UNIVERSITY OF CALIFORNIA,
IRVINE

Using Language Sample Analysis to Identify Developmental Language Disorder in Bilingual Children

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Education

by

Michelle Nichols Ramos

Dissertation Committee:
Professor Elizabeth D. Peña, Co-Chair
Professor Penelope Collins, Co-Chair
Professor Lisa M. Bedore

2024

DEDICATION

To

my former students and colleagues in San Marcos -

this work was done with you in mind

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank the research assistants who worked on the coding that was critical to all three studies of this dissertation. I deeply appreciate their patience, attention to detail, and ability to find the fun in what can be a very tedious process. Their feedback and reflections were also valuable for decision-making and ultimately strengthened the research. Thank you to Deya, Johanna, Destiny, Evelyn, and Sharon. I would also like to acknowledge the many HABLA research assistants who worked on the original projects that produced the data I used—the countless hours of data collection and transcription that they contributed allowed me to pursue questions that would not have been feasible otherwise.

I am also incredibly grateful to my committee co-chairs, Dr. Elizabeth Peña and Dr. Penelope Collins, for their guidance and encouragement throughout this process. This was an unexpectedly and uniquely challenging time in history to pursue a Ph.D., but both of them nudged us forward while always holding space to be human. Thank you to Dr. Peña and Dr. Bedore for their generosity with their data and patience in helping me collect the files I needed. A special thanks to the faculty who provided input as mentors and the members of my advancement to candidacy committee - Dr. Katherine Rhodes, Dr. Julie Washington, and Dr. Stephen Schueller. I would also like to thank the members of the HABLA and Write Stuff labs for all their feedback over the last five years. Being part of such a welcoming, uplifting community of fellow scholars made the journey all the lighter.

Finally, I want to express my appreciation to my family. Thank you to Eli, who didn't hesitate when I told him I wanted to pursue a doctorate and continued to roll with punches until it was finished. I am so grateful for him being there to hear out my frustrations, whether he understood them or not, and to celebrate the wins along the way with me. Thank you to Yuni for hanging in there while mommy was busy and for teaching me to keep work in its place.

This research was supported in part by grants from the National Institutes of Health (NIDCD) Contract #N01-DC-8210, #5 K23 DC00141, and #1 R01 DC007439-01. The text of Chapter 2 in this dissertation is a reprint of the material as it appears in Ramos et al. (2022), used with permission from the American Speech-Language and Hearing Association. The co-authors listed in this publication are Penelope Collins and Elizabeth D. Peña.

VITA

Michelle Nichols Ramos

2003-07	Undergraduate Research Assistant, Harold Goodglass Aphasia Research Center, Boston University
2007	B.A. in Linguistics, Boston University
2007-09	Graduate Student Research Assistant, Language Acquisition, Poverty, and Culture Lab, San Diego State University
2009	M.A. in Speech-Language Pathology, San Diego State University
2009-24	Bilingual Speech-Language Pathologist
2019-20	Community Research Fellow, Orange County Educational Advancement Network, University of California, Irvine
2019-24	Graduate Research Assistant, Human Abilities in Bilingual Language Acquisition Lab, University of California, Irvine
2022	Graduate Research Assistant, Language Variation and Academic Success Lab, University of California, Irvine
2023	Teaching Assistant, School of Education, University of California, Irvine
2024	Ph.D. in Education, University of California, Irvine

FIELD OF STUDY

Education, Human Development in Context

PUBLICATIONS

Ramos, M. N., Collins, P., & Peña, E. D. (2022). Sharpening Our Tools: A Systematic Review to Identify Diagnostically Accurate Language Sample Measures. *Journal of Speech, Language, and Hearing Research*, *65(10)*, 3890-3907.

Prado, Y., Ramos, M. N., Peña, E., & Zavala, J. (2022). Dual-language engagement: Concerted cultivation of Spanish use among students, teachers, and parents. *Bilingual Research Journal*, 1-21.

Pratt, A. S., Anaya, J. B., Ramos, M. N., Pham, G., Muñoz, M., Bedore, L. M., & Peña, E. D. (2022). From a distance: Comparison of in-person and virtual assessments with adult–child

dyads from linguistically diverse backgrounds. *Language, Speech, and Hearing Services in Schools, 53*(2), 360-375.

Brownell, H., Lundgren, K., Cayer-Meade, C., Nichols, M., Caddick, K., Spitzer, J. (2007) Assessing quality of metaphor interpretation by right hemisphere damaged patients. *Brain and Language*, *103*(1), 197-198.

ABSTRACT OF THE DISSERTATION

Using Language Sample Analysis to Identify Developmental Language Disorder in Bilingual Children

by

Michelle Nichols Ramos

Doctor of Philosophy in Education

University of California, Irvine, 2024

Professor Elizabeth D. Peña, Co-Chair

Professor Penelope Collins, Co-Chair

Accurately identifying developmental language disorder (DLD) in students who speak a home language other than English has proven to be an enduring challenge. Consequently, students with DLD miss out on critical interventions, and those who are acquiring their two languages in a typical manner are placed in settings that do not meet their educational needs. While many factors underlie the issue, language assessment practices are central to the problem and, therefore, to the solution. Standardized tests remain the preferred method despite repeated recommendations against their use with this population due to bias, lack of validation, and limited availability of instruments in the home language. Language sample analysis (LSA), on the other hand, has often been promoted as best practice, but there is a need for empirically based guidance for selecting and interpreting measures that are accurate indicators of DLD.

To address this need, three studies were conducted as part of this dissertation. Study 1 is a systematic review of studies that examined the diagnostic accuracy of language sample measures in English revealed several LSA measures and composites that reach 80% sensitivity and specificity for children ages 3 to 10 and at least one measure for all years except age six that reaches 90% or greater. Studies 2 and 3 explored the diagnostic accuracy of a set of LSA measures in English and Spanish, respectively, for use with Spanish-English bilingual 5- and 6-year-olds when adjusting for participants' relative exposure to each language. Percent grammatical utterances and errors per C-unit in English each yielded 94% diagnostic accuracy for children with at least 70% English exposure. In Spanish, the best model included errors per C-unit and MLU, but it fell just short of the desired 80% threshold for sensitivity and specificity. Results of the three studies are discussed in regard to their implications for clinical practice, language-specificity of clinical markers of DLD, usage-based theory, and future directions for research.

INTRODUCTION

Approximately two children in every classroom have developmental language disorder (DLD), a neurodevelopmental condition affecting the use of language for understanding others and expressing oneself (Bishop et al., 2017; Tomblin et al., 1997). While children with DLD are underidentified and underserved, students who speak a language other than English are at increased risk for misidentification of DLD and negative educational outcomes related to inappropriate instruction (McGregor, 2020; Norbury et al., 2016; Sullivan, 2011). Among many contributing factors, the lack of access to valid and accurate assessment tools for evaluating bilingual students represents a significant barrier for service providers (Arias & Friberg, 2017; Bedore & Peña, 2008; Guiberson & Atkins, 2012). Language sample analysis (LSA) is often recommended as the gold standard assessment for this population, but there is a lack of clarity on the diagnostic performance of LSA generally and especially limited evidence available regarding its classification accuracy for bilinguals. This dissertation contributes to the repertoire of available assessment tools by exploring the diagnostic potential of LSA in Spanish and English for identification of DLD in bilingual children.

DLD, which has been referred to in the literature using various terms such as specific language impairment (Bishop et al., 2017), is characterized by difficulties learning and using the rules of language that cannot be explained by intellectual, developmental, or physical disabilities (Tomblin et al., 1997). Children demonstrate particular difficulty acquiring grammatical morphology that can be observed in their production of morphosyntactic forms (Leonard, 2014a; Rice & Wexler, 1996). DLD affects 7 to 10% of the population, and the diagnosis is associated with a greater risk of reading disability, math difficulty, emotional and mental health issues, delinquency in youth, and unemployment later in adulthood (Brownlie et al., 2004; Conti-Ramsden & Botting, 2008; Conti-Ramsden et al., 2018; Dubois et al., 2020; Law et al., 2009; St Clair et al., 2019; Winstanley et al., 2021; Young et al., 2002).

Despite the prevalence of DLD and its social and educational consequences, proper identification of students who speak a language other than English has proven to be persistently problematic in the United States. The specific trend toward under- or overidentification has shifted over time and continues to vary by state, district, and grade level (e.g., Artiles et al., 2002; Samson & Lesaux, 2009; Waitoller et al., 2010), but the evidence points to consistent misidentification and inappropriate servicing of this demographic group (Skiba et al., 2008). At the national level, bilingual students are 50% less likely than their monolingual peers to be deemed eligible for language intervention services (Collins et al., 2014; Morgan et al., 2017). Given the academic risks associated with DLD, failure to provide critical intervention services is a tremendous concern for a rapidly growing population that is already vulnerable in terms of academic outcomes and persistence in school (Johnson, 2019; Polat et al., 2016). At the same time, bilingual students are disproportionately placed in special education under the qualifying category of Speech and Language Impaired (SLI; Sullivan, 2011). While overidentification might seem justifiable (or even desirable) to some as a perceived means of providing extra help to struggling students (e.g., Artiles et al., 2010; Kritikos, 2003), in practice, placement in special education services often results in restricted access to English Language Development services (Kangas, 2014, 2018), participation in general education (Cioè-Peña, 2017), and dual language programs (Cioè-Peña, 2017) - the very supports and settings they need for academic success (Collier & Thomas, 2017; Genesee et al., 2005).

The contributing factors to the trend of misidentification of DLD in bilingual children are multifaceted, including school demographics, interpretation of legal statutes, teacher attributions of academic difficulty, professional development, referral processes, and instructional programming (e.g., Artiles et al., 2010, Samson & Lesaux, 2009; Skiba et al., 2008). Of these, assessment practices are frequently identified as central to the issue (e.g., Abedi, 2004; Barrera, 2006; Bedore & Peña, 2008). Bilingual speakers are rarely represented in test norming samples, and scores obtained from English assessments normed on monolingual speakers are likely to reflect English proficiency rather than truly measuring the construct presumed by the test (i.e., impairment). Even if a test does not show evidence of psychometric bias, misguided interpretation of test scores can lead to inappropriate application of the results. Accurate and timely differential diagnosis of DLD in bilingual students is crucial for providing students the instruction and services that are appropriate for their needs and to which they are legally entitled.

The Role of Assessment Practices

The consequences of inadequate assessment tools are magnified by clinicians' strong preference for them. SLPs' reliance on omnibus standardized language tests as a central component of their assessment battery is well-documented (e.g., Fulcher-Rood et al., 2018, 2019; Selin et al., 2019). It is largely motivated by practical needs related to ease as well as setting-based obligations (e.g., legal statutes, insurance requirements). Standardized tests offer broad sampling of language skills, straightforward administration, and a final metric that can be quickly interpreted and objectively compared against institutional eligibility criteria - appealing features within increasingly litigious and resource-constrained practice settings (Blood et al., 2002; Ferney-Harris et al., 2009; Fulcher-Rood et al., 2018; Katz et al., 2010; Sylvan, 2014).

The preference for norm-referenced tests is also rooted in philosophical traditions within the field. This approach to identification of DLD aligns with early efforts to establish objective inclusionary criteria for the disorder based on scores from comprehensive test batteries (Aram et al., 1993; Plante, 1998; Tomblin et al., 1996). Impairment was defined based on a quantifiable distance from age expectations (e.g., Stark & Tallal, 1981) or performance below a designated cutoff (e.g., Paul, 1995; Silva et al., 1983), which operationalized a view of language impairment as a general delay in language development (both in onset and rate) represented by the low end of a normal distribution (Rice, 2003). It also established a precedent for diagnostic assessment focused on comprehensive evaluation of language skills across domains (i.e., semantics, morphology, syntax) and modalities (i.e., receptive and expressive), as well as the application of arbitrary cutoff scores that continues to be common in clinical practice (Plante, 1998; Spaulding et al., 2006).

In contrast with a view of DLD as general language delay, recognition of specific delays displayed by children with DLD within their overall language development has led to a diagnostic approach based on clinical markers (e.g., Ash & Redmond, 2014; Conti-Ramsden, 2003; Plante, 2004; Rice & Wexler, 1996). Clinical markers are linguistic forms that children with typically developing language and those with DLD produce with maximally differing accuracy, characterized by a bimodal distribution (Rice, 2003). According to input-based theories of language acquisition, such as Usage-Based Theory

(Bybee, 2006; Langacker, 1987; Tomasello, 2001), certain structures within a language require a greater amount of input accumulated through experience in order to develop an adult-like representation as a result of the statistical properties of that language (e.g., frequency, perceptual and psychological salience, predictability, similarity to previously acquired forms). These structures, which in English primarily involve verb tense morphemes (Rice & Wexler, 1998), are acquired later by typically developing children. They also pose pronounced difficulty for children with DLD, who struggle to extract relevant patterns from ambient language input and thus need a greater amount of experience and total input to reach the same level of mastery as their peers (Leonard, 2014b). The resulting gap in performance between groups lends these forms as clinical markers of impairment that provide more reliable identification of DLD than arbitrary test cutoff scores.

Along with the shift toward identification based on clinical markers, greater emphasis has been placed on empirical evidence of the diagnostic validity of assessment instruments (Dollaghan, 2004; Friberg, 2010; Plante & Vance, 1994; Spaulding et al., 2006), which in turn has raised additional concerns around the ubiquitous use of standardized tests. The accuracy with which an indicator (i.e., test, task, measure) classifies individuals as affected or unaffected by a condition of interest is commonly quantified using its rate of true positives, or "sensitivity," and true negatives, or "specificity" at a given cutoff point (Dollaghan, 2004; Drobatz, 2009). A widely accepted standard of accuracy is 80% sensitivity and specificity for a measure to be clinically useful, with 90% or greater considered "good" accuracy (Plante & Vance 1994). Many of the most popular language tests fall short of this threshold or have not been validated (Betz et al., 2013; Price et al., 2010; Spaulding et al., 2006).

LSA is often promoted as the gold standard for identifying DLD (e.g., Dunn et al., 1996; Evans, 1996; Miller et al., 2016), especially for bilingual children (Gutiérrez-Clellen & Simón-Cereijido, 2009; Rojas & Iglesias, 2009), but current approaches to LSA employ a general delay rather than a clinical markers paradigm, prompting broad and potentially cumbersome analyses that may overlook the features that separate typical language from DLD. Among desirable features such as ecological validity and efficiency (Costanza-Smith, 2010), one of LSA's advantages is its flexibility of administration and analysis, allowing the clinician to choose from a variety of elicitation tasks and to derive multiple measures of language skill from the sample in order to obtain a comprehensive representation of a child's performance, similar to an omnibus test, that is also easily adapted to the goals of the assessment (Costanza-Smith, 2010). The guidance for conducting and interpreting LSA typically involves comparison to developmental norms (e.g., Prath, 2018) or reference databases (e.g., Pezold et al., 2020) without specifying which measures are the most reliable for diagnosis. The flexibility and open-ended nature of such an approach becomes a double-edged sword as the SLP must contend with a tremendous decision load, given the potential scope of the analysis. Without validation of its diagnostic accuracy, simply increasing the use of LSA in place of tests will not avoid the problem of unreliable identification.

Applying a clinical markers approach to LSA by focusing on a narrower set of highly informative measures would help facilitate appropriate interpretation for diagnostic decisions and ease the burden of the LSA process. A cohesive account of the evidence to date is needed to evaluate the clinical usefulness of LSA measures and identify the most accurate indicators of impairment in order to inform best clinical practice. Whereas current improvisational methods demand expertise, knowledge, and time that many SLPs feel they lack (Klatte et al., 2022; Pavelko et al., 2016), clear evidence-based recommendations for analyzing and interpreting language sample measures would provide the requisite knowledge and reduce time spent making novel decisions - important elements for greater clinical uptake of LSA.

Study 1: A Systematic Review of LSA Diagnostic Accuracy

There is a lack of clarity on which language sample measures can accurately identify DLD and how to interpret them for a diagnosis. The current body of evidence is distributed across several publications, and prior reviews and meta-analyses to date have focused on a particular population, a limited number of LSA measures, or have not focused strictly on LSA measures (Dollaghan & Horner, 2011; Eisenberg & Guo, 2016; Pawlowska, 2014; Shahmahmood et al., 2016). Study 1 examines the scope and strength of available evidence of the diagnostic accuracy of LSA for identifying DLD by conducting a systematic review that focuses on language sample–derived measures and participants representing a wide age range and diverse linguistic backgrounds. The following research questions were addressed: 1) What is the range of LSA measures that have been examined in studies of diagnostic accuracy for identifying DLD using English language samples? 2) Which measures have acceptable diagnostic accuracy, and under what conditions (e.g., age range, sample length, elicitation task)? Results of the review provide SLPs with an easily accessible reference summarizing clinically informative LSA measures and, for the purposes of this dissertation, highlight measures that meet the criterion of acceptable accuracy to inform the selection of measures to be analyzed in Study 2.

Availability of Bilingual Assessors

Best practice for identifying DLD in bilingual speakers is to assess both of the child's languages, and though assessments in the home language are being incorporated with increasing consistency, testing in English continues to be more prevalent (Arias & Friberg, 2017). The capacity within the field to follow best practice is severely limited, not only by test availability but also availability of bilingual personnel (Caesar & Kohler, 2007; Guiberson & Atkins, 2012; Santhanam et al., 2019). Over 400 languages are spoken by families of US students, and between 5 and 225 languages are spoken in any given state (U.S. Department of Education, n.d.). However, only 8% of SLPs nationwide speak a language other than English, with competence reported in only 83 spoken languages, as well as ASL and other sign languages (ASHA, 2022). More than 60% of bilingual SLPs practice in California, New York, Texas, and Florida (ASHA, 2022). Conducting assessments in collaboration with an interpreter is often recommended to address this gap, but SLPs report doing so infrequently, lacking confidence, and experiencing challenges accessing interpreter services due to both human and financial resource constraints (Arias & Friberg, 2017; Guiberson & Atkins, 2012; Santhanam et al., 2019).

The challenge in identifying DLD based on L2 English alone is related to the overlap between the grammatical error patterns that are most indicative of impairment in monolingual English speakers and the linguistic patterns that are typical of second language acquisition (Paradis, 2008; Simon-Cereijido & Gutiérrez-Clellen, 2007). Additionally, performance among children acquiring English is quite heterogeneous as they produce different types of errors at different rates based on their amount of exposure to English (Bedore et al., 2012). This may lead to misidentification of DLD in cases where a child who is typically developing but has had more limited experience in the target language demonstrates lower accuracy than a child who has DLD but enough experience to have reached critical mass for acquiring that form.

Patterns of early and late-acquired forms observed in monolingual language acquisition are similar in bilingual acquisition but are moderated by cross-linguistic transfer and interference effects. For example, among the English inflectional forms that are later acquired (i.e., more difficult or complex), bilinguals with various home languages master copula and auxiliary verbs much faster relative to tense inflections (Paradis & Blom, 2016; Paradis, 2005). Acquisition of third person singular -*s*, on the other hand, varies depending on the inflectional richness of the home language (Blom et al., 2012). These findings demonstrate how schemas formed in one language may serve to facilitate learning of similar linguistic forms in the other (Bybee, 2008; Ellis, 2008; Gathercole & Hoff, 2007) in the same way that children transfer knowledge between related forms within the same language (Abbot-Smith & Behrens, 2006). Bilingual children with DLD follow a similar developmental progression as typical bilingual peers but at a slower rate (Jacobson & Yu, 2018) and may not transfer knowledge between languages as readily (Blom & Paradis, 2015). Just as an individual's prior experience predicts their overall language development, current use of L1 and L2 as a measure of language experience best predicts a child's language proficiency in and dominance across the two languages (Bedore et al., 2012).

With multiple sources of variability, it is critical to shift from thinking about disordered language and second language profiles separately to considering how disorder presents within second language variation (Bedore et al., 2018). Language assessment based on performance must consider the dynamic interaction between language typology, experience, and ability. Consideration for the characteristics of the student population being served and the capacities and needs of clinicians can inform research questions that are relevant and applicable to practice, such as the possibility of a valid English-only approach to bilingual assessment. Focusing on English measures that can accurately identify DLD in bilingual children has the potential for broad impact on practice among the majority of SLPs who are monolingual and have limited access to support from speakers of the many languages represented in the US.

Prior Research on English LSA Measures for Bilinguals

A substantial body of research has been conducted on the diagnostic accuracy of LSA measures for monolingual speakers of mainstream English, but only two studies have previously explored composites of English LSA measures for identifying DLD in bilingual speakers. Ooi and Wong (2012) examined MLU in words, a Malaysian English adaptation of IPSyn Total, and lexical diversity *D* with Malaysian Cantonese-English speakers ages 3;8 to 5;11. The composite fell short of acceptable with 78% sensitivity and specificity. Smyk (2012) examined a composite of MLU in words, errors per T-unit, number of different words, and percent maze words with Spanish-English bilingual children ages 5;3 to 8. It yielded 83% overall diagnostic accuracy, but since the disaggregated metrics were not reported, findings should be cautiously interpreted as suggestive but not conclusive. In addition to limitations in reporting, several of the LSA measures examined in these studies were found to have inconsistent or inadequate diagnostic accuracy even with monolingual speakers, and variation in relative exposure to each language was not accounted for in the analysis.

Study 2: Validation of English LSA Measures for Bilinguals

Only two studies have examined the diagnostic accuracy of English LSA measures with bilingual speakers, and findings were either clinically inadequate or inconclusive as reported. The LSA measures used in these studies did not necessarily correspond to those found to have the best diagnostic accuracy with monolingual speakers and also did not account for variation in relative exposure to each language. Study 2 builds on previous findings by focusing on LSA measures with evidence of good diagnostic accuracy, exploring their usefulness with a bilingual population representing a continuum of language experience, and testing the performance of each measure at various cut points to identify its maximum accuracy level. The following research questions were addressed: 1) What is the optimal classification accuracy for identifying DLD in Spanish-English bilingual 5- and 6-year-olds using percent ungrammatical utterances, errors per utterance, MLU in words, and subordination index calculated from English narratives? 2) Is classification accuracy improved by adjusting for language exposure? 3) Is classification accuracy improved by using a combination of these measures? Testing the potential of English-only assessment practices to accurately classify language ability in bilingual children will help improve the validity of language assessments in settings where testing in the home language is not feasible, benefitting both the monolingual practitioner and the child.

Availability of Bilingual Assessment Tools

Assessment practices for bilingual children are improving but still fall short of the standards set by the Individuals with Disabilities Education Act and the American Speech-Language and Hearing Association (Arias & Friberg, 2017), which include multiple sources of evidence and assessment in the child's native language. The preference for standardized language tests that SLPs demonstrate with their general caseload is also apparent in their practices specifically with bilingual students, which presents a challenge as there is a persistent shortage of standardized tests in languages other than English, especially tests that have been validated for the identification of DLD (Arias & Friberg, 2017; Castilla-Earls et al., 2020; Guiberson & Atkins, 2012; Huang et al., 1997). As a culturally responsive assessment, LSA can minimize bias related to cultural and linguistic variation found in tests (Kraemer & Fabiano-Smith, 2017; Rojas & Iglesias, 2009; Stockman, 1996), and the ability to conduct LSA in the home language is also not limited by the availability of a published instrument in the same way that standardized testing is. While SLPs are reportedly more likely to use LSA for such cases (Arias & Friberg, 2017; Pavelko et al., 2016), the obstacle is a lack of familiarity with the language-specific characteristics of DLD that would be necessary for clinical interpretation (Guiberson & Atkins, 2012) and of empirical data to guide those decisions even for the most commonly spoken languages, such as Spanish in the United States.

Translating the tests and procedures that are available in English is one strategy that has been employed in order to meet the need for home language assessment, though it is ultimately an ineffective one due to the fact that the profile and clinical markers of DLD are language-specific and may no longer be captured post-translation (Arnold & Matus, 2000; Bedore & Peña, 2008; Bracken & Barona, 1991; Peña, 2007). For example, English-speaking children make frequent errors in producing verb tense and agreement markers as well as complex sentences, while the errors typically associated with DLD involve word order in Germanic languages and verb aspect markers in Mandarin and Cantonese (Leonard, 2014b). Clinical markers vary according to the unique structural properties of the language being acquired and which features in that language are optional, especially complex, or deviate from a usual pattern (Leonard, 2014b). In both English and Spanish, verb morphology is affected and can serve to differentiate typical from atypical development. However, while verb tense morphemes, such as regular past tense -*ed* and 3rd person singular -*s*, are strong indicators of DLD in English (Rice & Wexler, 1996), analogous forms in Spanish are not clinically useful largely due to their regular and transparent nature. Instead, subject-verb agreement and verb aspect, specifically use of subjunctive, are most clinically informative, along with noun-based morphology, such as clitic pronouns (e.g., Bedore & Leonard, 2001; Castilla-Earls et al., 2021; Jacobson & Schwartz, 2002). Accurate identification of DLD for linguistically diverse children depends on accounting for these cross-linguistic differences and the language-specific profile of DLD.

