 10,* 382–393.

Rhein, D. (2013). Designing a response to intervention plan for English-language learners using the results of language testing. *Journal of Border Educational Research, 9*(1), 15–22.

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom.* New York, NY: Holt, Rinehart & Winston.

Sackett, D. L., Straus, S. E., Richardson, W. S., Rosenberg, W., & Haynes, R. B. (2000). *Evidence-based medicine: How to practice and teach EBM.* Edinburgh, Scotland: Churchill Livingstone.

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*(2), 148–162.

Sciberras, E., Westrupp, E. M., Wake, M., Nicholson, J. M., Lucas, N., Mensah, F., . . . Reilly, S. (2015). Healthcare costs associated with language difficulties up to 9 years of age: Australian population-based study. *International Journal of Speech-Language Pathology, 17*(1), 41–52.

Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical Evaluation of Language Fundamentals–Fourth Edition (CELF-4).* San Antonio, TX: Pearson Education Inc.

Semel, E., Wiig, E. H., & Secord, W. A. (2006). *Clinical Evaluation of Language Fundamentals–Fourth Edition, Spanish Version (CELF-4 Spanish).* San Antonio, TX: Pearson Education Inc.

Simon-Cereijido, G., & Gutiérrez-Clellen, V. F. (2007). Spontaneous language markers of Spanish language impairment. *Applied Psycholinguistics, 28*(02), 317–339.

Skeat, J., Wake, M., Ukoumunne, O. C., Eadie, P., Bretherton, L., & Reilly, S. (2014). Who gets help for pre-school communication problems? Data from a prospective community study. *Child: Care, Health and Development, 40*(2), 215–222.

Smyk, E., Restrepo, M. A., Gorin, J., & Gray, S. (2013). Development and validation of the Spanish–English language proficiency scale (SELPS). *Language, Speech, and Hearing Services in Schools, 44*(July), 252–266. https://doi.org/10.1044/0161-1461(2013/12-0074)

Tape, T. (2003). *The area under an ROC curve.* University of Nebraska Medical Center, Department of General Internal Medicine. Retrieved from http://gim.unmc.edu/dxtests/roc3.htm

Tavizón, J. M. (2014). *The Spanish Language Proficiency of Sequential Bilingual Children and the Spanish–English Language Proficiency Scale* (Unpublished Master of Science thesis). Brighman Young University, Provo, UT.

Thordardottir, E., Kehayia, E., Mazer, B., Lessard, N., Majnemer, A., Sutton, A., . . . Chilingaryan, G. (2011). Sensitivity and specificity of French language and processing measures for the identification of primary language impairment at age 5. *Journal of Speech, Language, and Hearing Research, 54*(2), 580–597.

Uccelli, P., & Páez, M. M. (2007). Narrative and vocabulary development of bilingual children from kindergarten to first grade: Developmental changes and associations among English and Spanish skills. *Language, Speech, and Hearing Services in Schools, 38*(3), 225–236.

Washington, J. A., & Craig, H. K. (1999). Performances of at-risk, African American preschoolers on the Peabody Picture Vocabulary Test–III. *Language, Speech, and Hearing Services in Schools, 30*(1), 75–82.

Wright, S. C., Taylor, D. M., & Macarthur, J. (2000). Subtractive bilingualism and the survival of the Inuit language: Heritage-versus second-language education. *Journal of Educational Psychology, 92*(1), 63.

Zimmerman, I., Steiner, V. G., & Pond, R. E. (1992). *Preschool Language Scale–Third Edition* (PLS-3). San Antonio, TX: The Psychological Corporation.