In the case of bilinguals who are acquiring two distinct linguistic systems, bilingual children with DLD show a similar profile of weaknesses in each language as monolingual speakers with DLD (Leonard, 2014b) but with additional variation related to experience with the languages being acquired. Bilingual children rely on the same learning mechanisms to develop emerging modules of each language (Hernandez et al., 2005; Wulff & Ellis, 2018), but the amount of time and experience the child has in each language as well as the properties of the ambient input become especially relevant (e.g., Bedore et al., 2012;

Blom et al., 2012; Bohman et al., 2010; Thordardottir, 2015). Because the usage exemplars a bilingual child is exposed to are divided between two languages, building a catalog of exemplars large enough to reach “critical mass” for mastering a particular form in a given language requires more time overall compared to monolingual counterparts (Gathercole, 2007). The timeframe for acquiring specific linguistic forms is largely determined by the relative amount of exposure to each language over time in interaction with the distributional properties of the two languages (Paradis, 2010). Examination of potential indicators of impairment must account for a child’s experience with the language in order to properly index expectations of performance to exposure and to explore the possibility that the indicators themselves may differ by level of language exposure in terms of their diagnostic utility (Castilla-Earls et al., 2016; Coloma et al., 2016).

Better understanding of clinical indicators of DLD across different languages would enable clinicians to conduct more valid and accurate assessments in the child’s home language, whether administering them themselves or in collaboration with an interpreter. LSA offers some unique advantages that can be leveraged for more rapid change in practice in the midst of the scarcity of home language assessments. Research to validate LSA methods for diagnosis of DLD across different languages would provide the desired expertise, and incorporating widely available materials and clinically feasible methods, as well as thoroughly reporting relevant procedures and metrics, would allow recommended practices to be readily implemented upon dissemination.

Prior Research on Spanish LSA Measures

While Spanish language sample data has been used frequently to compare the performance of bilingual speakers with DLD to those with typically developing language, few studies have examined the accuracy of language sample measures for classifying Spanish speakers based on ability status. One study identified a combination of LSA measures that achieved fair diagnostic accuracy for Spanish samples elicited using narrative retell and narrative generation tasks (Kapantzoglou et al., 2017). Forty 4- to 5-year-old bilingual children who were initially identified as having typically developing language or DLD using the Bilingual English-Spanish Assessment (Peña, et al., 2018) were accurately classified using combinations of lexical diversity *D*, mean length of utterance in words (MLU-w), subordination index, and grammaticality as errors per C-unit. The combination of grammaticality and lexical diversity *D* yielded 90% sensitivity and 85% specificity, for an overall classification accuracy of 87.5%, when applied to story retell data. Accuracy was lower though still acceptable when grammaticality and subordination index were applied to story generation data, with only 80% sensitivity and 85% specificity.

When examined independently, these and other measures were found to be inadequate for identifying DLD on their own. MLU-w yielded only 58% sensitivity and 74% specificity in a study of Spanish-dominant 4- and 5-year-olds (*N*=48) using a narrative retell sample (Simón-Cereijido & Gutiérrez-Clellen, 2007). Better sensitivity was achieved for other participant samples consisting of 30 bilingual children (85%; Lazewnik et al., 2019) and 55 3- to 6-year-old monolingual Spanish-speaking children (81%; Grinstead et al., 2013), but specificity remained inadequate at 57.1% and 76%, respectively. Percent ungrammatical utterances nearly reached acceptable accuracy with 79% sensitivity and 100% specificity for Spanish-dominant (i.e., >80% input at home) 4- and 5-year-olds (Simón-Cereijido & Gutiérrez-Clellen, 2007) but fell well below the 80% threshold for near-monolingual (90% or greater exposure to Spanish) 3- to 5-year-olds (59% sensitivity, 67%

specificity; Guiberson et al., 2015). Errors per T-unit as well as subordination index were examined with monolingual 3- to 6-year-olds' conversational samples and presumably found to be inadequate as only results from measures that met the 80% criterion were reported (Grinstead et al., 2013). Number of different word roots (NDW) yielded acceptable sensitivity for this participant group (85%) but inadequate specificity (72%; Grinstead et al., 2013).

The main limitations in these studies involved the participant sample and analysis. Most of these studies included a monolingual or near-monolingual participant sample (i.e., greater than 80% exposure to Spanish; Grinstead et al., 2012; Simón-Cereijido & Gutiérrez-Clellen, 2007; Guiberson et al., 2015). The two studies that included bilingual participants did not clearly quantify or did not report participants' degree of bilingualism. For example, Kapantzoglou et al. (2017) described their bilingual participants only as non-native speakers of English whose first language was Spanish, while Lazewnik et al. (2019) reported mean years of exposure for their Spanish-English dual language learners but not relative input/output in each language. These studies also did not include language exposure in the analysis, so varying levels of diagnostic accuracy across different levels of language exposure would have been obscured. Discriminant function analysis, which examines diagnostic accuracy at a single cutoff score, was the primary analytic method across studies, so measures with borderline acceptable accuracy (e.g., 79% sensitivity/100% specificity for ungrammaticality in Simón-Cereijido & Gutiérrez-Clellen, 2007) could not be explored further to see if a more optimal cutoff yielded better, clinically useful accuracy.

Study 3: Validation of Spanish LSA Measures

A limited number of studies have examined the diagnostic accuracy of LSA measures in Spanish, and findings were inconsistent. Degree of bilingualism among participants varied between studies, and language exposure was not accounted for in analyses. Furthermore, conclusions regarding the clinical usefulness of the measures examined were limited by analytic methods that tested only one cutoff value. Study 3 extends previous findings by exploring the usefulness of promising LSA measures across a broader range of language experience and using more informative methods (e.g., ROCs to test various cut points). The following research questions were addressed: 1) What is the optimal classification accuracy for identifying DLD in Spanish-English bilingual 5- and 6-year-olds using percent ungrammatical utterances, errors per utterance, MLU in words, and subordination index calculated from Spanish narratives? 2) Is classification accuracy improved by adjusting for language exposure? 3) Is classification accuracy improved by using a combination of these measures? Building on the limited evidence base for diagnostic LSA in Spanish will guide SLPs who are bilingual or are working with an interpreter in selecting and interpreting appropriate measures that could be applied to a broader bilingual client base.

Summary

Correcting the trend of misidentification of DLD will require a shift in assessment practice. Given the growing population of bilingual children in U.S. schools and their vulnerability to misidentification of DLD, the need for assessment methods that are culturally and linguistically appropriate as well as psychometrically sound is critical. The three studies in this dissertation examine the existing evidence for the diagnostic accuracy of English LSA measures and further explore the diagnostic potential of a set of LSA measures in English and in Spanish for bilingual children along a continuum of language experience. Analyses examine the effect of adjusting for language exposure on diagnostic accuracy and test a range of cutoff scores to identify optimal performance. To promote uptake in clinical practice, the language elicitation procedures and materials used are publicly available, and a popular software package was chosen for transcribing samples and calculating the LSA measures of interest, though use of this software is not necessary for replication. SLPs would benefit from greater clarity regarding which LSA measures offer the best diagnostic accuracy for their general caseload and specifically for bilingual students with varying levels of experience in each of their languages. Findings also inform future directions for studies to replicate and extend findings across different participant samples, language pairs, and LSA methods.

Chapter 1: Systematic Review of LSA Measures (Study 1)

Published in *Journal of Speech, Language, and Hearing Research* as:

Ramos, M. N., Collins, P., & Peña, E. D. (2022). Sharpening Our Tools: A Systematic Review to Identify Diagnostically Accurate Language Sample Measures. *Journal of Speech, Language, and Hearing Research*, *65*, 3890-3907. https://doi.org/10.1044/2022_JSLHR-22-00121

Abstract

Purpose: This systematic review provides a comprehensive summary of the diagnostic accuracy of English language sample analysis (LSA) measures for the identification of developmental language disorder.

Method: An electronic database search was conducted to identify English publications reporting empirical data on the diagnostic accuracy of English LSA measures for children age 3 or older.

Results: Twenty-eight studies were reviewed. Studies included between 18 and 676 participants ranging in age from 3;0 to 13;6. Analyzed measures targeted multiple linguistic domains, and diagnostic accuracy ranged from less than 25% to greater than 90%. Morphosyntax measures achieved the highest accuracy, especially in combination with length measures, and at least one acceptable measure was identified for each 1-year age band up to 10 years old.

Conclusion: Several LSA measures or combinations of measures are clinically useful for the identification of developmental language disorder, though more research is needed to

replicate findings using rigorous methods and to explore measures that are informative for adolescents and across diverse varieties of English.

Background

Within the field of speech-language pathology, language sample analysis (LSA) is often promoted as the gold standard for assessing language (Miller et al., 2016) and the "cornerstone of any clinical assessment battery" (Evans, 1996, p. 207) to identify language impairment, more recently termed developmental language disorder (DLD; Bishop et al., 2017). Yet, LSA is not deployed as such in typical clinical practice. Speech-language pathologists (SLPs), on the whole, do not conduct LSA regularly nor adhere to consistent procedures (Fulcher-Rood et al., 2018), relying instead on standardized language tests (Fulcher-Rood et al., 2019; Selin et al., 2019) despite concerns raised around their inadequate accuracy for identifying DLD (Betz et al., 2013) and cultural and linguistic bias in test design (Castilla-Earls et al., 2020; Horton-Ikard, 2010). Among the barriers to greater adoption of LSA is a lack of clarity on the diagnostic value of a language sample, specifically which measures are the most accurate indicators of impairment and how to interpret them to determine a diagnosis of DLD. Although there is a growing body of evidence addressing these questions, it is distributed across several publications and is not readily available for easy reference by clinicians. Therefore, in this review, we seek to consolidate the existing evidence of the diagnostic accuracy of LSA into a single resource to guide the selection and interpretation of these measures in clinical practice and to inform future research and policy-directed advocacy efforts.

DLD affects approximately 7-10% of children and is characterized by difficulties learning and using the rules of language in the absence of intellectual, developmental, or physical disabilities that would explain the disorder (Tomblin et al., 1997). Core characteristics of DLD include particular difficulty acquiring grammatical morphology, which is often observed in children's morphosyntactic productions (Leonard, 2014a; Rice & Wexler, 1996). In English, typical errors associated with DLD involve verb tense marking and agreement errors as well as difficulty producing complex sentences. Although this profile of language disabilities has been referred to as DLD in recent years, other terms are used in the literature including specific language impairment, language impairment, and primary language impairment (Bishop et al., 2017).

Several features of LSA are well-suited for clinical purposes and merit its status as the gold standard of assessment tools (for a more detailed discussion, see Costanza-Smith, 2010). A significant amount of information about a child's language ability can be extracted from a short sample, making LSA efficient and highly adaptable to the goals of an assessment (Heilmann et al., 2010). A notable strength of LSA over standardized assessments is its ecological validity, or generalizability to everyday function (Hewitt et al., 2005). This quality appeals the most to clinicians, who report that they typically use LSA for information about functional performance in naturalistic contexts (Fulcher-Rood et al., 2018). Because language samples are elicited through naturalistic interactions (e.g., conversation or storytelling), they can be collected in familiar and culturally responsive ways, minimizing the bias present in many standardized tests (Kraemer & Fabiano-Smith, 2017; Stockman, 1996). The variety of methods for eliciting the sample offers the flexibility of administration that is useful in situations not as conducive to standardized testing, such as the recent shift to remote testing due to COVID-19 (Manning et al., 2020). Further, LSA data are informative not only for identifying impairment but also for planning treatment and monitoring progress (Costanza-Smith, 2010; Price et al., 2010).

Though SLPs generally endorse LSA as a valuable assessment tool, in practice, they show a strong preference for standardized language tests (Fulcher-Rood et al., 2018, 2019; Selin et al., 2019), with nearly a third not using LSA at all for assessment (Pavelko et al., 2016). Attention has been drawn to issues of misdiagnosis when using standardized tests with inadequate diagnostic accuracy or non-empirical cutoff scores (Betz et al., 2013; Price et al., 2010; Spaulding et al., 2006), and it is equally important to scrutinize LSA against the same standard if it is to be promoted as best practice. The commonly recommended practice of comparing LSA results to developmental norms (Heilmann, 2010; Prath, 2018) or database norms (e.g., SALT reference databases; Castilla-Earls et al., 2020; Pezold et al., 2020; Rojas & Iglesias, 2009) is useful for characterizing language samples but just as susceptible to classification errors without consideration of the diagnostic accuracy of the measures. Guarded descriptions of LSA as supplemental or supporting evidence for clinical decisions (Pezold et al., 2020; Price & Jackson, 2015; Rojas & Iglesias, 2009) and limited acceptance of LSA data within institutional eligibility criteria (Pavelko et al., 2016) reflect ambivalence toward the diagnostic value of LSA, signaling a need to clarify the status of the evidence to date.

Synthesis and evaluation of the evidence available for diagnostic LSA are critical for its validation as an evidence-based practice and also for guiding clinical practice. The high variability in how LSA is implemented (Pavelko et al., 2016) suggests that standard practice is heavily influenced by individual decision-making. The improvisation involved in using self-designed protocols or none at all demands greater expertise and time - the most commonly cited barriers to implementing LSA (Klatte et al., 2022; Pavelko et al., 2016) - and thus undermines rather than increases efficiency. Technology and accompanying protocols have enabled tremendous improvement in the LSA process through the systematization and automation of its more tedious aspects (e.g., digital recording, increasingly accurate and accessible speech-to-text capability, dedicated analysis software; Pezold et al., 2020). However, the most consequential decision of a diagnostic assessment - how to interpret LSA results for a determination of impairment - remains largely at the clinician's discretion, who must choose from dozens of possible measures with limited consensus on their diagnostic usefulness or interpretation to guide that decision. Given the high stakes associated with diagnostic and eligibility decisions in increasingly litigious settings (Sylvan, 2014), the safer option often is to avoid using LSA for its perceived subjectivity. If LSA cannot serve the purpose for which it is conducted, the time and effort required even for streamlined procedures are likely to outweigh any value added (Klatte et al., 2022).

Clearly outlined selection criteria informed by evidence could help reduce the number of novel decisions a clinician must make during analysis and increase confidence in interpretation, thereby capturing the advantage of standardized tests (Sylvan, 2014). To enable SLPs to select the most trustworthy LSA measures for diagnosis and gauge an appropriate level of confidence in their selection and interpretation (Spaulding et al., 2006), criteria should include the client's age, language background, and elicitation procedures used and detail the accuracy metrics and associated cutoff score for available measures accordingly. Such guidelines could help to ease the burden of LSA as a task and ensure the accuracy of LSA as a tool, thereby fostering the perception of LSA as an efficient, informative, and defensible assessment - a true gold standard.

Prior Reviews

Previous systematic reviews of the diagnostic accuracy of LSA have focused on specific populations or sets of measures (Dollaghan & Horner, 2011; Eisenberg et al., 2001; Eisenberg & Guo, 2016) or included LSA measures among other language assessments in their analyses (Dollaghan & Horner, 2001; Pawlowska, 2014; Shahmahmood et al., 2016). The evidence for MLU indicates that it can provide supportive evidence of a disorder but on its own is not adequate for diagnosing DLD in preschool children (Eisenberg et al., 2001). Measures of morphosyntactic diversity and development (i.e., Tense Marker Total - quantifies the types of verb tense morphemes produced; Developmental Sentence Scoring (DSS) Total - rates the developmental level of forms used in eight linguistic categories) were also found inadequate for identifying impairment in this age group (Shahmahmood et al., 2016). In contrast, measures of morphosyntactic accuracy have yielded acceptable to good diagnostic accuracy for children in preschool through early elementary (Eisenberg & Guo, 2016; Shahmahmood et al., 2016). These included Percent Grammatical Utterances, the Sentence Point Score from the DSS, and the Finite Verb Morphology Composite (FVMC). PGU expresses grammaticality as a percentage of total utterances that are correct, while the Sentence Point Score is an average of points awarded per utterance for grammaticality (i.e., one point for a grammatical utterance, zero for an ungrammatical utterance). The FVMC reflects the accuracy of four clinical markers in obligatory contexts - third-person singular present – *s*, regular past tense – *ed*, and copula and auxiliary BE.

The FVMC and Tense Marker Total were also included in a meta-analysis along with two other morphosyntactic LSA measures; however, the author was unable to determine the diagnostic value of the measures due to heterogeneity across studies (Pawlowska, 2014). The additional measures were Percent Verb Tense, which calculates the accuracy of all obligatory verb tense marking, and Productivity Score, which reflects the diversity of contexts in which morphemes are produced. When used with Spanish-English bilingual children, meta-analysis results indicated that the FVMC and an obligatory subject measure were diagnostically suggestive at best and not recommended as individual measures (Dollaghan & Horner, 2011).

Purpose

The purpose of the current study is to examine the scope and strength of available evidence of the diagnostic accuracy of LSA for identifying DLD, which is used in this review to broadly refer to language impairment inclusive of prior terminology. A cohesive account of the evidence base is necessary to inform guidance for best clinical practice and provide a comprehensive summary of clinically useful LSA measures for SLPs' easy reference. To that end, this review builds on previous reviews and meta-analyses by limiting the scope to only language sample-derived measures while expanding it to include any such measure and participants representing a wide range of ages and diverse linguistic backgrounds. The following questions were addressed:

1. What is the range of LSA measures that have been examined in studies of diagnostic accuracy for identifying DLD using English language samples?
2. Which measures have acceptable diagnostic accuracy and under what conditions (e.g., age range, sample length, elicitation task)?

Methods

Literature Search Strategy

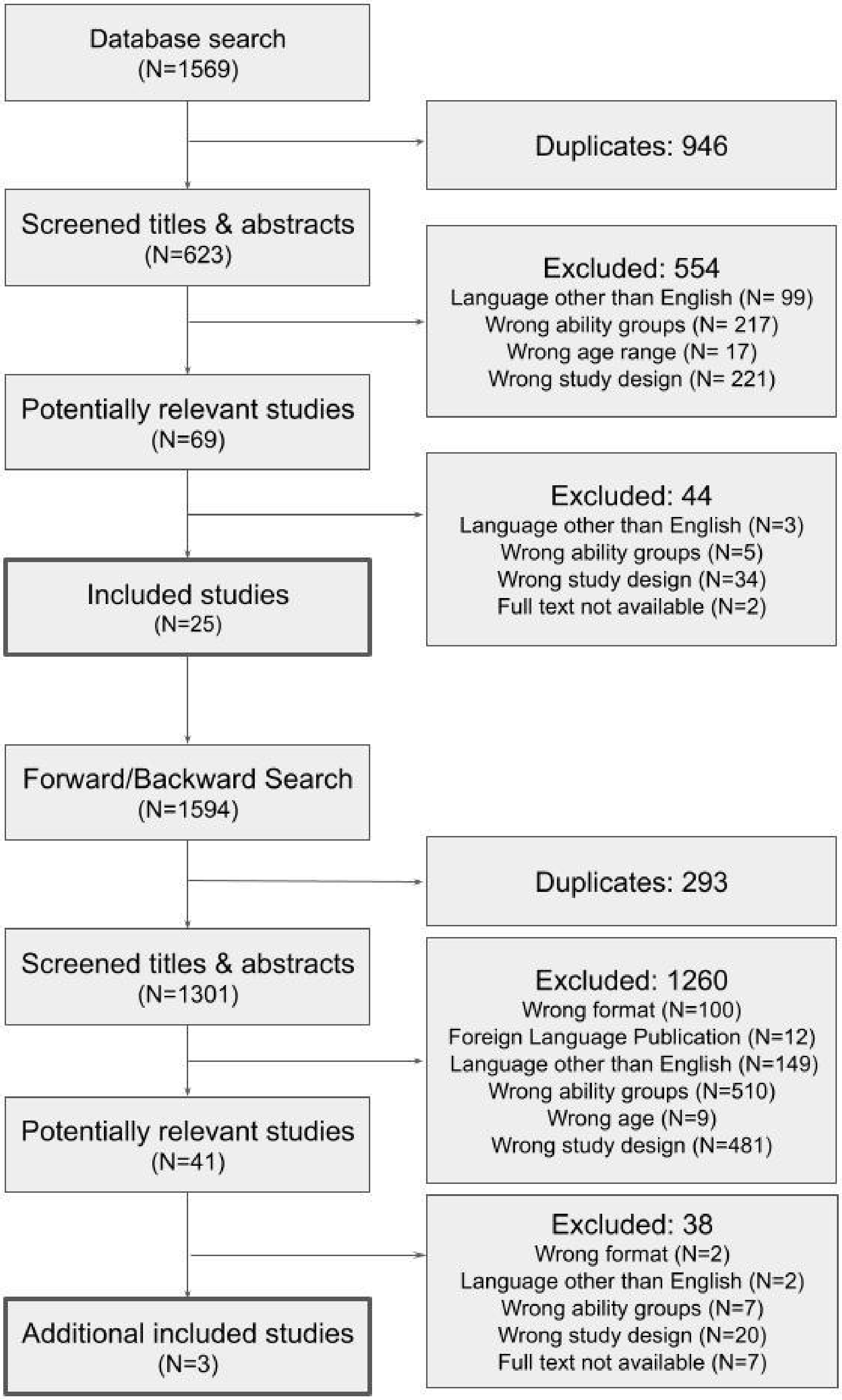
An electronic search for English-language publications reporting on the diagnosis of DLD using language sample analysis was conducted in January 2021 using the databases SCOPUS, PubMed, Web of Science, APA PsycINFO, ERIC, Medline Complete, and ProQuest Dissertations & Theses Global. To search these databases, we used a combination of terms representing the constructs of *developmental language disorder* (*language impair**, *language disorder*, DLD, SLI*), *LSA* generally and its individual measures (*language sample*, index of productive syntax, developmental sentence scoring, mean length of utterance, productivity, type-token ratio, number of different word*, subordination index, argument structure, lexical measure*, grammaticality, grammar measure*, syntax measure*, syntactic measure**), and various metrics of *diagnostic accuracy* (*sensitivity AND specificity, diagnos*, classif*, identif*, predict*, discrim*, likelihood ratio*). These three sets of terms were joined by the Boolean operator 'AND', and terms within each set were joined by 'OR.' The combination of these terms were applied to the title, abstract, keywords, and subject terms fields. Results were filtered for English as the language of publication, and the year of publication was not restricted.

Study Selection Criteria

The search and selection process is summarized in the PRISMA chart in Figure 1.1. Titles and abstracts of the 623 unique results returned by the database searches were screened for relevance based on the following inclusion criteria: an empirical study published as a journal article, thesis or dissertation, conference paper or chapter in an edited volume; language sample data was elicited in English; the participant sample included participants both with and without DLD; DLD was a primary diagnosis without comorbidities (e.g., studies with participants with language impairment secondary to another diagnosis were excluded); participants were age 3 to 18 (e.g., studies that only included toddlers younger than 36 months were excluded), and the study design and analytic methods addressed diagnostic accuracy (e.g., studies that only examined the statistical significance of group differences were excluded).

Figure 1.1

Search and Selection Process


Database search
(N=1569)
Duplicates: 946
Screened titles & abstracts
(N=623)
Excluded: 554
Language other than English (N= 99)
Wrong ability groups (N= 217)
Wrong age range (N= 17)
Wrong study design (N= 221)
Potentially relevant studies
(N=69)
Excluded: 44
Language other than English (N=3)
Wrong ability groups (N=5)
Wrong study design (N=34)
Full text not available (N=2)
Included studies
(N=25)
Forward/Backward Search
(N=1594)
Duplicates: 293
Screened titles & abstracts
(N=1301)
Excluded: 1260
Wrong format (N=100)
Foreign Language Publication (N=12)
Language other than English (N=149)
Wrong ability groups (N=510)
Wrong age (N=9)
Wrong study design (N=481)
Potentially relevant studies
(N=41)
Excluded: 38
Wrong format (N=2)
Language other than English (N=2)
Wrong ability groups (N=7)
Wrong study design (N=20)
Full text not available (N=7)
Additional included studies
(N=3)

The first author developed a coding manual of keywords for inclusionary and exclusionary criteria, which was reviewed and revised with the other authors. For example, keywords for inclusion based on target diagnosis were *developmental language disorder/DLD*, *specific language impairment/SLI, primary language impairment/PLI*, *language impaired/impairment*, and keywords for exclusion were *autism spectrum disorder/ASD*, *Asperger's*, *Fragile X Syndrome*, *Down's Syndrome*, *hearing impaired/impairment*, *Alzheimer's/dementia*, *aphasia*, *ADD/ADHD*, *phonological delay/disorder*, and *speech sound disorder.* The first author trained an undergraduate research assistant on the coding manual with 10 studies, followed by joint screening of 13 studies. They then screened and compared decisions for batches of 25 studies until agreement reached 90%, after which they double-screened and compared every 4 batches to prevent drift. Ultimately, the first author screened all studies, and the research assistant independently screened 25% of studies, with 94% agreement. Discrepancies were resolved through discussion. Five hundred fifty-four (554) studies were excluded during this phase.

The full text of the remaining 69 studies was examined to confirm the inclusion criteria and additional criteria that 1) procedures for calculating LSA measures were transparent and could be performed in a clinical setting (e.g., machine learning models were excluded), and 2) for assessment batteries that also included standardized tests or probes, diagnostic accuracy data was disaggregated by measure (i.e., diagnostic accuracy was reported for the LSA measures separately from the other non-LSA assessment measures). The first author screened all studies, and the research assistant screened 20% of the texts, with 93% reliability. Discrepancies were resolved through discussion. Forty-four (44) studies were excluded in this phase.

Forward and backward citation chaining from the twenty-five remaining studies was conducted using SnowGlobe (McWeeny et al., 2021) as well as the reference and 'Cited By' lists exported from SCOPUS for 4 studies that were incompatible with SnowGlobe. This process yielded 1301 unique results that had not appeared in the electronic database search results. These studies were screened using the previously described procedures and criteria. The first author screened all studies, and the research assistant screened 20% of the studies, with 94% reliability for titles and abstracts and 100% reliability for full texts. One thousand two hundred sixty (1260) and 38 studies were excluded during these phases, respectively.

The following data points were extracted from included studies and compiled in Google Sheets: participant sample size, participant age range, participant language background, reference standard, language sample elicitation task, average and/or range of language sample length, LSA measures analyzed, LSA measure cutoff score(s), sensitivity, specificity, overall diagnostic accuracy, positive likelihood ratio, negative likelihood ratio, and confidence intervals. The first author coded all studies, and the second author coded 25% of studies, with 90% reliability. Discrepancies were resolved through discussion.

Results

Twenty-eight (28) studies were ultimately included in this review (see Figure 1.1 for the complete selection process) and are listed in Table 1.1. Language sample elicitation tasks across the corpus included play, narrative tell and retell, conversation, and expository. The size of participant samples ranged from 18 to 676 children, with an average of 159 participants. Participant age also varied significantly across studies, ranging from age 2 to 13;6, though 4-, 5-, and 6-year-olds were included most frequently (14, 16, and 14 studies, respectively), followed by 3- and 7-year-olds (10 studies each). Participants of 25 studies were monolingual speakers of mainstream English (ME) from the United States and Canada, one of which also included British English speakers. Three studies included speakers of African American English, two of which also included speakers of Southern White English. Two studies included bilingual speakers of English.

Table 1.1

Studies Included for Review

Source	N	Age Range	Elicitation Task	Analyzed Measure(s)
Bedore & Leonard, 1998	38	3;7-5;9	Play Picture description	Mean Length of Utterance Noun Morphology Composite Verb Morphology Composite
Castilla-Earls & Fulcher-Rood, 2018	100	4;0-6;11	Narrative retell	Grammaticality & Utterance Length Instrument
Charest et al., 2020	377	4-9 years	Narrative tell	Moving Average Type-Token Ratio Number of Different Words
Dunn et al., 1996	242	2;6-6;11	Play	Mean Length of Utterance Percent structural errors
Eisenberg & Guo, 2013	34	3;0-3;11	Picture description	Percent Grammatical Utterances Percent Sentence Point Percent Verb Tense Usage
Fletcher & Peters, 1984	29	3;4-6;11	Play Picture description Narrative retell	Unmarked Verb Forms Verb Types
Gavin et al., 1993	47	2;0-4;2	Conversation	Stage 1 Major Utterances

			Play	Three-Element Noun Phrases Verb Phrase Errors
Gladfelter & Leonard, 2013	55	4;0-5;6	Play	Finite Verb Morphology Composite Tense and Agreement Productivity Score Tense Marker Total
Guo & Eisenberg, 2014	36	3;0-3;11	Play	Finite Verb Morphology Composite Tense and Agreement Productivity Score
Guo & Schneider, 2016	129	6 & 8 years	Narrative tell	Errors per C-unit Finite Verb Morphology Composite Percent Grammatical C-units
Guo et al., 2019	377	4-9 years	Narrative tell	Percent Grammatical Utterances
Guo et al., 2020	377	4-9 years	Narrative tell	Finite Verb Morphology Composite
Heilmann et al., 2010	488	3;0-13;6	Conversation	10 SALT measures
Hewitt et al., 2005	54	5;5-6;7	Conversation Narrative retell	IPSyn Total Mean Length of Utterance Number of Different Words
Hoffman, 2009	48	8-10 years	Narrative tell	Proportion “restricted” utterances
Klee et al., 2017	48	2;0-4;0	Play	Lexical diversity *D* Mean Length of Utterance

Liles et al., 1995	114	7;6-12;6	Narrative retell	Cohesive ties
				Mean # of subordinate clauses per T-unit
				Mean # of words per subordinate clause
				Percent of grammatical T-units
Moyle et al., 2011	100	5;5-9;9	Conversation	Mean Length of Utterance (morphemes)
			Expository	Noun Morphology Composite
				Verb Morphology Composite
Oetting & McDonald, 2001	93	4-6 years	Play	35 nonmainstream patterns
Oetting et al., 2021	106	5 years	Play	8 tense/agreement forms
Ooi & Wong, 2012	18	3;8-5;11	Play	IPSyn Total
			Conversation	Lexical diversity *D*
				Mean Length of Utterance (words)
Overton et al., 2021	37	<6 years	Play	DSS Total
				IPSyn
Pavelko & Owens, 2019	306	3;0-7;11	Conversation	Clauses per Sentence
				Mean Length of Utterance (SUGAR)
				Total words
				Words per Sentence
Rudolph et al., 2019	676	6;11-7;3	Play	Finite Verb Morphology Composite

Scheffel, 1997	37	8;4-13;2	Expository (Map task)	Expansions References to map Total turns Total words
Schneider et al., 2006	377	4;0-9;11	Narrative retell	Story Grammar score
Smyk, 2012	73	5;3-8	Narrative	Errors per T-unit Mean Length of Utterance Number of Different Words Percent maze words
Souto et al., 2014	112	4;0-5;10	Play	DSS Sentence Point DSS Total Finite Verb Morphology Composite Mean tense/agreement Mean Top 5 tense/agreement

Note. LARSP = Language Assessment Remediation and Screening Procedure. SALT = Systematic Analysis of Language Transcripts. IPSyn = Index of Productive Syntax. DSS = Developmental Sentence Scoring

RQ 1: LSA Measures Examined for Diagnostic Accuracy

Because of the plethora of analyses that can be conducted from a language sample, the first research question explored which measures have been examined for diagnostic accuracy in order to establish the scope of evidence that is available for LSA. Reviewed studies examined a wide range of language sample measures across the domains of morphology, syntax, semantics, discourse, and pragmatics. These measures are summarized in Table 1.2.

Table 1.2

Description and Frequency of LSA Measures Analyzed in Included Studies

LSA Measure	Frequency	Description
Morphosyntax: Accuracy		
DSS Sentence Point [a]	1/28 (4%)	Total points awarded to grammatical sentences (1 point if no errors)
Errors per T-unit [b]	1/28 (4%)	Number of grammatical errors divided by total T-units
(Finite) Verb Morphology Composite [a c d e f g h i j]	9/28 (32%)	% of correct productions in obligatory contexts of regular past tense, 3rd person singular present, copula BE, and auxiliary BE. Modifications also included auxiliary DO[a] or irregular past tense[f.]
Nonmainstream patterns [k]	1/28 (4%)	Total occurrences of 35 grammatical surface features that are possible in Southern African American English and/or Southern White English
Noun Morphology Composite [c d]	2/28 (7%)	% of correct productions in obligatory contexts of possessive -s, plurals, articles
Omitted bound morphemes (SALT) [l]	1/28 (4%)	Number of obligatory morphemes that were omitted
Omitted words (SALT) [l]	1/28 (4%)	Number of obligatory words that were omitted
Percentage Grammatical T-units [m] /Utterances [n]	2/28 (7%)	Number of grammatical utterances divided by total utterances
Percent Structural Errors [o]	1/28 (4%)	% of utterances that contain a morphological or syntactic error (e.g., word order, omitted morpheme, omitted word, telegraphic speech)
Percent Verb Tense Usage [e]	1/28 (4%)	% of correct production in obligatory contexts of tense marking including: copula/auxiliary BE, auxiliary DO, bound tense markers, irregular past or 3rd person verb forms

Tense & Agreement Forms [p]	1/28 (4%)	% of occurrence in possible contexts of mainstream overt, nonmainstream overt, and zero forms of 8 targets (past tense regular, past tense irregular, verbal -s habitual, verbal -s nonhabitual, 4 auxiliary BE forms)
Unmarked Verb Forms [q]	1/28 (4%)	Number of lexical verbs produced without premodification or inflection
Verb Phrase Errors [r]	1/28 (4%)	Number of errors occurring within verb phrases
Morphosyntax + Semantics: Accuracy		
Errors per C-unit [h]	1/28 (4%)	Number of grammatical errors* divided by total C-units
Percent Grammatical Utterances[e] / C-Units[h,s]	3/28 (11%)	% of utterances not containing any coded errors*
Percent Sentence Point [e]	1/28 (4%)	% of utterances awarded a point for containing no errors* (excluding C-units with a missing subject or missing main verb)
Proportion ‘restricted’ utterances [t]	1/28 (4%)	% of utterances with a complete clause (i.e., subject and predicate) and one or more syntactic or semantic errors
Utterance errors (SALT) [l]	1/28 (4%)	Number of utterances that contained a syntactic error, three or more word-level omissions/errors, or did not make sense
Word errors (SALT) [l]	1/28 (4%)	Number of incorrect productions of lexical items
Morphosyntax: Proficiency		
Clauses per Sentence [x]	1/28 (4%)	Number of clauses in the sample divided by total sentences
DSS Total [a u]	2/28 (7%)	Total points across all utterances divided by total utterances (Sentence Point plus 1-8 points awarded for each form produced within 8 categories: main verb, indefinite pronouns/noun modifiers, personal pronouns, secondary verbs, negatives, conjunctions,

		interrogative reversals, and Wh-questions)
IPSyn Total [u,v,w]	3/28 (11%)	Total ratings across four categories: noun phrases, verb phrases, question and negation, and sentence structure. Each structure is rated for frequency in the sample: 0=never, 1=once, 2=twice or more
Mean subordinate clauses per T-unit[m]	1/28 (4%)	Number of subordinate clauses divided by total T-units
Mean tense/agreement [a]	1/28 (4%)	Sum of DSS Main Verb category scores for each utterance divided by total utterances that earned at least a score of 1 for this category
Mean top 5 tense/agreement [a]	1/28 (4%)	Average of the five highest scores in the DSS Main Verb category
(Tense & Agreement) Productivity Score [f,g]	2/28 (7%)	Number of different uses (i.e., with different subjects, different lexical verbs inflected, different morphemes within the category) up to 5 of morphemes in 5 categories (copula, auxiliary BE, auxiliary DO, 3rd person singular, regular past), with 0-25 possible points
Tense Marker Total [f]	1/28 (4%)	Number of forms occurring at least once in samples from a set of 15 targets (cop/aux/3PS/reg past/DO), with 0-15 possible points
Morphosyntax: Length		
GLI: Length [n]	1/28 (4%)	Length of each utterance is rated as one of 3 intervals (≤3 words, 4-7 words, or ≥8 words) and a weighted average is calculated
Mean Length of Utterance (MLU) [b,c,d,l,o,v,w,x,y]	9/28 (32%)	Number of free and inflectional morphemes[c,d,l,v,y] or words[w,x] divided by total utterances/sentences
Mean Length of Utterance: SUGAR[x]	1/28 (4%)	Number of morphemes divided by total utterances (18 derivational morphemes, and each word in a proper name = 1 morpheme; all contractions, *hafta*, *wanna*, and *gotta* = 2; *gonna* = 3)

Mean words per subordinate clause [m]	1/28 (4%)	Number of words within subordinate clauses divided by total subordinate clauses
Stage 1 Major Utterances [r]	1/28 (4%)	1-word utterances produced as commands, questions, or statements
Three-Element Noun Phrases [r]	1/28 (4%)	Noun phrases with 3 words (i.e., determiners, modifiers, prepositions)
Semantics		
Lexical diversity *D* [w y]	2/28 (7%)	Repeated ratio of number of different words to total words calculated using CLAN software's *vocd* program
Moving Average Type-Token Ratio [z]	1/28 (4%)	Average of type-token ratios (ratio of different word types to total word tokens) calculated for successive 100-word cuts of the transcript
Number of Different Words [b l v z]	4/28 (14%)	Number of different word roots produced in the sample. Alternative calculations used the first 200 words of the sample[z], the first 41 utterances[z], and 50 utterances[v]
Verb Types [q]	1/28 (4%)	Number of different/unique verbs produced in the sample
Pragmatics/Discourse		
Between-utterance pauses (SALT) [l]	1/28 (4%)	Total seconds of pausing between two utterances (no speech for ≥2s)
Complete cohesive ties [m]	1/28 (4%)	Total intersentential cohesive ties (conjunctive, reference, lexical, ellipsis) that were complete (i.e., information referred to by the cohesive marker is easily found and defined without ambiguity)
Expansions [aa]	1/28 (4%)	Whether child's response to examiner's question about a nonexistent map feature expanded on features of the map/discovered
Mean turn length [l]	1/28 (4%)	Total main body words divided by total conversational turn
Percentage maze words [b l]	2/28 (7%)	% of words that were reduplications, revisions,

		filled pauses, or false starts
References [aa]	1/28 (4%)	Total number of map features mentioned by the child
Story Grammar [ab]	1/28 (4%)	Total points awarded for inclusion of Story Grammar elements based on a story-specific rubric (character(s), setting, initiating event, etc.)
Total Number of Words [x] [aa]	2/28 (7%)	Total words produced in the sample (including unintelligible words[a])
Total turns [aa]	1/28 (4%)	Total number of conversational turns in the sample

Note. Frequency indicates the number of studies and the percentage of the total included studies that analyzed the measure. DSS = Developmental Sentence Scoring. SALT = Systematic Analysis of Language Transcripts. IPSyn = Index of Productive Syntax. GLI = Grammaticality & Length Instrument. SUGAR = Sampling Utterances and Grammatical Analysis Revised. * Coded errors included missing verb, missing obligatory argument/constituent, pronoun substitution, tense marking, grammatical morphemes (articles, plural -s, obligatory present participle -ing, prepositions), and lexical/other. [a] Souto et al., 2014. [b] Smyk, 2012. [c] Bedore & Leonard, 1998. [d] Moyle et al., 2011. [e] Eisenberg & Guo, 2013. [f] Gladfelter & Leonard, 2013. [g] Guo & Eisenberg, 2014. [h] Guo & Schneider, 2016. [i] Guo et al., 2020. [j] Rudolph et al., 2019. [k] Oetting & McDonald, 2001. [l] Heilmann et al., 2010. [m] Liles et al., 1995. [n] Castilla-Earls & Fulcher-Rood, 2018. [o] Dunn et al., 1996. [p] Oetting et al., 2021. [q] Fletcher & Peters, 1984. [r] Gavin et al., 1993. [s] Guo et al., 2019. [t] Hoffman, 2009. [u] Overton et al, 2021. [v] Hewitt et al., 2005. [w] Ooi & Wong, 2012. [x] Pavelko & Owens, 2019. [y] Klee et al., 2017. [z] Charest et al., 2020. [aa] Scheffel, 1997. [ab] Schneider et al., 2006.

Morphosyntax

Morphosyntactic measures constituted the broadest category with more than 15 unique measures. *Morphosyntactic accuracy* was measured as overall grammaticality or error frequency (i.e., proportion of grammatically correct utterances in a sample, errors per utterance) or the production of specific grammatical forms or types of errors (e.g., Verb Morphology Composite, Unmarked Verbs). In studies comparing diagnostic accuracy across dialects of English, frequency of occurrence of grammatical patterns of interest was calculated across possible contexts or total utterances rather than obligatory contexts (Oetting & McDonald, 2001; Oetting et al., 2021). Some grammaticality measures also reflected semantic accuracy, such at Utterance Errors from SALT or Percent Sentence Point (Eisenberg & Guo, 2013).

In addition to accuracy, morphosyntax was also measured in terms of *proficiency* - used here to refer to expertise with morphosyntactic production - and *length*. These measures quantified the range or diversity of forms (e.g., Tense Marker Total, Tense and Agreement Productivity Score), developmental sophistication (e.g., DSS, IPSyn), and complexity (e.g., clauses per sentence, DSS Coordination score) of participants' morphosyntactic production. Length was most often examined at the utterance level (i.e., MLU), and generally calculated in either words or morphemes. Some unique variations included mean words per subordinate clause rather than per utterance (Liles et al., 1995), categorical rating of length (i.e., 1-3 words, 4-7 words, etc.; Castilla-Earls & Fulcher-Rood, 2018), and the inclusion of a wider range of structures, such as derivational morphemes -*ly* and -*ful* (Pavelko & Owens, 2019).

Semantics

Semantic measures focused on either overall diversity (e.g., type-token ratio, number of different words; Charest et al., 2020) or diversity within specific word classes (e.g., Verb Type; Fletcher & Peters, 1984). Number of different words was analyzed for different calculation methods (Charest et al., 2020) and in combination with other measures (Hewitt et al., 2005; Smyk, 2012). Type-token ratio (Charest et al., 2020) and lexical diversity *D* (Klee et al., 2017; Ooi & Wong, 2012) are based on the proportion of total words that are unique instances.

Pragmatics and Discourse

Measures of pragmatics included number (Scheffel, 1997) and length of turns (Heilmann et al., 2010). Length of the sample, or total number of words, was also examined in two studies (Scheffel, 1997; Pavelko & Owens, 2019). Discourse quality in terms of clarity and organization was measured based on specific references to details in the elicitation materials (Scheffel, 1997), story grammar components (Schneider et al., 2006), and cohesive ties (Liles et al., 1995). Discourse fluency was measured using the proportion of maze words to total words (Heilmann et al., 2010; Smyk, 2012), between-utterance pause length, and words per minute (Heilmann et al., 2010).

RQ 2: Diagnostic Accuracy of LSA Measures

To determine diagnostic accuracy, measures of interest are used to predict whether each participant belongs to the DLD or typically developing group based on whether the value of that measure (or weighted composite of measures) falls above or below a particular cutoff. That predicted status is then compared with their actual status as was determined at the outset of the study using a chosen reference measure, often a prior diagnosis by an SLP or a standardized test. The diagnostic accuracy of the measure is the percentage of participants whose predicted language ability status correctly matches their actual status, and is often calculated separately for accurate identification of DLD (i.e., sensitivity) and accurate identification of typical language (i.e., specificity). A commonly accepted threshold of "acceptable" diagnostic accuracy is 80% sensitivity and specificity or greater, and 90% or greater is considered "good" (Plante & Vance, 1994). Results of the reviewed studies are summarized in Table S1 of Supplemental Material S1 (see Appendix).

Morphosyntax Measures: Accuracy

Measures of grammaticality were generally found to have acceptable diagnostic accuracy, with more specific measures of morphosyntactic accuracy reaching acceptable to good accuracy. Percent Sentence Point yielded 100% sensitivity and 82% specificity for 3-year-olds using picture description as an elicitation task (Eisenberg & Guo, 2013), which is within the range found for 4- and 5-year-olds using play-based samples (93% sensitivity/94% specificity and 100% sensitivity/100% specificity; Souto et al., 2014) and narratives (83% sensitivity/96% specificity and 100% sensitivity/82% specificity; Guo et al., 2019). Comparable accuracy was achieved when measuring grammaticality as the percentage of grammatical T- or C-units or inversely as the proportion of utterances with errors. Using 3-year-olds' picture description samples (Eisenberg & Guo, 2013) and 4- to 10-year-olds' narratives (Guo & Schneider, 2016; Guo et al., 2019; Hoffman, 2009), adequate sensitivity and specificity (83-100% and 82-96%) was achieved, with good accuracy for 9-year-olds (90% sensitivity/specificity). Similarly, errors per C-unit yielded 91% sensitivity and 82% specificity for 6-year-olds' narrative samples, and 94% sensitivity and 80% specificity for 8-year-olds (Guo & Schneider, 2016).

The Finite Verb Morphology Composite, which targets forms considered to be clinical markers of DLD, was examined in several studies and generally had acceptable to good diagnostic accuracy moderated by age and sample length. For 3;0-3;11 children's play samples, a sample of 100 utterances is needed to achieve at least acceptable accuracy (83% sensitivity/89% specificity), as shorter samples of only 50 utterances yielded inadequate sensitivity of 67% (Guo & Eisenberg, 2014). The inclusion of additional tense and agreement forms in the measure, as with percent verb tense usage, also results in acceptable diagnostic accuracy for this age group (100% sensitivity/82% specificity; Eisenberg & Guo, 2013). For 4- and 5-year olds, FVMC yielded good sensitivity (91-100%) and specificity (93-100%) across studies using play-based elicitation (Gladfelter & Leonard, 2013; Souto et al., 2014) and narrative (Guo et al., 2020). Bedore and Leonard (1998) found acceptable accuracy for the verb composite alone with their sample of children ranging in age from 3;7 to 5;9 (84% sensitivity/100% specificity), which seems consistent with the pattern of acceptable improving to good accuracy moving up through the preschool ages.

Findings for children ages five and older are inconsistent across studies but suggest an age-related ceiling for the clinical usefulness of FVMC. Guo and Schneider (2016) and Guo et al. (2020) found that FVMC diagnostic accuracy decreases with increasing age (82% sensitivity/90% specificity for 6-year-olds' narrative samples; 85% sensitivity/86% specificity for 7-year-olds; 76% sensitivity/80% specificity for 8-year-olds; 80% sensitivity/76% specificity for 9-year-olds). Moyle et al. (2011) found inadequate accuracy with their sample, which included children from age 5;5 to 9;9 and thus appears consistent with this age-related pattern. One study's results deviated significantly from these, finding very poor sensitivity (26-35%) for 5;11-6;3 children's conversational samples when compared against three different reference measures - MLU, the Peabody Picture Vocabulary Test-Revised, and nonword repetition (Rudolph et al., 2019).

Three reviewed studies analyzed the diagnostic accuracy of language sample measures based on dialect-specific grammatical patterns, building on previous research investigating clinical markers of language disorder within linguistic variation (Oetting et al., 2016). A model comprised of 35 nonmainstream dialectal patterns yielded acceptable diagnostic accuracy (87% sensitivity/94% specificity) for 4- to 6-year-old speakers of Southern African American English (SAAE) and rural Southern White English (SWE), but a reduced model of 4 patterns did not perform as well (74% sensitivity; Oetting & McDonald, 2001). A reduced dialect-specific composite of 5 patterns also yielded acceptable accuracy for SWE speakers, but not for SAAE speakers (75% specificity). Eight tense and agreement forms previously found to be diagnostically useful within an elicitation probe fell short of acceptable levels for 5-year-old SAAE and SWE speakers, with the exception of past tense using strategic scoring for SWE speakers (89% sensitivity/specificity; Oetting et al., 2021).

Morphosyntax Measures: Proficiency

Measures of morphosyntactic developmental level or productivity demonstrated more limited diagnostic usefulness. For the TAPS, as with the FVMC, samples of only 50 utterances yielded inadequate diagnostic accuracy of 94% sensitivity and 50% specificity for 3-year-olds (Guo & Eisenberg, 2014). Samples of 100 utterances still fell short of acceptable (89% sensitivity/78% specificity), but improved when the group was disaggregated into younger and older 3-year-olds (88% sensitivity/specificity for 3;0-3;5; 90% sensitivity/80% specificity for 3;6-3;11), generating an age-specific cutoff. Using a cutoff score of 87 on the SPELT-P2 rather than a primarily clinical reference criterion also yielded good accuracy (100% sensitivity/specificity for 3;0-3;11), though this was based on only a subset of participants. Diagnostic accuracy did not reach acceptable levels for 4-year-olds despite samples of more than 100 utterances (67% sensitivity/88% specificity) but did for 5-year-olds (80% sensitivity/80% specificity; Gladfelter & Leonard, 2013). Instead, the related Tense Marker Total identified 4-year-olds more accurately (83% sensitivity/88% specificity; Gladfelter & Leonard, 2013). The DSS Total and the IPSyn Total were found to be inadequate for both ME (Hewitt et al., 2005; Souto et al., 2014) and AAE speakers under age 6 (Overton et al., 2021), as were the subscales that were evaluated for ME speakers. Syntactic complexity, however, yielded acceptable accuracy (83% sensitivity/91% specificity) for conversational samples with 3- to 7-year-olds, as did their total number of words (86% sensitivity/84% specificity; Pavelko & Owens, 2019).

Length Measures

Many studies have examined MLU, both independently and combined with other measures. Alone, its accuracy varies significantly. Bedore & Leonard (1998) found MLU to nearly reach good accuracy with children ages 3;7-5;9 (95% sensitivity/89% specificity), but this was not replicated with the validation sample (100% sensitivity/68% specificity). The replication findings are consistent with other studies, which found at least one of the accuracy metrics to be inadequate (i.e., 67% sensitivity for children ages 5;5-6;7 in Hewitt et al., 2005; 72% sensitivity for children ages 5;5-9;9 in Moyle et al., 2011). Modifications to the way MLU is typically calculated, as with the Sampling Utterances and Grammatical Analysis Revised (SUGAR) protocol, resulted in better accuracy (86% sensitivity/86% specificity) with 3- to 7-year-olds' conversational samples (Pavelko & Owens, 2019).

Semantics, Pragmatics, and Discourse Measures

Measures of semantics and pragmatics or discourse were generally found to be diagnostically inadequate, falling below the 80% standard in one or both metrics. Number of different words yielded poor sensitivity (20-44%) across two studies of 4- to 9-year-olds even when calculated in various ways (Charest et al., 2020; Hewitt et al., 2005). Moving average type-token ratio similarly yielded only 26% sensitivity for this age range (Charest et al., 2020). A measure of story grammar yielded 70% sensitivity and 84% specificity for 4;0-9;11 children's narratives (Schneider et al., 2006). One study used an expository task involving description of a route on a map to analyze discourse and pragmatic behaviors measured by total words, number of turns, references to map, and number of expansions in response to prompt, and these measures collectively yielded 75% and 60% specificity for children ages 8;4 to 13;2 (Scheffel, 1997).

Composite Measures

Several models of combined measures also achieved acceptable levels of diagnostic accuracy, all of which included either MLU or a grammaticality measure. For very young children of 2 to 4 years, MLU combined with lexical diversity and an age factor yielded 86% sensitivity and 91% specificity (Klee et al., 2017), and for 3- to 7-year-olds, MLU and clausal density together yielded 97% sensitivity and 82% specificity (Pavelko & Owens, 2019). When MLU was combined with the noun morphology composite for children ages 3;7 to 5;9, diagnostic accuracy was nearly good (89% sensitivity and 100% specificity; Bedore & Leonard, 1998) and better than the noun and verb composites together (84% sensitivity/100% specificity) or the combination of all three measures (89% sensitivity/95% specificity). The Grammaticality & Length Instrument (GLi), which includes a grammaticality score and a categorical average of utterance length, yielded 83% sensitivity and 92% specificity for 4;0-6;11 children's narrative retell samples (Castilla-Earls & Fulcher-Rood, 2018). Unmarked Verbs + Verb Types, two categories from the Language Assessment, Remediation and Screening Procedure (LARSP; Crystal et al., 1976), together yielded 89% sensitivity and 90% specificity for children ages 3;4 to 6;11 (Fletcher & Peters, 1984) and outperformed any other combination of measures considered in the study. Though the MLU + Noun composite results could not be replicated with 5;5-9;9 (Moyle et al., 2011), a comprehensive model of 10 measures from the Systematic Analysis of Language Transcripts (SALT; Miller & Iglesias, 2008) Standard Measures Report - MLU in morphemes, mean turn length, omitted words, omitted bound morphemes, word errors, utterance errors, number of different word roots, words per minute, percentage of maze words, and between-utterance pauses - yielded acceptable diagnostic accuracy for conversational samples from children in this age range and even younger (87% sensitivity/specificity for 3;0-5;11, 80% sensitivity/85% specificity for 6;0-9;11; Heilmann et al., 2010).

None of the composite models tested with older children 10- to 13-years-old reached acceptable diagnostic accuracy. The comprehensive SALT model, which performed well with younger children, achieved only 77% sensitivity with 10;0 to 13;6 (82% specificity; Heilmann et al., 2010). A combination of grammaticality by T-unit, clausal density, average length of subordinate clause in words, and total cohesive ties yielded 82% overall diagnostic accuracy with 9;0-11;4 participants' narrative retell samples (and only 77% when used with participants 8;6-12;6), but disaggregated metrics were not reported to verify whether the threshold of at least 80% sensitivity and specificity was met (Liles et al., 1995). Similarly, pragmatics and discourse measures had poor accuracy for this age range (8;4-13;2; Scheffel, 1997), though they have not been evaluated with younger children to be able to distinguish age from measure-related effects.

Two composites were explored for bilingual speakers of English. For Malaysian Cantonese-English speakers ages 3;8 to 5;11, a composite of MLU in words, a Malaysian English adaptation of IPSyn Total, and lexical diversity *D* fell short of acceptable (78% sensitivity and specificity; Ooi & Wong, 2012). For Spanish-English bilingual children ages 5;3 to 8, a composite of MLU in words, errors per T-unit, number of different words, and percent maze words yielded 83% overall diagnostic accuracy, but since the disaggregated metrics were not reported, findings should be cautiously interpreted as suggestive but not conclusive (Smyk, 2012).

Best Diagnostic Accuracy

Examining diagnostic accuracy by age, there are multiple options for monolingual speakers of mainstream English with at least acceptable accuracy for each year interval between age 3 and 10 and at least one measure or model with good accuracy for each year interval except 6 (see Table 1.3). For 3-year-olds, studies found that MLU combined with a verb composite score (referred to as the FVMC in later studies; Bedore & Leonard, 1998) or age combined with 3 LARSP categories (Gavin et al., 1993) can achieve at least 90% sensitivity and specificity using conversation or play-based language samples, though more modest levels were found for the LARSP model in the validation study (91% sensitivity/80% specificity). For 4- and 5-year-olds, both the traditional and modified FVMC yield good accuracy with play (Gladfelter & Leonard, 2013; Souto et al., 2014) and narrative samples (Guo et al., 2020), as did the DSS Sentence Point with play samples

(Souto et al., 2014). Though none of the measures examined with 6-year-olds reached 90% sensitivity and specificity, several reached the 80% threshold of acceptable: FVMC, PGCU, Errors per C-unit, MLU + Clauses per Sentence, GLi, Unmarked Verbs + Verb Types, and a combination of 10 SALT measures.

For 7- to 10-year-olds, Liles et al.'s (1995) model combining measures of cohesive ties, grammaticality, subordinate clauses per T-unit, and clause length based on 7;6-10;6 children's narrative samples yielded 97% overall accuracy. With the exception of PGU for 9-year-olds (Guo et al., 2019), this was the one set of measures that reached good accuracy for children older than 6. However, since its sensitivity and specificity cannot be evaluated separately, other measures that still have acceptable accuracy may be preferable, such as FVMC or the SUGAR model for 7-year-olds and age-appropriate grammaticality measures (errors per C-unit, PGU, percent "restricted" utterances) for 8- to 10-year-olds. Beyond age 10, none of the measures or models definitively met the threshold for acceptable diagnostic accuracy, as previously discussed.

One single measure and one composite of measures yielded acceptable accuracy for 5-year-old and 4- to 6-year-old speakers of SWE, respectively: strategic scoring of past tense and a combination of zero irregular past, auxiliary DO, zero irregular 3rd, and subject-verb agreement of *don't*. Analyzing a set of 35 nonmainstream features achieved acceptable accuracy for AAE speakers, but more parsimonious models were inadequate. None of the models examined with speakers of English as an additional language were clinically useful.

Table 1.3

LSA Measures with Best Diagnostic Accuracy by Age

Measure	Elicitation Task	Materials	Sensitivity	Specificity	Overall	Cutoff
Mainstream English Speakers						
3-year-olds						
Age + Stage 1 Utterances + VP Errors + 3-Element NP [a]	Conversation/Play	Toys	91%	92%	-	Yes
FVMC + MLU [b]	Play/Picture description	Toys, picture sequences	95%	95%	-	No
4-year-olds						
FVMC modified (4;0-4;6) [c]	Play	Toys	100%	100%	-	Yes
FVMC [d]	Play	Toys	93%	94%	-	Yes
FVMC [e]	Narrative tell	Picture sequences (ENNI)	92%	94%	94%	Yes
DSS Sentence Point [d]	Play	Toys	93%	94%	-	Yes
5-year-olds						
FVMC modified (5;0-5;6) [c]	Play	Toys	92%	93%	-	Yes
FVMC [d]	Play	Toys	91%	93%	-	Yes
FVMC [e]	Narrative tell	Picture sequences (ENNI)	100%	90%	92%	Yes

Measure	Elicitation Task	Materials	Sensitivity	Specificity	Overall	Cutoff
DSS Sentence Point [d]	Play	Toys	100%	100%	-	Yes
6-year-olds						
MLU (SUGAR) + Clauses/Sentence [f]	Conversation	Personal topics (SUGAR protocol)	97%	82%	-	Yes
Errors per C-unit [g]	Narrative tell	Picture sequences (ENNI)	91%	82%	85%	Yes
Unmarked Verb Forms + Verb Types [h]	Conversation/Narrative	Toys, board game, picture sequence, wordless picture book	89%	90%	-	Yes
FVMC [e, g]	Narrative tell	Picture sequences (ENNI)	82%	90%	89%	Yes
Percent Grammatical C-units [g, l]	Narrative tell	Picture sequences (ENNI)	82%	90%	89%	Yes
10 SALT measures [i]	Conversation	Personal topics (SALT protocol)	80%	85%	-	No
Grammaticality & Utterance Length Instrument [j]	Narrative retell	Wordless picture book	83%	92%	-	No
7-year-olds						
Cohesive ties+ Grammaticality + Subordinate clauses/T-unit +Words/subordinate clause[k]	Narrative retell	Movie	-	-	98%	No
Percent Grammatical Utterances [l]	Narrative tell	Picture sequences (ENNI)	92%	88%	89%	Yes
FVMC [e]	Narrative tell	Picture sequences (ENNI)	85%	86%	86%	Yes
MLU (SUGAR) + Clauses per Sentence [f]	Conversation	Personal topics (SUGAR protocol)	97%	82%	-	Yes

Measure	Elicitation Task	Materials	Sensitivity	Specificity	Overall	Cutoff
8-year-olds						
Cohesive ties+ Grammaticality + Subordinate clauses/T-unit +Words/subordinate clause[k]	Narrative retell	Movie	-	-	98%	No
Errors per C-unit[g]	Narrative tell	Picture sequences (ENNI)	94%	80%	84%	Yes
Percent "restricted" utterances[m]	Narrative tell	Wordless picture book	83%	88%		Yes
Percent Grammatical Utterances[l]	Narrative tell	Picture sequences (ENNI)	88%	84%	85%	Yes
9-year-olds						
Percent Grammatical Utterances[l]	Narrative tell	Picture sequences (ENNI)	90%	90%	90%	Yes
Cohesive ties+ Grammaticality + Subordinate clauses/T-unit +Words/subordinate clause[k]	Narrative retell	Movie	-	-	98%	No
10-year-olds						
Cohesive ties+ Grammaticality + Subordinate clauses/T-unit +Words/subordinate clause[k]	Narrative retell	Movie	-	-	98%	No
Percent "restricted" utterances[m]	Narrative tell	Wordless picture book	83%	88%	-	Yes

Measure	Elicitation Task	Materials	Sensitivity	Specificity	Overall	Cutoff
African American English & Southern White English Speakers (4-6-year-olds)						
35 nonmainstream patterns [n]	Play	Toys (gas station, picnic/park, baby dolls, food, Legos, beads), picture scenes	87%	94%	90%	No
Southern White English Speakers (4-6-year-olds)						
Irregular past+Auxiliary DO+Irregular third+Infinitive TO+*Don't* Agreement [n]	Play	Toys (gas station, picnic/park, baby dolls, food, Legos, beads), picture scenes	87%	95%	-	No
Past tense (Strategic Scoring) [o]	Play	Toys (gas station set, picnic/park set, baby doll set), action pictures (visiting doctor's office; fishing, grocery shopping, washing a car)	89%	89%	89%	Yes

Note. Em dashes indicated data not reported. VP = Verb Phrase. NP = Noun Phrase. MLU = Mean Length of Utterance. FVMC = Finite Verb Morphology Composite. DSS = Developmental Sentence Scoring. SUGAR = Sampling Utterances and Grammatical Analysis Revised. SALT = Systematic Analysis of Language Transcripts. [a] Gavin et al., 1993. [b] Bedore & Leonard, 1998. [c] Gladfelter & Leonard, 2013. [d] Souto et al., 2014. [e] Guo et al., 2020. [f] Pavelko & Owens, 2019. [g] Guo & Schneider, 2016. [h] Fletcher & Peters, 1984. [i] Heilmann et al., 2010. [j] Castilla-Earls & Fulcher-Rood, 2018. [k] Liles et al., 1995. [l] Guo et al., 2019. [m] Hoffman, 2009. [n] Oetting & McDonald, 2001. [o] Oetting et al., 2021.

Quality of Evidence

The 15 publications reporting the measures in the previous section were examined for design features that indicate the quality or strength of the evidence using a checklist published by Dollaghan (2004). A one-gate design that recruits all participants from the same population is more likely to result in a participant sample that represents a continuum of ability or severity than a 2-gate design that recruits from different sources (e.g., TD from a local school and DLD from a clinic). Selection of a valid and accurate gold standard reference measure and blinded administration of the measure to all participants by independent examiners ensures that group assignment reflects accurate and objective classification of impairment status. Positive and negative likelihood ratios (LR+ and LR-) are diagnostic accuracy metrics that are less vulnerable to small participant samples, though there is less consensus on a recommended threshold (Klee et al., 2017). Intermediate values of above 4.0 for LR+ and below 0.4 for LR- are suggested as a minimum to be considered conclusive, with values of 10.0 for LR+ and .2 for LR- indicating high likelihood of accurate classification in the corresponding ranges (Dollaghan, 2004). Additionally, confidence intervals indicate how precise the diagnostic accuracy is likely to be across different groups. Clinical feasibility for LSA measures can include whether the cutoff value or regression equation for the measure(s) was reported, the number of measures that must be calculated, the length of the LSA transcript required, and access to required materials.

All studies used a 2-gate design or did not report this information clearly. Twelve studies used a clinical criterion (i.e., a previous diagnosis by an SLP and/or current enrollment in language therapy) as the gold standard reference measure either alone, in addition to a standardized test or a parent report measure, or confirmed by such a measure (see Table S2 in Appendix). A clinical criterion is widely regarded as an appropriate gold standard (Dollaghan & Campbell, 1998). Three studies used standardized tests as the reference measure, namely the TOLD-P2 or PLS-3 (Bedore & Leonard, 1998), the SPELT-3 (Castilla-Earls & Fulcher-Rood, 2018), and the DELV-NR (Oetting et al., 2021). Of the tests used in these studies, only the Test for Examining Expressive Morphology and DELV-NR have evidence of at least acceptable diagnostic accuracy with the age group and cutoff scores used (Eisenberg & Guo, 2013; Nitido & Plante, 2020; Spaulding et al., 2006). None of the studies clearly reported whether administration of measures was blinded.

Six studies reported likelihood ratios and two reported confidence intervals. We calculated these for the remaining studies based on true and false positives and negatives except for two studies that did not report adequate data. Positive and negative likelihood ratios all met the threshold to be considered diagnostically conclusive (i.e., >4.0 and <.4, respectively), but the confidence intervals of only two measures fell completely within this range for both ratios (FVMC for 5-year-olds in Guo et al., 2020; MLU-SUGAR + CPS in Pavelko & Owens, 2019). This likely reflects the small participant samples (Dollaghan, 2004), as nearly all studies included fewer than 25 participants per ability group within each age interval.

Eight studies reported the cut score(s) or the regression equation used in determining the diagnostic accuracy of the measure (see Table 1.3). The cutoff for individual measures from two additional studies could be derived based on the midpoint between group means (Gladfelter & Leonard, 2013; Souto, et al., 2014). Of these, analyses required calculation of only one to two LSA measures, except for the LARSP model which required three measures and child age. The analysis samples ranged from 33 to over 375 utterances long. Procedures that elicit 50 to 100 utterances will be more feasible in clinical practice (Heilmann, 2010; Pavelko et al, 2016) than those requiring more time-intensive elicitation or significantly longer samples (e.g., 1-hour protocol in Fletcher & Peters, 1984). The elicitation methods and materials are clearly described and publicly available for fidelity of implementation in a clinical setting for all but two studies.

Discussion

Many different language measures spanning different language domains have been analyzed for their accuracy in identifying children with DLD. While the body of evidence is far from complete, the extant data is substantial enough to focus future research efforts and offer some actionable guidance to clinicians. The most consistently useful measures tend to measure verb inflection accuracy, or at least include such a measure in a composite - a finding that is consistent both with the findings of prior reviews and with our understanding of morphosyntax as a core deficit of DLD. Our expanded scope for participant age revealed that the clinical utility of these measures extends beyond the preschool to early elementary range previously examined. The FVMC yielded greater than 90% diagnostic accuracy for 4- and 5-year-olds across at least three studies, and at least acceptable accuracy of 80% for slightly younger (as did variations of this measure, such as the PVT) and slightly older children. Measures of overall grammaticality (e.g., percent grammatical C-units, errors per C-unit, DSS Sentence Point) yielded consistently acceptable accuracy across ages 3 through 10 and evidence of good accuracy in some cases.

While our results reiterate previous findings that measures of length are not consistently adequate on their own, evidence from the composite models in our included studies shows they may enhance the accuracy of verb morphology measures, especially for certain age groups. For example, the models that achieved good accuracy of 90% or greater for the youngest participants were Bedore and Leonard's (1998) verb morphology composite combined with MLU and Gavin et al.'s (1993) model, which included verb phrase errors along with frequency of single word utterances and three-element noun phrases - arguably measures which reflect length - and a factor to account for age. The GLi, which combines a grammaticality measure with a categorical measure of length, yielded acceptable accuracy for 4- to 6-year-olds, and although more accurate measures are available for this age range, the GLi offers the advantage of more rapid administration using shorter samples and calculations that are easily done by hand - an appealing feature for both clinicians and researchers.

LSA has been specifically recommended as a culturally relevant assessment approach (Kraemer & Fabiano-Smith, 2017; Stockman, 1996), but only 5 studies identified for this review included speakers of nonmainstream English dialects or other languages in addition to English. Strategic scoring of regular past tense and a set of 5 dialect patterns can both yield acceptable accuracy for speakers of Southern White English, but a set of 35 dialect patterns is needed for speakers of Southern African American English. Given the number of variables required in the analysis, standardized tests and probes that have demonstrated comparable accuracy are likely to be more clinically feasible at this time while this line of research develops (Oetting et al., 2021).

While the evidence may be too limited to make specific recommendations for immediate clinical application (Oetting et al., 2021), findings that diagnostic accuracy of a given measure does not generalize across dialects underscore caution against using assessment measures with populations for which they have not been validated. Findings also illustrate the importance of adopting a *disorder within dialect* framework (Oetting et al., 2016), such as the finding that strategically scored regular past tense - a structure that might typically be disregarded as characteristic of language difference rather than evidence of disorder - was one of the best for differentiating impairment in speakers of certain dialects. Attention to the unique presentation of disorder within the context of linguistic variation is also relevant for speakers of English as an additional language (Bedore et al., 2018). Measures tested with this population fell short of adequately differentiating children with impairment, which is consistent with the previous meta-analysis of diagnostic accuracy of bilingual assessments (Dollaghan & Horner, 2011) and with guidance that assessing a child in both of their languages is the best approach (Gutierrez-Clellen & Simon-Cereijido, 2009).

Limitations

One limitation of the current study is that, although the search terms allowed for the inclusion of a wide age range, the actual age range represented in the reviewed studies is fairly narrow. A substantial number of LSA measures have been tested with children between age 3 and 6, but very few studies, which examined a limited selection of measures or composites, were available for children past the age of 9 and none for children older than 13. Considering that the accuracy of measures varies by age even among young children, as we see with FVMC, we cannot assume that “good” measures will still be useful if they have not been tested on children of that age. This mirrors the larger trend in speech-language pathology research, and so calls to expand research on adolescent language also apply in this case.

The quality of the evidence identified in this review also suggests limitations in the generalizability of diagnostic accuracy results to the larger population. When participant samples are small, single cases of misclassification can dramatically alter sensitivity and specificity values and potentially over- or underestimate actual diagnostic accuracy. Additionally, since most studies relied on a 2-gate design, diagnostic accuracy may be artificially high compared with a prospective, 1-gate community sample representing a broader range of performance, as clinicians are likely to encounter in practice. While some measures have cumulative evidence across studies to merit more confidence in the results (e.g., FVMC, PGU, MLU), those examined with a single study using a 2-gate design and/or a small sample require more caution, pending further studies. The findings of this review can guide SLPs in conducting LSA according to the best available evidence, though they should continue to use multiple sources of converging evidence for identification of DLD and stay apprised of how ongoing research informs recommendations for LSA measure selection and interpretation.

The current study limited the scope of the review to only English language sample data. This allowed for a more comprehensive synthesis of the patterns of findings within a single language. However, because clinical markers of DLD are language-specific (Leonard, 2014b), the diagnostic accuracy level and corresponding cutoff scores or equations found cannot be generalized to other languages, even if the measure can be readily applied (e.g., grammaticality). To facilitate best practices of assessing bilingual students in both of their languages using diagnostic LSA, future research should examine the diagnostic accuracy of LSA measures in languages other than English and compare the accuracy of measures cross-linguistically.

Implications for Future Research

A critical need for future studies to address is the coverage of accurate LSA measures based on age and linguistic variety. The evidence available indicates that the measures that best identify elementary-age children are not as sensitive to impairment at older ages, even when incorporating more developmentally appropriate measures such as syntactic complexity (Nippold et al., 2008). Additional research focused on early and late adolescents is needed to test a wider range of LSA measures and composites using developmentally sensitive elicitation tasks that are more likely to elicit group differences (Nippold et al., 2008). More research is also needed to identify valid and accurate LSA measures across diverse populations. The potential of acceptably accurate measures to achieve good diagnostic accuracy (e.g., for 6-year-olds) should also be explored through inclusion in a composite model or alternative methodology (e.g., different elicitation tasks, varying length of language sample, ROC analysis vs discriminant function).

While some measures have been examined across multiple studies using a variety of elicitation methods, such as the FVMC, others have yet to be replicated. Future studies should aim to validate extant findings while incorporating rigorous designs, such as larger participant samples and choosing current test versions that have good diagnostic accuracy as the reference standard, in order to identify LSA measures that are robust as well as accurate. These studies should also be sure to report information needed for practical application, namely cutoff scores. Implementation studies that explore the clinical feasibility of the protocols used in the existing evidence base are needed to inform practice-relevant methods for future diagnostic accuracy studies, as well as to identify the remaining barriers to routine use of LSA in clinical practice that dissemination of evidence alone does not overcome (Rabin & Brownson, 2017).

Clinical Application

Despite the limitations and gaps that remain to be addressed, SLPs can apply the findings of this review in current practice by incorporating the LSA measures identified as having evidence of clinical utility, albeit preliminary, into their assessments with similar clients. Clinicians can refer to Table 1.3 to identify the measure(s) that would provide the most accurate diagnostic classification for the client's age. For measures with an available cutoff score, Supplemental Material S2 includes a summary of procedures and interpretation guidelines (see Appendix). A software-specific tutorial on how to automatically generate each measure is beyond the scope of this study; however, they appear to be generally compatible with the functionality of popular programs using either embedded commands (e.g., MLU; see Pezold et al., 2020 Supplemental Material S2, p. 1) or custom codes (Pezold et al., 2020, Supplemental Material S1 Section 2, p. 5-6). Future tutorials should explore the options for coding transcripts and computing the measures highlighted in this review across different software programs to enable clinicians to take full advantage of computer-assisted LSA using the most efficient procedures.

Conclusion

This systematic review highlights the availability of several LSA measures and composites that can accurately differentiate monolingual mainstream English-speaking preschool and elementary-age children with DLD from those who are typically developing. Further research is needed, however, to identify measures that are useful with adolescents and speakers of diverse varieties of English and to both replicate and build upon previous findings in order to strengthen the evidence base for and clinical feasibility of diagnostic LSA. Nevertheless, findings of acceptable levels of diagnostic accuracy across multiple studies and measures reinforce recommendations to incorporate LSA as an informative, ecologically valid tool in clinical assessments, and clinicians can use the evidence reviewed here to guide and justify their interpretation of LSA results for diagnostic decisions.

References

References marked with an asterisk indicate studies included in this systematic review.

American Speech-Language-Hearing Association (n.d.). Spoken Language Disorders. (Practice Portal). Retrieved October 27, 2020, from https://www.asha.org/PRPSpecificTopic.aspx?folderid=8589935327

*Bedore, L. M., & Leonard, L. B. (1998). Specific language impairment and grammatical morphology: a discriminant function analysis. *Journal of Speech, Language, and Hearing Research*, *41*(5), 1185–1192.

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. DOI: 10.1044/2017_LSHSS-17-0027

Betz, S. K., Eickhoff, J. R., & Sullivan, S. F. (2013). Factors influencing the selection of standardized tests for the diagnosis of specific language impairment. . *Language, Speech, and Hearing Services in Schools*, *44*(2), 133–146. DOI: 10.1044/0161-1461(2012/12-0093)

Bishop, D. V., Snowling, M. J., Thompson, P. A., Greenhalgh, T., & the 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. DOI: 10.1111/jcpp.12721

Castilla-Earls, A., Bedore, L., Rojas, R., Fabiano-Smith, L., Pruitt-Lord, S., Restrepo, M. A., & Peña, E. (2020). Beyond scores: Using converging evidence to determine speech and language services eligibility for dual language learners. *American Journal of Speech-Language Pathology*, *29*(3), 1116–1132. DOI: 10.1044/2020_AJSLP-19-00179

*Castilla-Earls, A., & Fulcher-Rood, K. (2018). Convergent and divergent validity of the grammaticality and utterance length instrument. *Journal of Speech, Language, and Hearing Research*, *61*(1), 120–129. DOI: 10.1044/2017_JSLHR-L-17-0152

*Charest, M., Skoczylas, M. J., & Schneider, P. (2020). Properties of lexical diversity in the narratives of children with typical language development and developmental language disorder. *American Journal of Speech-Language Pathology*, *29*(4), 1866–1882. DOI: 10.1044/2020_AJSLP-19-00176

Costanza-Smith, A. (2010). The clinical utility of language samples. *Perspectives on Language Learning and Education*, *17*(1), 9–15. DOI: 10.1044/lle17.1.9

Crystal, D., Fletcher, P., & Garman, M. (1976). The grammatical analysis of language disability: A procedure for assessment and remediation. *The Grammatical Analysis of Language Disability: A Procedure for Assessment and Remediation*, *1*.

Dollaghan, C. A. (2004). Evidence-based practice in communication disorders: What do we know, and when do we know it? *Journal of Communication Disorders*, *37*(5), 391-400.

Dollaghan, C., & Campbell, T. F. (1998). Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research*, *41*(5), 1136-1146.

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. DOI: 10.1044/1092-4388(2010/10-0093)

*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. DOI: 10.1044/jshr.3903.643

Dunn, L. M., & Dunn, L. M. (1981). Peabody Picture Vocabulary Test – Revised: Manual for Forms L and M (PPVT-R). Circle Pines, MN: AGS.

Eisenberg, S. L., Fersko, T. M., & Lundgren, C. (2001). The use of MLU for identifying language impairment in preschool children. *American Journal of Speech-Language Pathology*, *10*(4), 323–342. DOI: 10.1044/1058-0360(2001/028)

*Eisenberg, S. L., & Guo, L. Y. (2013). Differentiating children with and without language impairment based on grammaticality. *Language, Speech, and Hearing Services in Schools*, *44*(1), 20–31. DOI: 10.1044/0161-1461(2012/11-0089)

Eisenberg, S.L., & Guo, L. Y. (2016). Using language sample analysis in clinical practice: Measures of grammatical accuracy for identifying language impairment in preschool and school-aged children. *Seminars in Speech and Language*, *37*(2), 106–116. DOI: 10.1055/s-0036-1580740

Evans, J. (1996). Plotting the complexities of language sample analysis: Linear and non-linear dynamical models of assessment. *Assessment of Communication and Language*, *6*, 207–256.

*Fletcher, P., & Peters, J. (1984). Characterizing language impairment in children. *Language Testing*, *1*(1), 33–49. DOI: 10.1177/026553228400100104

Fulcher-Rood, K., Castilla-Earls, A. P., & Higginbotham, J. (2018). School-based speech-language pathologists' perspectives on diagnostic decision making. *American Journal of Speech-Language Pathology*, *27*(2), 796–812. DOI: 10.1044/2018_AJSLP-16-0121

Fulcher-Rood, K., Castilla-Earls, A., & Higginbotham, J. (2019). Diagnostic decisions in child language assessment: findings from a case review assessment task. *Language, Speech, and Hearing Services in Schools*, *50*(3), 385–398. DOI: 10.1044/2019_LSHSS-

18-0044

*Gavin, W. J., Klee, T., & Membrino, I. (1993). Differentiating specific language impairment from normal language development using grammatical analysis. *Clinical Linguistics & Phonetics*, *7*(3), 191–206. DOI: 10.3109/02699209308985557

*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. DOI: 10.1044/1092-4388(2012/12-0100)

*Guo, L. Y., & Eisenberg, S. (2014). The diagnostic accuracy of two tense measures for identifying 3-year-olds with language impairment. *American Journal of Speech-Language Pathology*, *23*(2), 203–212. DOI: 10.1044/2013_AJSLP-13-0007

*Guo, L. Y., Eisenberg, S., Schneider, P., & Spencer, L. (2019). Percent grammatical utterances between 4 and 9 years of age for the Edmonton Narrative Norms Instrument: Reference data and psychometric properties. *American Journal of Speech-Language Pathology*, *28*(4), 1448-1462. DOI: 10.1044/2019_AJSLP-18-0228

*Guo, L. Y., Eisenberg, S., Schneider, P., & Spencer, L. (2020). Finite verb morphology composite between age 4 and age 9 for the Edmonton Narrative Norms Instrument: Reference data and psychometric properties. *Language, Speech, and Hearing Services in Schools*, *51*(1), 128-143. DOI: 10.1044/2019_LSHSS-19-0028

*Guo, L. Y., & Schneider, P. (2016). Differentiating school-aged children with and without language impairment using tense and grammaticality measures from a narrative task. *Journal of Speech, Language, and Hearing Research*, *59*(2), 317–329. DOI: 10.1044/2015_JSLHR-L-15-0066

Gutiérrez-Clellen, V. F., & Simon-Cereijido, G. (2009). Using language sampling in clinical assessments with bilingual children: Challenges and future directions. *Seminars in Speech and Language*, *30*(4), 234-245. DOI: 10.1055/s-0029-1241722

Heilmann, J. J. (2010). Myths and realities of language sample analysis. *Perspectives on Language Learning and Education*, *17*(1), 4. DOI: 10.1044/lle17.1.4

*Heilmann, J. J., Miller, J. F., & Nockerts, A. (2010). Using language sample databases. *Language, Speech, and Hearing Services in Schools*, *41*(1), 84–95. DOI: 10.1044/0161-1461(2009/08-0075)

*Hewitt, L. E., Hammer, C. S., Yont, K. M., & Tomblin, J. B. (2005). Language sampling for kindergarten children with and without SLI: Mean length of utterance, IPSYN, and NDW. *Journal of Communication Disorders*, *38*(3), 197–213. DOI: 10.1016/j.jcomdis.2004.10.002

*Hoffman, L. M. (2009). The utility of school-age narrative microstructure indices: INMIS and the proportion of restricted utterances. *Language, Speech, and Hearing Services in Schools*, *40*(4), 365-375. DOI: 10.1044/0161-1461(2009/08-0017)

Horton-Ikard, R. (2010). Language sample analysis with children who speak non-mainstream dialects of English. *Perspectives on Language Learning and Education*, *17*(1), 16–23. DOI: 10.1044/lle17.1.16

Klatte, I. S., van Heugten, V., Zwitserlood, R., & Gerrits, E. (2022). Language Sample Analysis in Clinical Practice: Speech-Language Pathologists' Barriers, Facilitators, and Needs. *Language, Speech, and Hearing Services in Schools*, *53*(1), 1-16.

*Klee, T., Gavin, W. J., & Stokes, S. F. (2017). Utterance length and lexical diversity in American- and British-English speaking children: What is the evidence for a clinical marker of SLI? In R. Paul (Ed.), *Language disorders from a developmental perspective: Essays in honor of Robin S. Chapman* (pp. 103–140). Psychology Press. DOI: 10.4324/9781315092041-4

Kraemer, R., & Fabiano-Smith, L. (2017). Language assessment of Latino English learning children: A records abstraction study. *Journal of Latinos and Education*, *16*(4), 1–10. DOI: 10.1080/15348431.2016.1257429

Leonard, L. B. (2014). Specific language impairment across languages. *Child Development Perspectives*, *8*(1), 1-5.

*Liles, B. Z., Duffy, R. J., Merritt, D. D., & Purcell, S. L. (1995). Measurement of narrative discourse ability in children with language disorders. *Journal of Speech, Language, and Hearing Research*, *38*(2), 415–425. DOI: 10.1044/jshr.3802.415

Manning, B. L., Harpole, A., Harriott, E. M., Postolowicz, K., & Norton, E. S. (2020). Taking language samples home: Feasibility, reliability, and validity of child language samples conducted remotely with video chat versus in-person. *Journal of Speech, Language, and Hearing Research*, *63*(12), 3982-3990. DOI: 10.1044/2020_JSLHR-20-00202

Miller, J. F., Andriacchi, K., & Nockerts, A. (2016). Using language sample analysis to assess spoken language production in adolescents. *Language, Speech, and Hearing Services in Schools*, *47*(2), 99–112. DOI: 10.1044/2015_LSHSS-15-0051

Miller, J. F., & Iglesias, A. (2008) Systematic Analysis of Language Transcripts (SALT), English & Spanish (Version 9) [Computer software]. Madison, Wisconsin: University of Wisconsin-Madison, Waisman Center, Language Analysis Laboratory.

*Moyle, M. J., Karasinski, C., Weismer, S. E., & Gorman, B. K. (2011). Grammatical morphology in school-age children with and without language impairment: A discriminant function analysis. *Language, Speech, and Hearing Services in Schools*, *42*(4), 550–560. DOI: 10.1044/0161-1461(2011/10-0029)

Nippold, M. A., Mansfield, T. C., Billow, J. L., & Tomblin, J. B. (2008). Expository discourse in adolescents with language impairments: Examining Syntactic Development. *American Journal of Speech-Language Pathology*, *17*(4), 356-366. DOI: 10.1044/1058-0360(2008/07-0049)

Nitido, H., & Plante, E. (2020). Diagnosis of developmental language disorder in research studies. *Journal of Speech, Language, and Hearing Research*, *63*(8), 2777–2788. DOI: 10.1044/2020_JSLHR-20-00091

Oetting, J. B., Gregory, K. D., & Rivière, A. M. (2016). Changing how speech-language pathologists think and talk about dialect variation. *Perspectives of the ASHA Special Interest Groups*, *1*(16), 28–37. DOI: 10.1044/persp1.SIG16.28

*Oetting, J B, & McDonald, J. L. (2001). Nonmainstream dialect use and specific language impairment. *Journal of Speech, Language, and Hearing Research*, *44*(1), 207–223. DOI: 10.1044/1092-4388(2001/018)

*Oetting, J. B., Rivière, A. M., Berry, J. R., Gregory, K. D., Villa, T. M., & McDonald, J. (2021). Marking of tense and agreement in language samples by children with and without specific language impairment in African American English and Southern White English: Evaluation of scoring approaches and cut scores across structures. *Journal of Speech, Language, and Hearing Research*, *64*(2), 491–509. DOI:10.1044/2020_JSLHR-20-00243

*Ooi, C. C.-W., & Wong, A. M.-Y. (2012). Assessing bilingual Chinese-English young children in Malaysia using language sample measures. *International Journal of Speech-Language Pathology*, *14*(6), 499–508. DOI: 10.3109/17549507.2012.712159

*Overton, C., Baron, T., Pearson, B. Z., & Ratner, N. B. (2021). Using free computer-assisted language sample analysis to evaluate and set treatment goals for children who speak African American English. *Language, Speech, and Hearing Services in Schools*, *52*(1), 31–50. DOI: 10.1044/2020_LSHSS-19-00107

*Pavelko, S. L., & Owens, R. E. (2019). Diagnostic accuracy of the sampling utterances and grammatical analysis revised (SUGAR) measures for identifying children with language impairment. *Language, Speech, and Hearing Services in Schools*, *50*(2), 211–223. DOI: 10.1044/2018_LSHSS-18-0050

Pavelko, S. L., Owens, R. E., Ireland, M., & Hahs-Vaughn, D. L. (2016). Use of language sample analysis by school-based SLPs: Results of a nationwide survey. *Language, Speech, and Hearing Services in Schools*, *47*(3), 246–258. DOI: 10.1044/2016_LSHSS-15-0044

Pawlowska, M. (2014). Evaluation of three proposed markers for language impairment in English: A meta-analysis of diagnostic accuracy studies. *Journal of Speech, Language, and Hearing Research*, *57*(6), 2261–2273. DOI: 10.1044/2014_JSLHR-L-13-0189

Pezold, M. J., Imgrund, C. M., & Storkel, H. L. (2020). Using computer programs for language sample analysis. *Language, Speech, and Hearing Services in Schools*, *51*(1), 103–114. DOI: 10.1044/2019_LSHSS-18-0148

Plante, E., & Vance, R. (1994). Selection of preschool language tests: A data-based approach. *Language, Speech, and Hearing Services in Schools*, *25*(1), 15. DOI: 10.1044/0161-1461.2501.15

Prath, S. (2018, June 19). *The How and Why of Collecting a Language Sample*. *ASHA Leader Live*. leader.pubs.asha.org/do/10.1044/the-how-and-why-of-collecting-a-language-sample/full/

Price, J. R., & Jackson, S. C. (2015). Procedures for obtaining and analyzing writing samples of school-age children and adolescents. *Language, Speech, and Hearing Services in Schools*, *46*(4), 277–293. DOI: 10.1044/2015_LSHSS-14-0057

Price, L. H., Hendricks, S., & Cook, C. (2010). Incorporating computer-aided language sample analysis into clinical practice. *Language, Speech, and Hearing Services in Schools*, *41*(2), 206–222. DOI: 10.1044/0161-1461(2009/08-0054)

Rabin, B. A., & Brownson, R. C. (2017). Terminology for dissemination and implementation research. In R.C. Brownson, G.A. Colditz, & E.K. Proctor (Eds.) *Dissemination and implementation research in health: Translating science to practice* (pp. 19-45). Oxford University Press.

Rojas, R., & Iglesias, A. (2009). Making a case for language sampling. *ASHA Leader*, *14*(3), 10. DOI: 10.1044/leader.FTR1.14032009.10

*Rudolph, J. M., Dollaghan, C. A., & Crotteau, S. (2019). The finite verb morphology composite: Values from a community sample. *Journal of Speech, Language, and Hearing Research*, *62*(6), 1813–1822. DOI: 10.1044/2019_JSLHR-L-18-0437

Selin, C. M., Rice, M. L., Girolamo, T., & Wang, C. J. (2019). Speech-language pathologists' clinical decision making for children with specific language impairment. . *Language, Speech, and Hearing Services in Schools*, *50*(2), 283–307. DOI: 10.1044/2018_LSHSS-18-0017

Scarborough, H. S. (1990). Index of productive syntax. *Applied Psycholinguistics*, *11*(1), 1–22. DOI: 10.1017/S0142716400008262

*Scheffel, D. L. (1997; February 19-22). *The language of negotiation: Comparing children with language based learning disabilities and children with normally developing language* [Paper presentation]. LDA International Conference, Chicago, IL, United States.

*Schneider, P., Hayward, D., & Dubé, R. V. (2006). Storytelling from pictures using the Edmonton narrative norms instrument. *Journal of Speech-Language Pathology and*

Audiology, *30*(4).

Shahmahmood, T. M., Jalaie, S., Soleymani, Z., Haresabadi, F., & Nemati, P. (2016). A systematic review on diagnostic procedures for specific language impairment: The sensitivity and specificity issues. *Journal of Research in Medical Sciences: The Official Journal of Isfahan University of Medical Sciences*, *21*, 2016-2021. DOI: 10.4103/1735-1995.189648

*Smyk, E. (2012). *Second language proficiency in sequential bilingual children with and without primary language impairment* [Undergraduate thesis, Arizona State University]. ASU Electronic Theses and Dissertations. https://hdl.handle.net/2286/R.I.15159

*Souto, S. M., Leonard, L. B., & Deevy, P. (2014). Identifying risk for specific language impairment with narrow and global measures of grammar. *Clinical Linguistics & Phonetics*, *28*(10), 741–756. DOI: 10.3109/02699206.2014.893372

Spaulding, T. J., Plante, E., & Farinella, K. A. (2006). Eligibility criteria for language impairment: Is the low end of normal always appropriate? *Language, Speech, and Hearing Services in Schools*, *37*(1), 61–72. DOI: 10.1044/0161-1461(2006/007)

Stockman, I. J. (1996). The promises and pitfalls of language sample analysis as an assessment tool for linguistic minority children. *Language, Speech, and Hearing Services in Schools*, *27*(4), 355–366. DOI: 10.1044/0161-1461.2704.355

Sylvan, L. (2014). Speech-language services in public schools: How policy ambiguity regarding eligibility criteria impacts speech-language pathologists in a litigious and resource constrained environment. *Journal of the American Academy of Special Education Professionals*, *2*, 7-23.

Chapter 2: English LSA Measures (Study 2)

Background

Only 8% of speech-language pathologists in the US are bilingual (ASHA, 2022), which significantly constrains the capacity to assess bilingual children - who make up over 20% of the student population nationally - in their home language. Testing these children only in English is common but problematic due to the overlap between characteristics of typical second language acquisition and clinical markers of developmental language disorder (DLD). This frequently leads to overidentification and disproportionate placement in special education (Sullivan, 2011). As awareness of the potential bias of English tests increases, many clinicians opt to delay testing as a means of avoiding misdiagnosis. However, this approach leads to underidentification in those settings, especially among early school-age children (Collins et al., 2014; Morgan et al., 2017), and fails to comply with federal regulations such as Child Find. There is a great need for assessment methods that can accurately identify DLD in bilingual children based on their English performance. This study seeks to add to such efforts that are currently underway by exploring the diagnostic accuracy of English narrative language sample measures for Spanish-English bilingual children.

It is difficult to disentangle DLD from typical bilingual development based on L2 English for two reasons: the linguistic patterns associated with each overlap, and mastery of morphosyntactic forms varies according to exposure to English. In English, clinical markers of DLD include verb tense and subject-verb agreement errors (Leonard, 2014b). Children with DLD are less efficient at extracting these grammatical rules from the language input they receive and need more time than their typical peers to master them.

Typically developing bilingual children also tend to produce tense and agreement errors more often than same-age monolingual peers (Paradis, 2008; Simón-Cereijido & Gutiérrez-Clellen, 2007). Since bilinguals' language input is divided between two languages, it takes longer to accumulate the "critical mass" of English input needed to master a particular form (Gathercole, 2007). On the surface, it is unclear whether these errors reflect second language acquisition processes, impairment, or both. Additionally, bilinguals' linguistic experience is quite heterogeneous. While bilingual children both with and without DLD follow a similar developmental progression in acquiring English grammar (Jacobson & Yu, 2018), there is significant variability in the timeline to master different forms based on the amount of English a child is exposed to (Bedore et al., 2012; Paradis, 2010). This further contributes to misidentification of DLD when a child that is typically developing but has had more limited experience in English demonstrates lower morphosyntactic accuracy than a child who has DLD but enough experience to have reached mastery. With multiple sources of variability, it is critical to shift from thinking about profiles of difference and disorder separately to considering how disorder manifests within second language variation (Bedore et al., 2018). Assessing language based on performance requires us to consider the dynamic interaction between language typology, experience, and ability.

English-Only Assessment Methods

Identification of English measures that can accurately identify DLD in bilingual children has the potential for broad impact on practice among the majority of SLPs who are monolingual and have limited access to support from speakers of the many languages represented in the US. Efforts to date have explored processing measures (e.g., nonword repetition), language domains with shared knowledge across languages (e.g., narrative macrostructure), modifications to test construction (e.g., tailoring items, adjusted cutoff scores), and dynamic assessment. Dynamic assessment of English narratives is one approach that has been shown to accurately identify DLD in bilinguals (e.g., Peña et al., 2014), while others such as narrative macrostructure (e.g., Andreou & Lemoni, 2020) and nonword repetition (e.g., Schwob et al., 2021) on their own do not meet clinical standards of diagnostic accuracy for this population. Tailoring English test items to different levels of English exposure also results in accurate classification of bilingual children (e.g., Bedore et al., 2018; Jasso et al., 2020). LSA, which is a familiar task that can be readily implemented and adapted in clinical practice, offers another promising approach if it can be validated for English-only assessment.

Prior Research on English LSA Measures for Bilinguals

A substantial body of research has been conducted on the diagnostic accuracy of LSA measures for monolingual speakers of mainstream English, but only two studies have previously explored composites of English LSA measures for identifying DLD in bilingual speakers. For Malaysian Cantonese-English speakers ages 3;8 to 5;11, a composite of MLU in words, a Malaysian English adaptation of IPSyn Total, and lexical diversity *D* fell short of acceptable (78% sensitivity and specificity; Ooi & Wong, 2012). For Spanish-English bilingual children ages 5;3 to 8, a composite of MLU in words, errors per T-unit, number of different words, and percent maze words yielded 83% overall diagnostic accuracy, but since the disaggregated metrics were not reported, findings should be cautiously interpreted as suggestive but not conclusive (Smyk, 2012).

In addition to limitations in reporting, several of the LSA measures examined in these studies were found to have inconsistent or inadequate diagnostic accuracy even with monolingual speakers. For example, children with DLD do not always score lower than typical peers on IPSyn scores (Hewitt et al., 2005), and including IPSyn with MLU and number of different words yielded only 74% diagnostic accuracy. Lexical diversity *D* with MLU and age had acceptable diagnostic accuracy for very young children aged 2 to 4 (86% sensitivity, 91% specificity; Klee et al., 2017), but semantic measures, such as type-token ratio and number of different words, had poor sensitivity of <45% for 4- to 9-year-olds (Charest et al., 2020). In contrast, measures of grammatical accuracy, such as percent grammatical utterances (PGU), were found in several studies to have acceptable to good accuracy for monolingual speakers across a wide age range (e.g., Guo et al., 2019; Hoffman, 2009), and MLU and subordination index have enhanced the diagnostic accuracy of grammaticality measures (e.g., Bedore & Leonard, 1998; Liles et al., 1995). Applying LSA measures with evidence of clinical utility for monolinguals to the language samples of bilingual speakers may yield better classification.

Discriminant function analysis is commonly used in studies of diagnostic accuracy, but one limitation of this approach is that classification is based on only one cutoff value. Receiver operator characteristic (ROC) curves test sensitivity and specificity across all cutoff values. By testing the performance at multiple thresholds, ROC analysis allows identification of a measure's maximum possible diagnostic accuracy. Multivariate ROC analysis (multiROC) offers the same advantage when examining a panel of LSA measures by determining the optimal combined thresholds of two or more measures combined in a logical expression defining case positivity (i.e., A and B, A or B, etc.; Shultz, 1995). In cases where a panel of measures outperforms single measures, multiROC results will yield improved classification accuracy over univariate ROC and linear discriminant analysis (Wu et al., 2013).

Variation in relative exposure to each language has also not been adequately accounted for in previous studies. Studies have either recruited near-monolingual speakers, which limits the generalizability of findings, or pooled the entire participant sample in analyses, which may have attenuated the estimates of diagnostic accuracy (Janes & Pepe, 2008). Examining a continuum of bilingual experience and analyzing its effect on the diagnostic performance of LSA measures would provide more ecologically valid evidence for clinical application. Current language exposure provides a useful metric for quantifying bilingual experience and predicting language performance (Bedore et al., 2012). Covariate-adjusted ROC analysis (AROC) can reduce bias in classification accuracy estimates and reveal if applying different thresholds for different populations - in this case, different levels of English exposure - improves the performance of a measure (Janes et al., 2009).

Purpose

The current study builds on previous research by focusing on LSA measures with evidence of good diagnostic accuracy, exploring their usefulness with a bilingual population representing a continuum of language experience, and testing the performance of each measure at various cut points to identify its maximum accuracy level. Identifying English-only assessment practices that accurately classify language ability in bilingual children would greatly improve the validity of language assessments in settings where testing in the home language is not feasible, benefitting both the monolingual practitioner and the bilingual child. To that end, the following research questions were addressed:

1) What is the optimal classification accuracy for identifying DLD in Spanish-English bilingual 5- and 6-year-olds using percent grammatical utterances, errors per C-unit, MLU in words, and subordination index calculated from English narratives?

2) Is classification accuracy improved by adjusting for language exposure?

3) Is classification accuracy improved by using a combination of these measures?

Methods

Data Source

This study involved secondary analysis of existing language sample data drawn from two projects: Development of a Test for Hispanic Children in the US (DTHC) and Diagnostic Markers of Language Impairment (DM). For a detailed description of the participants and data collection, see Gutiérrez-Clellen et al. (2006), Gutiérrez-Clellen & Simón-Cereijido (2007), and Peña et al., (2018) for DTHC and Gillam et al., (2013) and Peña et al. (2011) for DM.

The goal of the DTHC project was to develop and validate a new diagnostic language test for the identification of DLD in Spanish-English speakers. 756 participants aged 4;0-6;11 were recruited via a one-gate design from school districts in California, Texas, and Pennsylvania serving primarily low-income students. The goal of the DM project was to examine clinical markers of DLD in bilingual children. Participants were recruited via a one-gate design across 12 elementary schools from three districts in Texas and Utah serving a high number of bilingual Latinx students, and 168 participants aged 5;0-6;5 completed follow-up testing as part of the longitudinal component of the project.

Current Study Participants

Participants were selected for the current study from the larger samples who were between age 5;0 and 6;11, were not missing data for determining language exposure or ability status, and produced a language sample in English even if they did not in Spanish. Table 2.1 summarizes participant characteristics. Language ability was classified as TD or DLD using standard scores on the Bilingual English-Spanish Assessment (BESA; Peña et al., 2018) Language Index Composite score based on a cutoff of 85 (-1SD), which has acceptable diagnostic accuracy for 5-year-olds (88% sensitivity/85% specificity) and good accuracy for 6-year-olds (96% sensitivity/94% specificity). Fifty-seven participants were classified as having DLD, of which two were excluded due to inadequate samples (i.e., no analyzable utterances of more than one word, no utterances in English). The remaining 55 were matched with TD participants from the same source project based on language exposure within 2% and age within 8 months[1]. A TD match could not be identified for nine DLD participants.

The final participant sample included 92 participants, with 28 participants from the DTHC study and 64 from the DM study. English exposure ranged from 20 to 100% (*M*=63.4, *SD*=21.8). Participants were also assigned categorically to the following language exposure

[1] All pairs but 3 were matched within 6 months of age

groups: high English exposure for those with 70% or more English exposure, balanced exposure for 30-69%, and high Spanish exposure for less than 30%.

Table 2.1

Study 2 Participant Characteristics

	Combined			DTHC			DM		
	Total	TD	DLD	Total	TD	DLD	Total	TD	DLD
N	92	46	46	28	14	14	64	32	32
Age in months (*SD*)	69.6 (5.44)	69.5 (5.38)	69.7 (5.56)	72.8 (6.47)	72.9 (6.99)	72.6 (6.16)	68.2 (4.28)	68.3 (4.82)	68.0 (3.73)
English exposure									
Mean (%)	63.4	63.4	63.3	79.9	79.9	79.9	56.1	56.2	56.0
Range (%)	20-100	20-100	20-100	22-100	23-100	22-100	20-85	20-84	20-85
Exposure group (*N*)									
English	32	16	16	22	11	11	10	5	5
Balanced	50	25	25	4	2	2	46	23	23
Spanish	10	5	5	2	1	1	8	4	4

Note. DTHC = Development of a Test for Hispanic Children in the US participants. DM = Diagnostic Markers of Language Impairment participants. TD = typically developing. DLD = developmental language disorder.

Materials

BIOS. Relative exposure to and use of Spanish and English was measured using the Bilingual Input-Output Survey (BIOS; Peña et al., 2018). The BIOS is a structured interview completed with a parent and teacher. Interviewees provide an hour-by-hour report of which language(s) a child currently hears and speaks, which is calculated into an overall percentage of input-output in each language, as well a year-by-year history of exposure to the two languages at home and school.

Language Samples. A story retell procedure was used to elicit the language samples across both source projects. Participants listened to a story told by the examiner while looking at a wordless picture book, then retold the same story while again looking at the book. DTHC samples were elicited were elicited using *One Frog Too Many* (Mayer, 1975), and DM samples were elicited using the books *Frog On His Own* (Mayer, 1973) and *A Boy, a Dog, and a Frog* (Mayer, 1967), which were counterbalanced by target language. All samples were audio-recorded and transcribed using Systematic Analysis of Language Transcripts (SALT; Miller & Chapman, 2008) software by a trained research assistant. Utterances were segmented into C-units, and mazes (i.e., word repetitions, filled pauses, etc.) were marked. Verbs were linked to their word roots. Morphemes were marked, which in English include plural -*s,* 3rd person singular -*s*, possessive -*'s*, regular past -*ed*, progressive -*ing,* past participle -*en*, and contracted words (e.g., -*n't*, *'ve*, etc.). Transcripts were checked by a second research assistant, and any discrepancies were resolved by an independent transcriber (see Bedore et al., 2010 for detailed procedures).

Procedures

Language Sample Coding. For the current study, transcripts were additionally coded for errors and subordination. The grammaticality of each utterance was judged in isolation, without reference to the story details or the child's previous utterances. Utterances that were judged to be ungrammatical were coded with the least number of changes needed to make a grammatical utterance. Errors were coded according to SALT conventions (see Table 2.2), which include morpheme omissions (e.g., omitted plural, omitted tense morpheme), word omissions (e.g., omitted preposition, omitted verb), word-level substitutions (e.g., pronoun case substitution, subject-verb agreement error), extraneous words, and word order errors. SALT conventions do not include coding of semantic errors. Errors related to pronouns (i.e., gender, animacy) or verb tense were determined based on consistency within an utterance. Each utterance was coded for the number of clauses it contains according to SALT conventions for subordination index (i.e., clauses per utterance). These coding procedures align with previous studies (e.g., Bedore et al., 2010).

Table 2.2

Coding Scheme for English Errors

Error Type	Code	Example
Morpheme omission		
Plural	WORD/*s	“a lot of frog” = a lot of frog/*s
Possessive -s	WORD/*z	“the boy pet” = the boy/*z pet
Tense markers		
Regular past -ed	WORD/*ed	“He call the frog” = He call/*ed the frog
3rd person singular -s	WORD/*3s	“The frog like the boy” = The frog like/*3s the boy
Progressive -ing	WORD/*ing	“was chase” = was chase/*ing
Word omission	*WORD	“Fell the window” = fell *out the window
Subject		
Article		
Preposition		
Verb		
Word-level substitution	[EW:WORD]	“Him fell.” = Him[EW:he] fell.
Pronoun case		
Subject-verb agreement		
Utterance-level errors		
Word order	[WO]	“The frog big left” = The frog big left [WO].
Extraneous words	[EW]	He had was jumping. = He had[EW] was jumping.

Training & Reliability. Coding for all transcripts was completed by the first author and trained research assistants (RA). RAs included an undergraduate student, two master's level graduate students in speech-language pathology, and one practicing SLP. All were bilingual Spanish-English speakers. RAs completed relevant self-paced online courses provided on the SALT website, which included coding of 3 practice transcripts. They met with the first author to review their practice transcripts and the coding guidelines for the current study. To establish initial inter-rater reliability, RAs coded at least 10 transcripts from a separate dataset and continued coding additional transcripts until reaching at least 85% agreement on 2 transcripts. To determine inter-rater reliability for coding of the current dataset, 20% of total samples were reviewed by a second coder using a consensus procedure of reviewing the original coding and noting disagreements, as reported by Guo et al. (2019). Discrepancies were discussed to reach agreement. Interrater reliability was 85%.

Measures

The SALT software was used to generate the following measures: mean length of utterance in words (MLUw), subordination index, total utterances containing error codes, total utterances, total morpheme omissions, total word omissions, total word codes, and total utterance codes. Percent grammatical utterances (PGU) was calculated as the inverse of the total utterances containing error codes. Errors per C-unit were calculated by dividing the sum of all error types (total morpheme omissions, total word omissions, total word codes, total utterance codes) by the total number of utterances. See Table 2.3 for a summary of these procedures. The following utterances were excluded from analyses: abandoned, unintelligible, single word, and utterances containing code-switching.

Table 2.3

Procedures for Calculating English LSA Measures

Measure	Calculation	Method/Source
Mean length of utterance in words	Total number of words in the sample divided by total number of utterances/C-units in the sample	SALT SMR: MLU-w
Errors per C-unit	Total number of omissions, substitutions, and word order errors coded in the sample divided by total number of utterances/C-units in the sample	Hand-calculated using • SALT SMR: Total Omitted Morphemes • SALT SMR: Total Omitted Words • SALT SMR: Total Word Codes • SALT SMR: Total Utterance Codes • SALT SMR: Total Utterances
Percent grammatical utterances	Total utterances with no coded errors divided by total utterances in the sample, converted to a percentage	Calculated as the inverse of: • SALT SMR: % Utterances Containing Error Codes
Subordination Index	Total clauses (independent and subordinate counted separately) in the sample divided by total utterances in the sample	SALT SMR: Subordination Index

Note. SMR = Standard Measures Report.

Analytic Strategy

The first research question explores the optimal classification accuracy for identifying DLD in Spanish-English bilingual 5- and 6-year olds using four LSA measures (PGU, errors per C-unit, MLUw in words, and subordination index) calculated from English narratives when adjusting for language exposure. To address this question, descriptive statistics including group means and standard deviations were calculated to examine the distribution of each LSA measure. Correlations between English exposure and LSA measures were calculated to verify appropriateness for inclusion as a covariate in subsequent analyses (Janes et al., 2009). Independent samples *t*-tests were run to verify DLD versus TD group differences on each LSA measure. Analyses were performed using jamovi (The jamovi project, 2024).

To examine classification accuracy, each LSA measure showing a statistically significant difference between TD and DLD participants was entered into a pooled receiver operator characteristic (ROC) analysis with the optimal cut point determined using the Youden index (Youden, 1950), which has been used in previous studies of diagnostic accuracy for identification of DLD (Oetting et al., 2021; Redmond et al., 2019). The following indicators of classification accuracy were generated and evaluated against established criteria for clinical usefulness: area under the curve (AUC; Youngstrom, 2014), sensitivity and specificity (Plante & Vance, 1994), positive and negative likelihood ratios (Dollaghan, 2004), and 95% confidence intervals. Analyses were performed using the ROCnReg package for R Studio, and curves were plotted using SPSS version 29.0.0.0.

To examine whether accounting for language exposure improves classification accuracy, LSA measures were then entered into covariate-adjusted ROCs with English exposure as a continuous variable included as a covariate (AROC; Janes & Pepe, 2008). Indicators of classification accuracy were again generated (i.e., AUC, sensitivity and specificity, positive and negative likelihood ratios, 95% confidence intervals) and evaluated against the pooled ROC and established criteria for clinical usefulness. Analyses were performed using the ROCnReg package for R Studio, and curves were plotted using SPSS version 29.0.0.0.

The third research question explored whether classification accuracy is improved by using a combination of the measures of interest. To evaluate whether combining one or more LSA measures improves on their individual classification accuracy, LSA measures were entered into multivariate receiver operator characteristic curves (multiROCs; Schultz, 1995). Indicators of classification accuracy were again generated (i.e., AUC, sensitivity and specificity, positive and negative likelihood ratios, 95% confidence intervals) and evaluated against the pooled ROC and established criteria for clinical usefulness. Analyses were performed using the multipleROC package for R Studio, and curves were plotted using SPSS version 29.0.0.0.

Results

Descriptive Analyses

Table 2.4 summarizes the descriptive statistics for the four LSA measures of interest - percent grammatical utterances (PGU), errors per C-unit, mean length of utterance in words (MLUw), and subordination index - as well as additional characteristics of participants' language samples, including total utterances, total words, and semantic measures. Bivariate correlations are presented in Table 2.5, and scatterplots of each LSA measure with English exposure are shown in Figures 2.1 through 2.4.

Table 2.4

Descriptive Statistics for English Language Sample-Derived Measures

				95% CI				
	Ability	*N*	Mean	Lower	Upper	*SD*	Min	Max
Total Utterances	DLD	46	32.80	27.50	38.10	17.83	4	100
	TD	46	29.39	26.60	32.20	9.55	10	52
Total Words	DLD	46	176.80	143	210	112.27	8	559
	TD	46	179.63	158	201	71.45	45	336
# Different Words	DLD	46	56.35	50.00	62.70	21.54	4	104
	TD	46	62.20	56.80	67.60	18.28	20	110
Type-Token Ratio	DLD	46	0.37	0.34	0.40	0.10	0.19	0.74
	TD	46	0.37	0.35	0.40	0.09	0.21	0.64
PGU	DLD	46	0.48	0.42	0.53	0.18	0.18	0.89
	TD	45	0.63	0.56	0.69	0.23	0.24	0.94
Errors per C-unit	DLD	46	0.68	0.59	0.76	0.29	0.11	1.45
	TD	45	0.46	0.37	0.55	0.30	0.06	1.05
MLUw	DLD	46	5.15	4.83	5.48	1.10	2.00	6.98
	TD	45	5.98	5.69	6.26	0.95	3.55	7.83
Subordination Index	DLD	46	0.93	0.87	1.00	.21	0.00	1.23
	TD	45	1.05	1.01	1.09	.13	0.70	1.43

Note. PGU = Percent Grammatical Utterances. MLUw = Mean Length of Utterance in words. DLD = Developmental Language Disorder. TD = typically developing.

Table 2.5

Bivariate Correlations between Study 2 Participant Characteristics and LSA Measures

		Ability	Exposure	Age	PGU	EPC	MLUw	SI
Ability	Pearson's r	—						
	p	—						
Exposure	Pearson's r	.003	—					
	p	.98	—					
Age	Pearson's r	-.01	.34 **	—				
	p	.89	.001	—				
PGU	Pearson's r	.34 ***	.38 ***	.21 *	—			
	p	< .001	< .001	.04	—			
EPC	Pearson's r	-0.35 ***	-0.34 **	-.20	-.94 ***	—		
	p	< .001	.001	.05	< .001	—		
MLUw	Pearson's r	0.38 ***	.25 *	.28 **	.35 ***	-.35 ***	—	
	p	< .001	.02	.01	< .001	< .001	—	
SI	Pearson's r	.31 **	.34 ***	.19	.28 **	-.25 *	.68 ***	—
	p	.003	< .001	.08	.01	.02	< .001	—

Note. Exposure = % English exposure. Age = age in months. PGU = Percent Grammatical Utterances. EPC = errors per C-unit. MLUw = Mean length of utterance in words. SI = Subordination Index.

* p < .05, ** p < .01, *** p < .001

Figure 2.1

Percent Grammatical Utterances by English Exposure

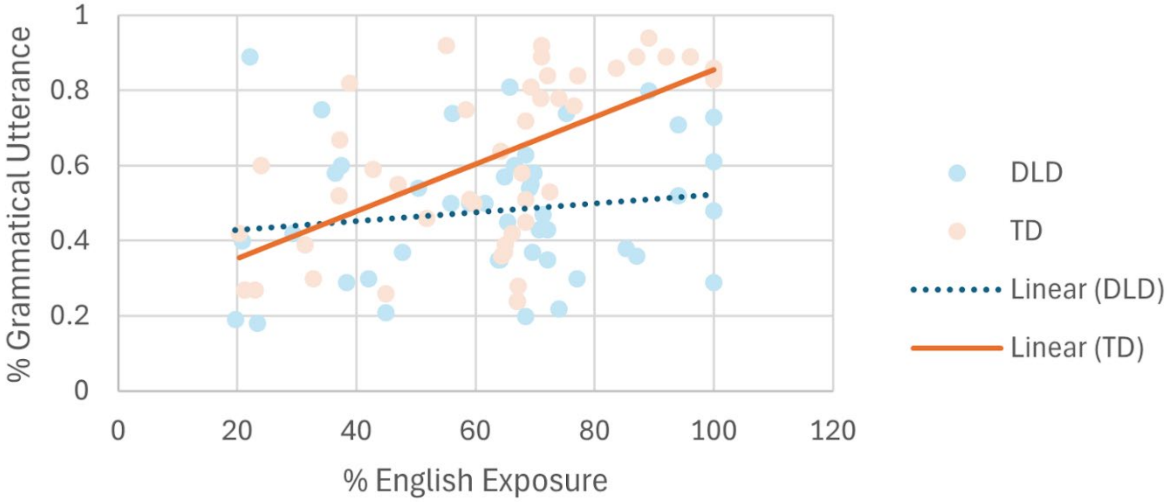


Figure 2.2

Errors per C-unit by English Exposure

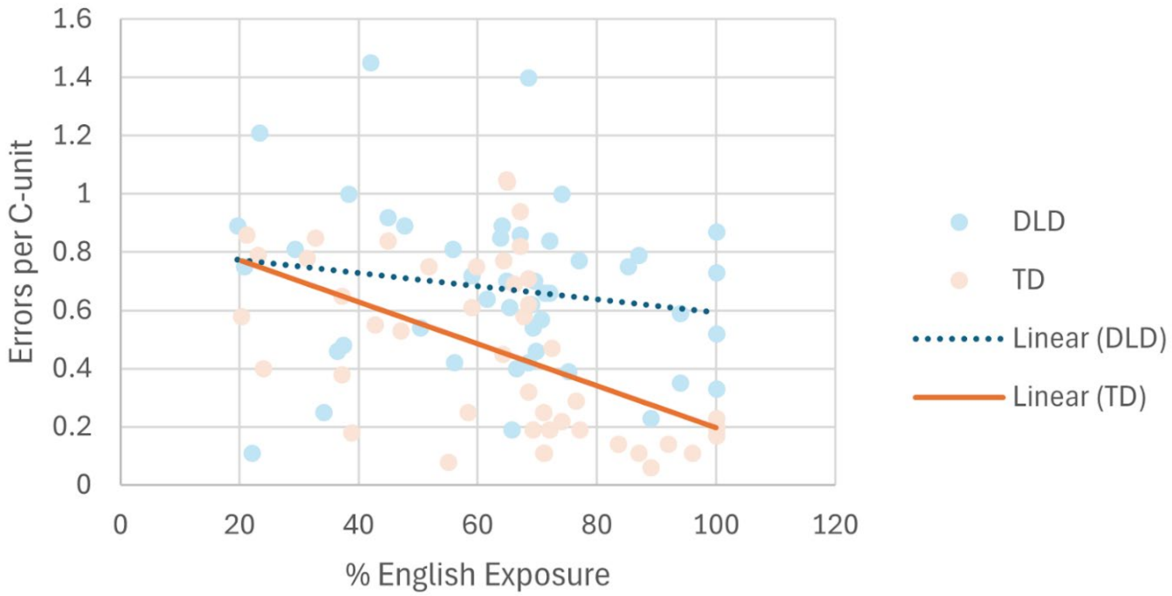


Figure 2.3

Mean Length of Utterance in Words by English Exposure

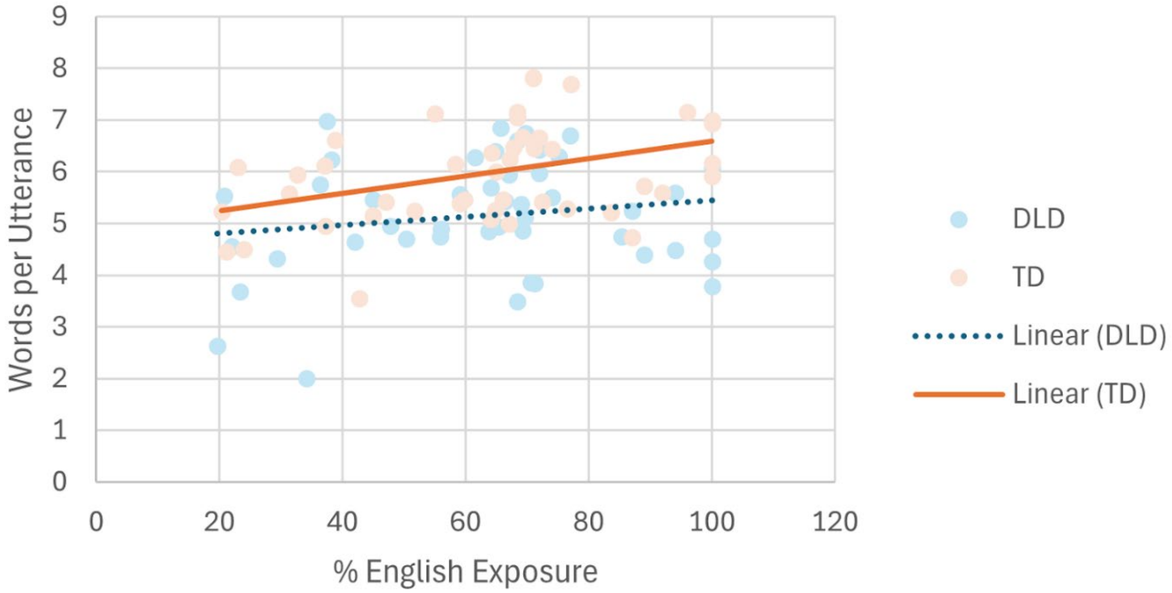
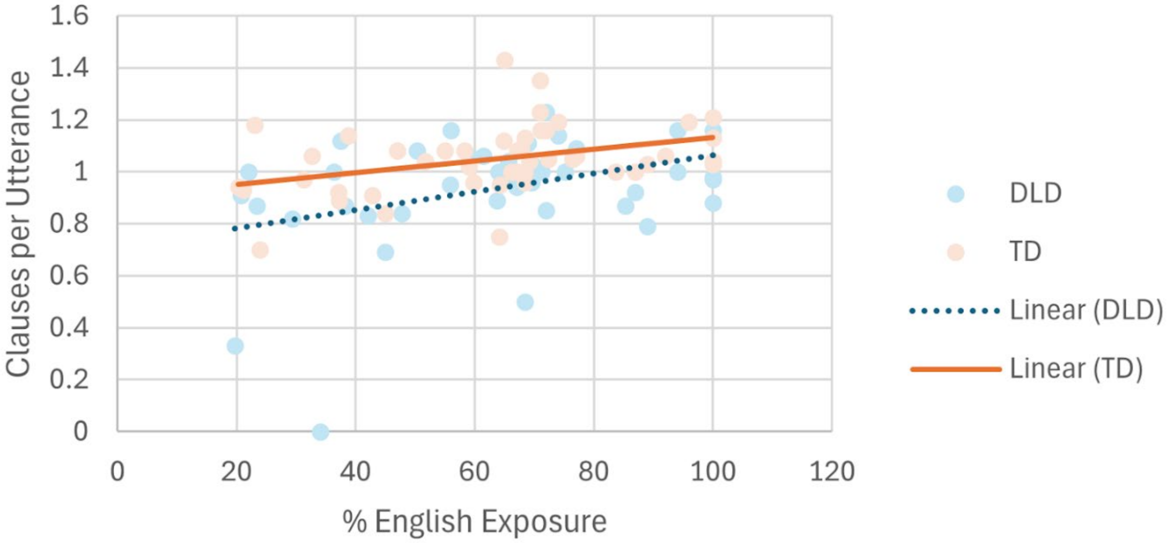


Figure 2.4

Subordination Index by English Exposure



Language exposure and ability were each moderately correlated with all four diagnostic LSA measures (positively with PGU, MLUw, and subordination and negatively with errors per C-unit) (*r* = .25-.38; *p* = <.001-.02). Age was moderately correlated with MLUw (*r* = .27, *p* = .01), PGU (*r* = .21, *p* = .048) and English exposure (*r* = .33, *p* = .001). Correlations between LSA measures were all statistically significant with moderate to large effect sizes (*r* = -.94-.68; *p* = <.001-.02). Length of the language samples measured in words or utterances was comparable between ability groups (*p* = .83 and *p* = .28). For all four diagnostic LSA measures, the TD group outperformed the DLD group, and differences were statistically significant based on independent samples *t*-tests (see Table 2.6).

Table 2.6

Independent Samples T-Tests for Age and English LSA Measures

	t	**df**	***p***	***d***
Age (months)	0.13	89.9	.89	.03
Total Utterances	1.15	68.8	.26	.24
Total Words	-0.14	76.3	.89	-.03
PGU	-3.42	86.5	< .001	-.71
EPC	3.51	89.9	< .001	.73
MLUw	-3.84	88.1	< .001	-.80
SI	-3.08	75.3	.003	-.64

Note. PGU = percent grammatical utterances. EPC = errors per C-unit. MLUw= mean length of utterance in words. SI = subordination index. $H_a\ \mu_0 \neq \mu_1$

Diagnostic Accuracy of LSA Measures With and Without Covariate Adjustment

Pooled and covariate-adjusted ROC results are presented in Table 2.7 for the entire participant sample, and Figures 2.5 through 2.9 show the corresponding plotted curves. Pooled ROC results for the high English and balanced exposure groups are presented in Table 2.8, and Figures 2.10 and 2.11 show the corresponding plotted curves.

Table 2.7

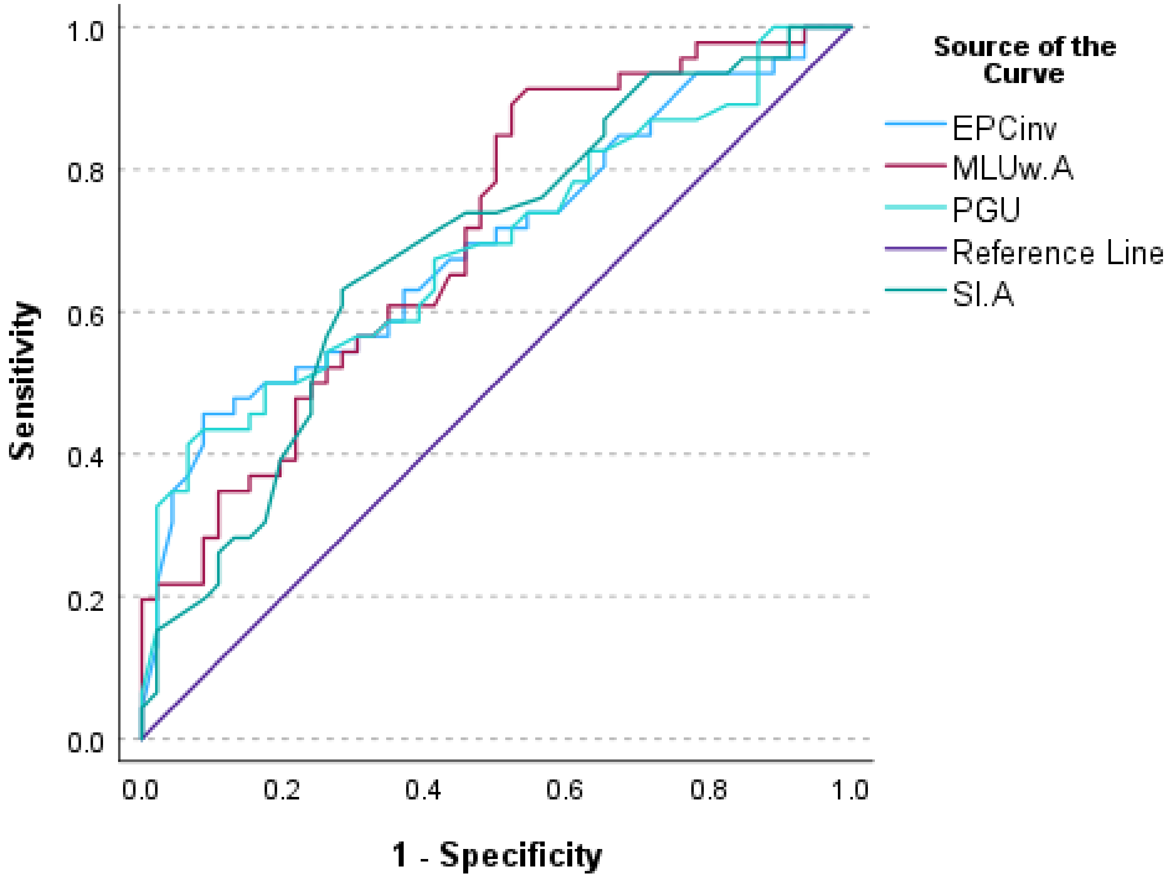
Pooled and Covariate-adjusted ROC Results in English

	AUC	Cutoff	Sensitivity	Specificity	LR+	LR-
Pooled						
PGU	.69 [.58, .79]	74.50%	**91%**	43%	1.60	0.21
EPC	.69 [.59, .80]	.32	**91%**	46%	1.69	0.20
MLUw	.71 [.60, .81]	4.95	48%	**89%**	4.36	0.58
SI	.69 [.57, .80]	1.00	72%	63%	1.95	0.44
AROC						
PGU	.70 [.59, .81]	-	-	-	-	-
EPC	.73 [.61, .83]	-	-	-	-	-
MLUw	.71 [.60, .81]	-	-	-	-	-
SI	.68 [.56, .78]	-	-	-	-	-

Note. AUC = area under the curve. LR+ = positive likelihood ratio. LR- = negative likelihood ratio. PGU = Percent Grammatical Utterances. EPC = errors per C-unit. MLUw = Mean length of utterance in words. SI = Subordination Index. Metrics that reached the 80% threshold for sensitivity and/or specificity are in bold text. Em dash = not applicable.

Figure 2.5

Pooled ROC Curves for Individual English LSA Measures



Note. EPCinv = errors per C-unit. MLUw.A = mean length of utterance in words. PGU = percent grammatical utterances. SI.A = subordination index.

Figure 2.6

Pooled & Covariate-Adjusted ROC Curves for English Percent Grammatical Utterances (PGU)

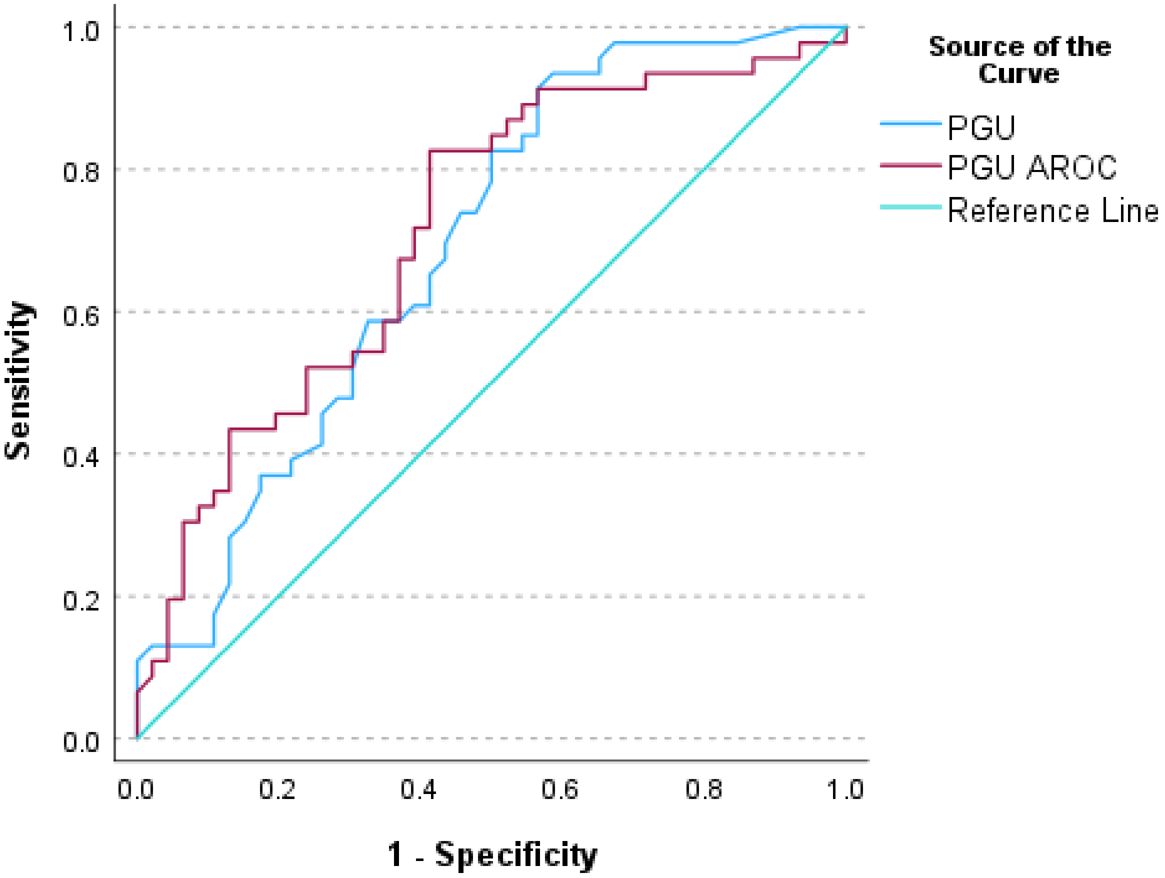


Figure 2.7

Pooled and Covariate-Adjusted ROC Curves for English Errors per C-unit (EPC)

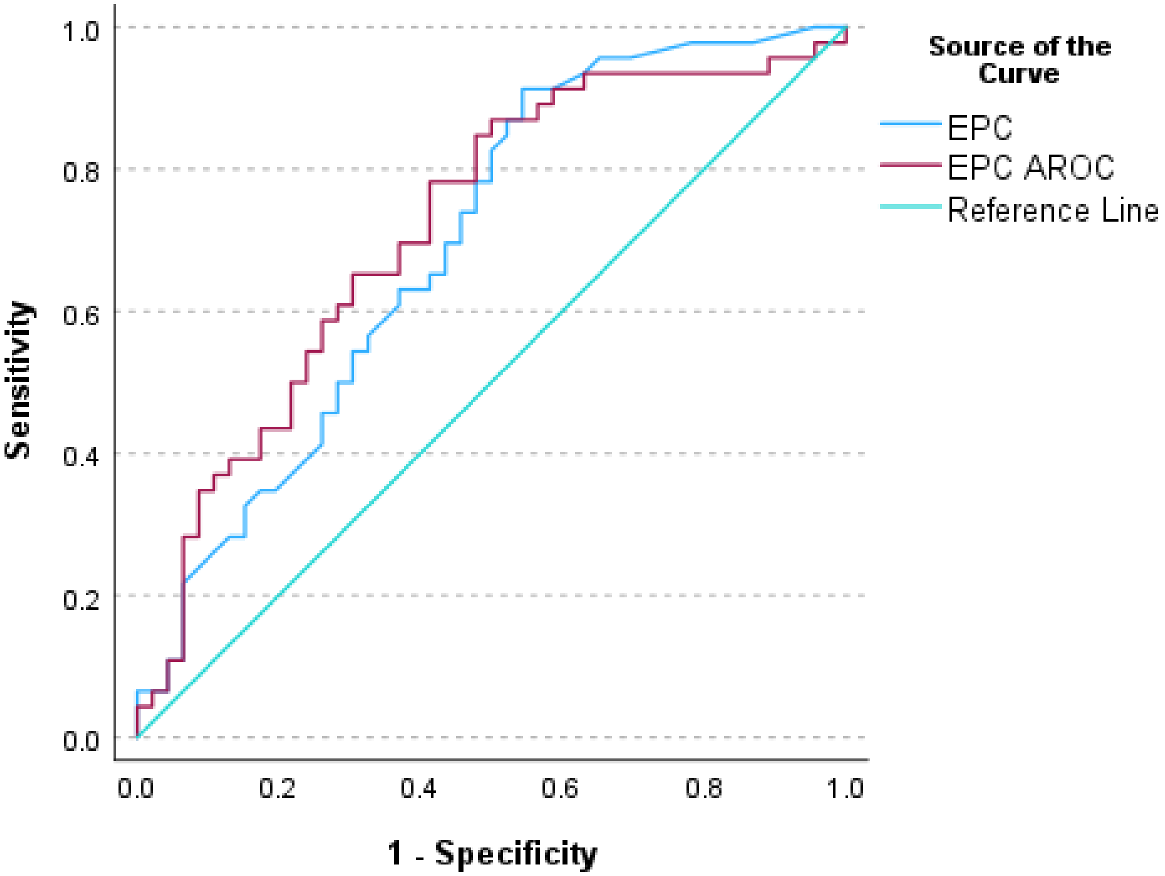


Figure 2.8

Pooled & Covariate-Adjusted ROC Curves for English Mean Length of Utterance in Words (MLUw)

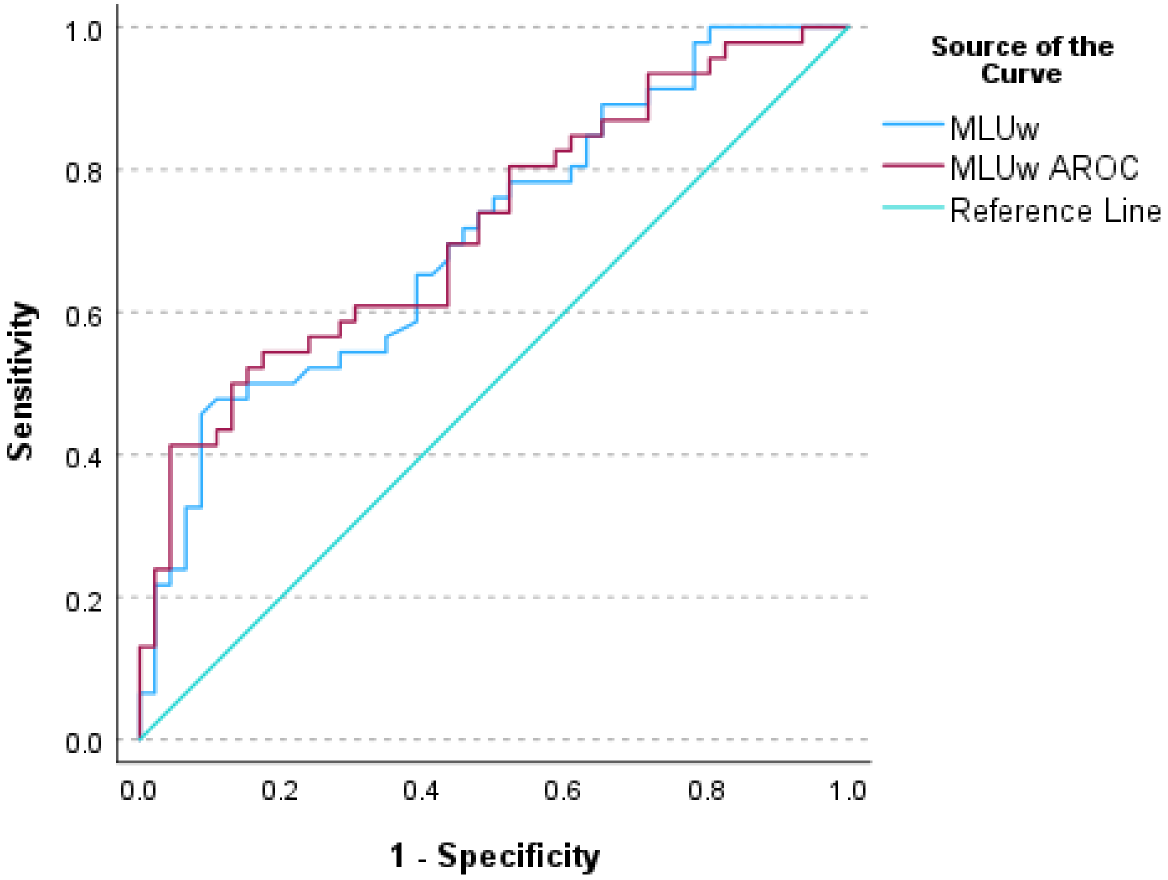


Figure 2.9

Pooled and Covariate-Adjusted ROC Curves for English Subordination Index (SI)

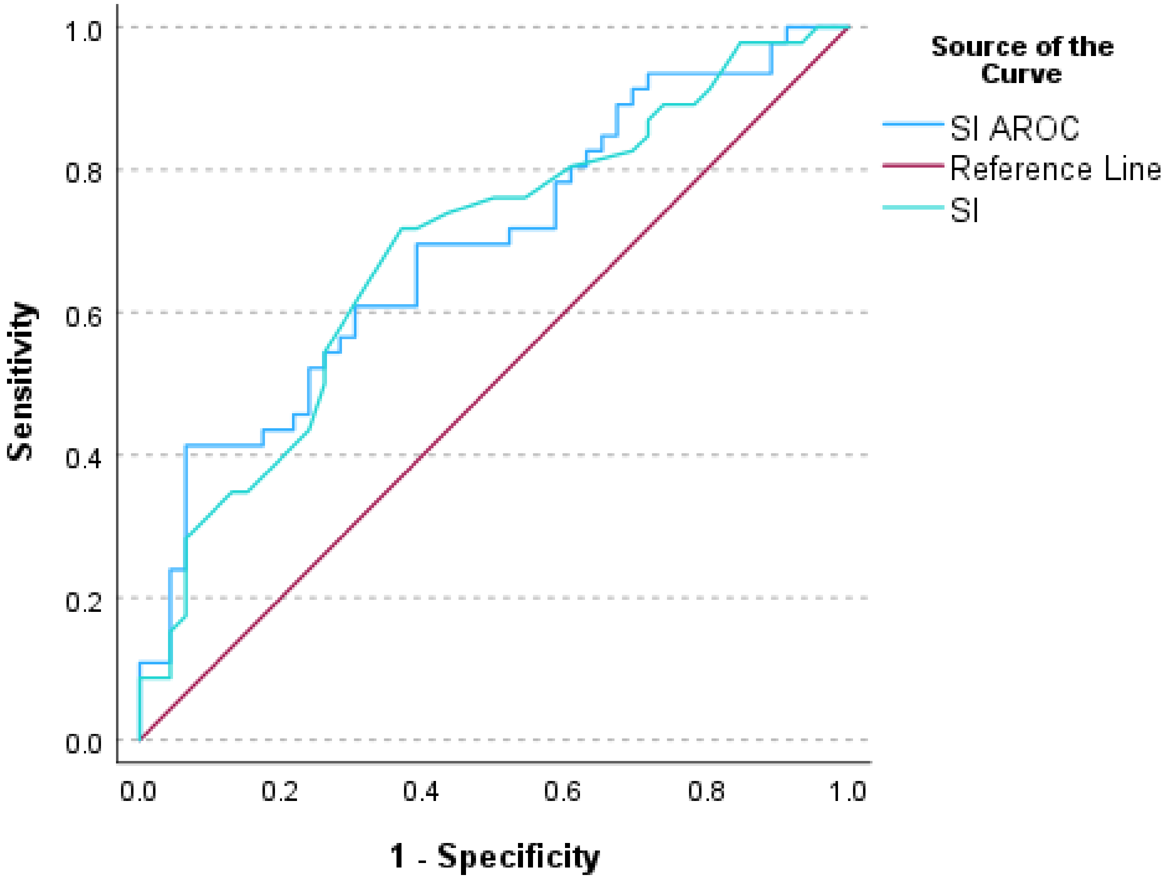


Table 2.8

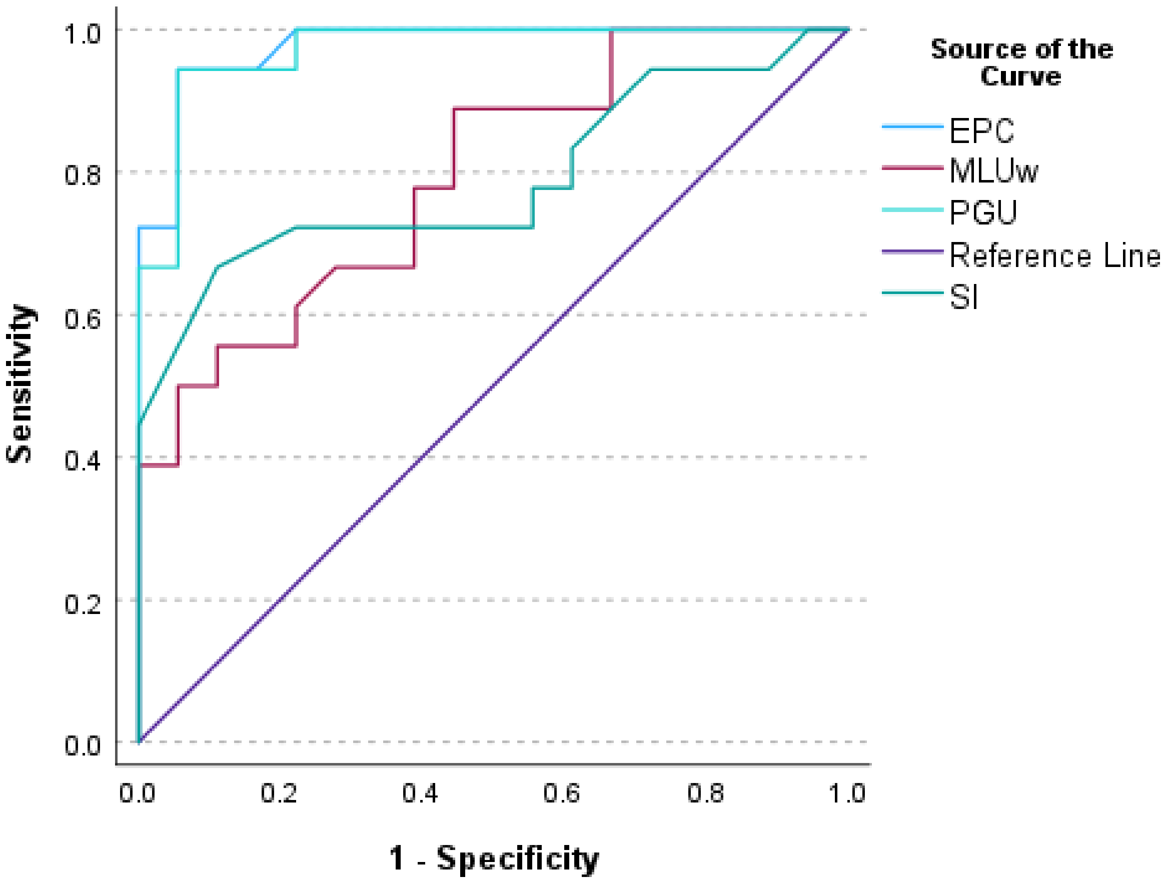
Pooled ROC Results in English by Exposure Group

	AUC	Cutoff	Sensitivity	Specificity	LR+	LR-
≥70% English Exposure						
PGU	.97 [.92, 1.0]	74%	**94%**	**94%**	15.67	0.06
EPC	.98 [.92, 1.0	0.29	**94%**	**94%**	15.67	0.06
MLUw	.82 [.66, .94]	5.59	71%	71%	2.45	0.41
SI	.79 [.59, .94]	1.00	71%	**88%**	5.92	0.33
30-69% English Exposure						
PGU	.52 [.38, .69]	63%	**88%**	28%	1.22	0.43
EPC	.55 [.38, .71]	0.82	76%	32%	1.12	0.75
MLUw	.66 [.50, .80]	4.94	40%	**96%**	10.00	0.63
SI	.64 [.48, .78]	0.99	56%	64%	1.56	0.69

Note. AUC = area under the curve. LR+ = positive likelihood ratio. LR- = negative likelihood ratio. PGU = Percent Grammatical Utterances. EPC = errors per C-unit. MLUw = Mean length of utterance in words. SI = Subordination Index. Metrics that reached the 80% threshold for sensitivity and/or specificity are in bold text.

Figure 2.10

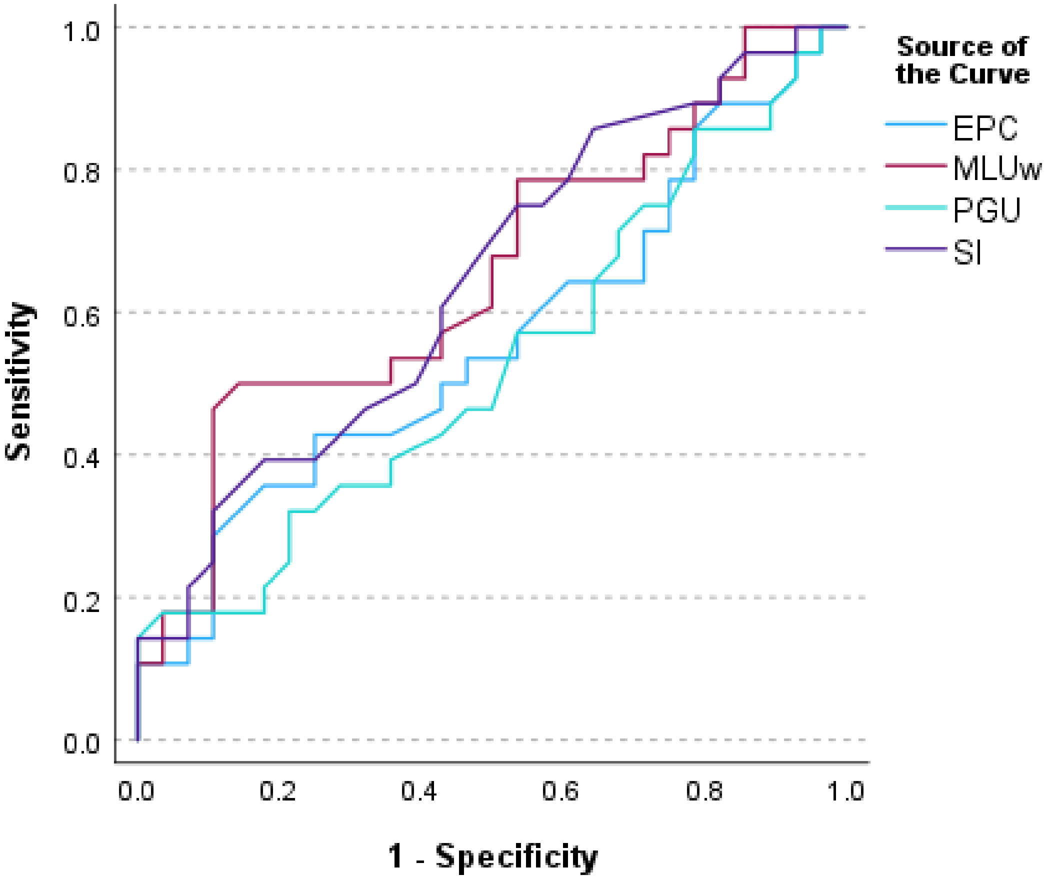
Pooled ROC Curves for Individual English LSA Measures for High English Exposure Group



Note. EPC = errors per C-unit. MLUw= mean length of utterance in words. PGU = percent grammatical utterances. SI = subordination index.

Figure 2.11

Pooled ROC Curves for Individual English LSA Measures for Balanced Exposure Group



Note. EPC = errors per C-unit. MLUw= mean length of utterance in words. PGU = percent grammatical utterances. SI = subordination index.

Percent Grammatical Utterances (PGU)

Based on pooled ROC analysis, PGU yielded poor classification accuracy (AUC=.69, 95% CI [.58, .79]), with 91% sensitivity and 43% specificity at the optimal threshold of 75% grammatical utterances. AROC analysis with language exposure as a covariate yielded a slightly larger AUC (.70, 95% CI [.57, .80]). The difference between the pooled and covariate-adjusted curves was not statistically significant (*p* = .31).

Errors per C-unit

Pooled ROC analysis of errors per C-unit yielded poor classification accuracy (AUC=.69, 95% CI [.58, .80]), with 91% sensitivity and 46% specificity at the optimal threshold of 0.32 errors per C-unit. AROC analysis with language exposure as a covariate yielded a larger AUC (.73, 95% CI [.61, .82]). The difference between the pooled and covariate-adjusted curves was not statistically significant (*p* = .26).

Mean Length of Utterance

Pooled ROC analysis of MLUw yielded poor classification accuracy (AUC=.71, 95% CI [.60, .81]), with 48% sensitivity and 89% specificity at the optimal threshold of 4.95 words per utterance. AROC analysis with language exposure as a covariate yielded a comparable AUC (.71, 95% CI [.61, .81], area difference *p* = .42).

Subordination Index

Pooled ROC analysis of subordination index yielded poor classification accuracy (AUC=.67, 95% CI [.58, .79]), with 72% sensitivity and 63% specificity at the optimal threshold of 1.00 clause per utterance. AROC analysis with language exposure as a covariate yielded an AUC that was slightly smaller (.68, 95% CI [.57, .78]), but the difference was not statistically significant (*p* = .88).

Diagnostic Accuracy of Combined LSA Measures

Since covariate-adjusted ROCs varied in whether they yielded better classification than pooled ROCs, multiROC analyses were run for the whole participant sample and separately for the high English and balanced exposure groups. Table 2.9 summarizes the results, and the multiROC curves are presented in Figure 2.12.

Table 2.9

Multivariate ROC Results in English

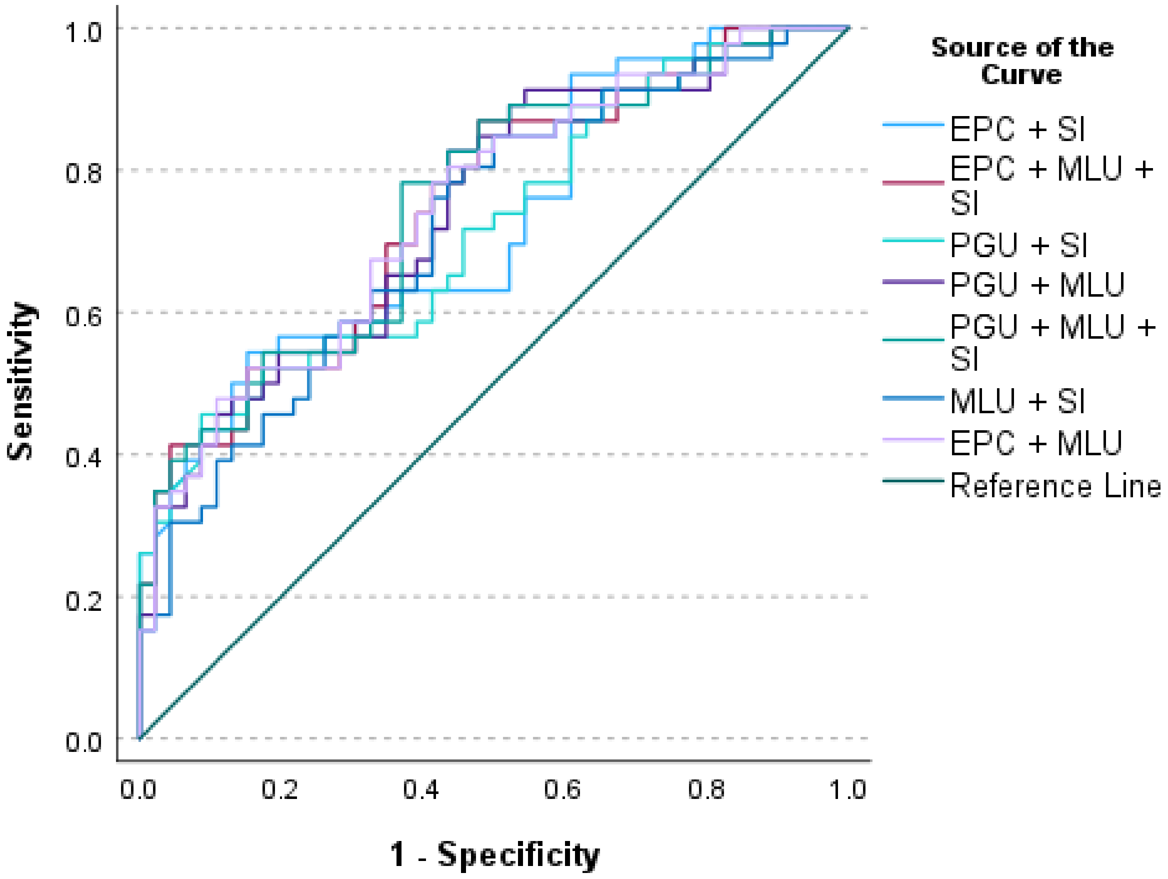
	AUC	Cutoffs	Sensitivity	Specificity	LR+	LR-
			All Exposure Levels			
PGU + MLUw	.74 [.64, .84]	42%, 5.46	**83%**	57%	1.93	0.30
PGU + SI	.72 [.61, .82]	76%, 1.05	46%	**91%**	5.11	0.59
PGU + MLUw + SI	.75 [.65, .85]	46%, 5.25, 1.04	78%	63%	2.11	0.35
MLUw + SI	.72 [.62, .82]	5.25, 1.04	76%	59%	1.85	0.41
EPC + MLUw	.75 [.65, .85]	0.69, 5.46	78%	59%	1.90	0.37
EPC + SI	.72 [.62, .83]	0.62, 1.13	54%	**85%**	3.60	0.54
EPC + MLUw + SI	.75 [.65, .85]	0.45, 5.09, 0.75	**87%**	52%	1.81	0.25
			≥70% English Exposure			
PGU + MLUw	.98 [.94, 1.0]	76%, 5.29	**94%**	**94%**	15.67	0.06
PGU + SI	.98 [.94, 1.0]	76%, 1.05	**94%**	**94%**	15.67	0.06
PGU + MLUw + SI	.98 [.94, 1.0]	.76%, 5.29, 1.05	**94%**	**94%**	15.67	0.06
MLUw + SI	.82 [.68, .96]	5.21, 1.0	**94%**	59%	2.29	0.10
EPC + MLUw	.98 [.95, 1.0]	0.29, 5.29	**94%**	**94%**	15.67	0.06
EPC + SI	.98 [.94, 1.0]	0.29, 1.05	**94%**	**94%**	15.67	0.06
EPC + MLUw + SI	.98 [.95, 1.0]	0.29, 5.29, 1.05	**94%**	**94%**	15.67	0.06

	AUC	Cutoffs	Sensitivity	Specificity	LR+	LR-
			30-69% English Exposure			
PGU + MLUw	.65 [.49, .80]	26%, 5.16	**92%**	39%	1.51	0.21
PGU + SI	.63 [.48, .79]	67%, 0.89	**92%**	35%	1.42	0.23
PGU + MLUw + SI	.66 [.51, .81]	24%, 5.0, 1.0	**84%**	46%	1.56	0.35
MLUw + SI	.67 [.52, .82]	5.0, 1.0	**84%**	46%	1.56	0.35
EPC + MLUw	.65 [.50, .80]	0.75, 5.25	**84%**	46%	1.56	0.35
EPC + SI	.62 [.47, .78]	0.65, 0.92	**92%**	39%	1.51	0.21
EPC + MLUw + SI	.65 [.50, .81]	0.82, 5.0, 1.0	**84%**	42%	1.45	0.38

Note. AUC = area under the curve. LR+ = positive likelihood ratio. LR- = negative likelihood ratio. PGU = Percent Grammatical Utterances. EPC = errors per C-unit. MLUw = Mean length of utterance in words. SI = Subordination Index. Metrics that reached the 80% threshold for sensitivity and/or specificity are in bold text.

Figure 2.12

English Multivariate ROC Curves



Note. EPC = errors per C-unit. MLUw= mean length of utterance in words. PGU = percent grammatical utterances. SI = subordination index.

High English Exposure

For bilinguals with at least 70% exposure to English, combining either PGU or errors per C-unit with MLUw, subordination index, or both yielded good diagnostic accuracy with 94% sensitivity and 94% specificity (AUC=.98). MLUw and subordination index yielded 94% sensitivity but only 59% specificity (AUC=.82).

Balanced Exposure

Models for the balanced exposure group, whose English exposure was between 30 and 70%, yielded acceptable to good sensitivity but poor specificity. PGU yielded 92% sensitivity when combined with either subordination index (35% specificity, AUC=.63) or MLUw (39% specificity; AUC=.65). Combining all three of these measures yielded 84% sensitivity and 46% specificity (AUC=.66), as did combining MLUw with subordination index (AUC=.67). Errors per C-unit yielded better specificity than PGU when combined with subordination index (92% sensitivity, 39% specificity, AUC=.62) or MLUw (46% specificity, AUC=.65), though sensitivity was not as high for MLUw model at 84%.

Discussion

After analyzing LSA measures that had previously been shown to be clinically useful for identifying DLD in monolingual English speakers, we were able to demonstrate that these same measures are also informative for bilingual speakers with a high level of English exposure but not for those with less than 70% exposure to English even when LSA measures are combined.

When applying a single cutoff value for the full continuum of language exposure, the optimal diagnostic accuracy of individual LSA measures was inadequate overall. PGU and errors per C-unit had good sensitivity but very low specificity, meaning a significant number of typically developing bilingual children would be classified as having DLD based on their performance relative to the cutoff. MLUw had acceptable (almost good) specificity but low sensitivity, meaning a significant number of children with DLD scored above the cutoff and would be classified as typical. Subordination index was inadequate in both.

Given the wide variation in participants' English exposure, their performance on these LSA measures was expected to vary substantially, likely necessitating more than one cutoff score for accurate classification. For this reason, we conducted covariate-adjusted ROC analyses, which allowed us to model each LSA measure's diagnostic accuracy using language exposure-specific cutoffs. The resulting AUCs, which summarize the overall adequacy of the measure by averaging the exposure-specific diagnostic accuracy levels, showed slight if any improvement over pooled ROC analyses. This suggested that using multiple cutoffs might yield better diagnostic accuracy for some participants, but classification would still be generally inadequate. This was confirmed through visual inspection of the distributions of participant scores on LSA measures and follow-up pooled ROCs for each exposure group. These revealed that, while diagnostic accuracy was still poor for the balanced exposure group, PGU and errors per C-unit both had excellent diagnostic accuracy for children with 70% or more English exposure.

The possibility that using a combination of LSA measures might yield better diagnostic accuracy than individual measures was addressed using multivariate ROC analyses. For the full continuum of English exposure, results indicated slight improvement of combined over individual LSA measures based on higher AUCs, but diagnostic accuracy for the whole participant sample still did not reach the desired 80% threshold. The model that maximized sensitivity and specificity at 78% and 63%, respectively, combined PGU, MLUw, and subordination index, though its AUC of .75 was not statistically different from using only PGU and MLUw (*p* = .49; 83% sensitivity, 57% specificity) or from using errors per C-unit instead with MLUw and subordination index (*p* = .83; 87% sensitivity, 52% specificity). When the participant sample was disaggregated, various combinations of LSA measures yielded good diagnostic accuracy for those with at least 70% English exposure, but all were comparable to using PGU or errors per C-unit on its own (*p* = .39-.65; 94% sensitivity and specificity).

Our findings regarding which measures are most informative are consistent with previous studies on monolingual English LSA, specifically that grammaticality measures are best. Monolingual 5- and 6-year-olds could be classified with acceptable to good diagnostic accuracy using errors per C-unit (Guo & Schneider, 2016), PGU (Guo et al., 2019), or other measures quantifying grammatical utterances (e.g., Sentence Point score; Souto et al., 2014). These same measures yielded excellent diagnostic accuracy for our participants with high English exposure using an empirically-derived cutoff. For bilinguals with less exposure to English, however, measures that reflect specific error types may be more informative than the grammaticality measures examined in the current study, which were based on error rate. For example, the finite verb morphology composite had excellent diagnostic accuracy for monolingual speakers ages 4 to 5 (Gladfelter & Leonard, 2013; Guo et al., 2020; Souto et al., 2014), and pairing it with MLU maintained excellent accuracy for children as young as 3 (Bedore & Leonard, 1998). Such a model might also be effective for bilinguals who are in early stages of English language acquisition.

Previous studies examining LSA diagnostic accuracy for bilingual speakers of English reported inadequate to possibly acceptable sensitivity and specificity, while the current study found good diagnostic accuracy for grammaticality measures. The measures examined likely contributed to differences in findings. For example, studies based on monolingual speakers have consistently demonstrated the diagnostic value of measures based on grammatical accuracy, while measures of length, morphosyntactic proficiency, and semantics have been less useful for classification. The LSA measures Ooi and Wong (2012) tested with Malaysian English bilinguals - IPSyn Total Score, MLU, and lexical diversity *D* - were very similar to those that Hewitt et al. (2005) found had only 74% overall diagnostic accuracy for monolinguals. Smyk (2012), in contrast, included errors per T-unit along with MLUw, number of different words, and mazes and obtained more promising results.

Other notable differences between studies include the age and language background of participants. The bilinguals in Ooi and Wong (2012) spoke Cantonese as their home language and may have produced different morphosyntactic patterns as a result of language transfer than our Spanish-speaking bilinguals. In much the same way that LSA measures have differential diagnostic accuracy across dialects of English (e.g., Oetting et al., 2021), such L1-based variation could affect the diagnostic performance of a measure when applied with speakers of different home languages. It is also difficult to compare the language exposure levels of our participants to those of previous studies without a common metric, so if there were significant differences, they likely contributed to the inconsistent findings. One implication of our results is that evaluating the diagnostic accuracy of measures that are influenced by language exposure without accounting for exposure in that analysis may produce results that underestimate or mask diagnostic accuracy.

Limitations & Future Directions

The current study was exploratory in nature, so validation of findings with another sample of bilingual children in future research is critical. A matched pairs design may not have captured enough of the variability in performance among our sample, especially at more balanced or high Spanish exposure levels, so a larger and/or community-based sample may be more informative. Attempts should also be made to ensure that an adequate number of participants are included at each level of exposure, as the skew toward high English exposure in our sample prevented deeper exploration of diagnosis along the continuum.

Specific coding decisions made in this study, though aligned with prior studies and SALT conventions, may have limited the capacity for measures to capture distinguishing characteristics of DLD in bilinguals' spontaneous language. For example, lexical and semantic errors were not included as grammatical errors, though they have been in some studies of PGU with monolinguals (e.g., Guo et al., 2019). Future studies could explore the impact of these decisions on diagnostic accuracy empirically.

Clinical Implications

Our results demonstrate that English LSA - specifically, using one of the measures of grammaticality - can be used to identify DLD in 5- to 6-year-old Spanish-English bilingual children who have at least 70% English exposure. It has been unclear at what level of exposure English assessments can be used with English-dominant bilinguals, and so the most defensible choice has been to test all bilinguals in both languages and reserve English-only testing for monolingual speakers. However, in settings where bilingual support is limited, such a guideline is particularly restrictive and, in practice, frequently not followed

(Arias & Friberg, 2017). Though the current findings do not apply to a substantial portion of bilingual students, the ability to use validated English LSA with at least some of them can alleviate the strain on bilingual resources just a bit.

Good diagnostic accuracy was achieved using PGU or errors per C-unit, as well as any combination of LSA measures that include one of them. The simplest to implement would be PGU since it requires less coding (a single judgment for each utterance rather than identification of all errors) and is automatically calculated by LSA software. PGU could also be easily hand-calculated from a transcript without the use of specialized software. Furthermore, PGU also has acceptable diagnostic accuracy with monolingual speakers across a broad age range, providing clinicians with a continuity in LSA procedures that can greatly streamline its implementation. Note that the recommended cutoffs differ for the two groups - our findings indicate an optimal cutoff of 74% compared to 80% and 83% for monolinguals aged 5 and 6, respectively (Guo et al., 2019).

Assessing bilingual children in both of their languages continues to be best practice, not only from the standpoint of identification of DLD (which our results demonstrate) but also for other equally important goals of an assessment, such as understanding a child's strengths, needs, and the linguistic resources at their disposal for functioning in their various contexts. The search for a valid English-only approach serves a practical purpose given the limitations of our field, but it is imperative to remember that results of such an assessment are ultimately an incomplete representation of a bilingual child's language skills. That is the case even for children with greater than 70% exposure to English, who can be classified accurately using an English language sample but who do not cease to be bilingual speakers.

It also bears emphasizing that poor diagnostic accuracy for bilinguals with more limited English exposure is a reflection of the *measure's* inadequacy to identify children's language ability and should in no way be interpreted as a problem related to the children's English skills. Framing the challenges of assessing bilingual children as them "not knowing enough English to be tested" reinforces deficit narratives and also misses the root cause that our assessment tools are not up to the task of capturing what children are doing with their language, given their experience and ability. Our focus should be on identifying the right tool for the job using the evidence available.

Conclusion

While DLD is often misidentified in bilingual children due to practices such as using English tests, language sample analysis is a naturalistic and culturally sensitive tool that may be more effective for identifying DLD in this population. We found that percent grammatical utterances and errors per C-unit derived from English narrative samples each had excellent diagnostic accuracy of 94% for identifying Spanish-English bilingual 5- and 6-year-olds whose current exposure to English was at least 70%. None of the measures or combinations of measures we examined reached acceptable levels of diagnostic accuracy for participants with less exposure to English. As our study was exploratory, future research should confirm findings with a separate sample of bilingual children that is adequately large at all levels of English exposure for detailed analyses. Further exploration of the diagnostic accuracy of other LSA measures and cross-linguistic analysis with Spanish measures would also be informative for clinical practice.

Chapter 3: Spanish LSA Measures (Study 3)

Background

Bilingual children are simultaneously at risk of underidentification and overidentification of developmental language disorder (DLD; Collins et al., 2014; Morgan et al., 2017; Sullivan, 2011), which results in inappropriate educational placement and disparities in academic and health outcomes. Assessment standards set by the Individuals with Disabilities Education Act and the American Speech-Language and Hearing Association include multiple sources of evidence and assessment in the child's native language, but typical practice continues to fall short of these (Arias & Friberg, 2017). One limitation is the scarcity of home language assessment tools, even in the most commonly spoken languages (i.e., Spanish in the United States). Language sample analysis (LSA) is frequently recommended for this population and is preferable to adaptation of English instruments, such as translating standardized test protocols. However, clinicians feel they lack the knowledge to analyze and interpret LSA across languages (Guiberson & Atkins, 2012), and extant research offers limited empirical evidence to guide those decisions. This study seeks to inform the clinical use of LSA in Spanish by validating four language sample-derived measures for the identification of DLD in bilingual children representing a range of language experience.

Availability of Bilingual Assessment Tools

Though routine bilingual assessment practices have been gradually moving toward recommended practices, conducting testing in the home language continues to present formidable challenges (Arias & Friberg, 2017). Even when a qualified practitioner or interpreter is available, there is a persistent shortage of standardized tests in languages other than English, especially tests that have been validated for the identification of DLD (Arias & Friberg, 2017; Castilla-Earls et al., 2020; Guiberson & Atkins, 2012; Huang et al., 1997). More routine implementation of language sample analysis (LSA), which is frequently recommended as a gold standard for bilingual children, would enable clinicians to improve identification of DLD in a way that bypasses this shortage. The ability to conduct LSA is not constrained by the availability of a published instrument in the same way that standardized testing is. The naturalistic interactions commonly used to elicit language samples are familiar across cultures (Heilmann & Westerveld, 2013) and easily adapted to any language, and evidence-based procedures can be implemented as soon as recommendations are disseminated. As a culturally responsive assessment, LSA can also minimize bias related to cultural and linguistic variation found in tests (Kraemer & Fabiano-Smith, 2017; Rojas & Iglesias, 2009; Stockman, 1996).

Unlike standardized test scores, however, interpretation of the data derived from a language sample is largely at the discretion of the clinician and thus requires adequate knowledge. One obstacle to greater uptake of LSA in the home language is a lack of familiarity with the language-specific characteristics of DLD needed for clinical interpretation of LSA results (Guiberson & Atkins, 2012). A potential solution would be to develop guidelines for selecting and interpreting LSA measures in various languages based on their ability to accurately differentiate DLD from typical developmental patterns. While the literature offers comparative data on performance related to ability (Castilla-Earls et al. 2021), very few studies provide evidence of the classification accuracy of specific LSA measures to validate their use for diagnostic purposes. Further investigation would not only give clinicians the knowledge to recognize features of DLD in spontaneous language across different languages, it would also perform the necessary validation of LSA measures in the intended language of administration.

While considerably more evidence is available regarding the diagnostic accuracy of LSA in English, like other assessment instruments, LSA must be validated for diagnostic purposes in the language in which it will be administered. This is because the characteristics of DLD are language-specific. Clinical markers vary according to the unique structural properties of the language being acquired (Leonard, 2014b). Simply translating the tests and procedures that are available in English may no longer capture clinically relevant features of DLD post-translation (Arnold & Matus, 2000; Bedore & Peña, 2008; Bracken & Barona, 1991; Peña, 2007). Similarly, LSA measures and cutoffs that accurately differentiate DLD in English may not be as effective when applied to another language if they don't reflect the unique clinical markers and distribution of performance within that language.

Prior Research on Spanish LSA Measures

While Spanish language sample data has been used frequently to compare the performance of bilingual speakers with DLD to those with typically developing language, few studies have examined its classification accuracy. Of these, only one study reported adequate diagnostic accuracy of at least 80% sensitivity and specificity. Kapantzoglou et al. (2017) found that 4- to 5-year-old bilingual children were accurately classified using grammaticality and lexical diversity *D* calculated from story retell data (90% sensitivity, 85% specificity, overall 87.5%) and using grammaticality and subordination index calculated from story generation data (80% sensitivity, 85% specificity;).

When examined independently, these and other measures were found to be inadequate for identifying DLD on their own. Percent ungrammatical utterances nearly reached acceptable accuracy with 79% sensitivity and 100% specificity for Spanish-dominant (i.e., >60% input at home) 4- and 5-year-olds (Simón-Cereijido & Gutiérrez-Clellen, 2007) but fell well below the 80% threshold for near-monolingual (90% or greater exposure to Spanish) 3- to 5-year-olds (59% sensitivity, 67% specificity; Guiberson et al., 2015). Errors per T-unit yielded 70% sensitivity and 100% specificity for Spanish-dominant 5- to 7-year-olds (Restrepo, 1998). Errors per T-unit was also examined with monolingual 3- to 6-year-olds' conversational samples, as was subordination index; presumably, both were found to be inadequate as results were only reported for measures that met the 80% criterion were reported (Grinstead et al., 2012). Number of different word roots (NDW) yielded acceptable sensitivity for this participant group (85%) but inadequate specificity (72%; Grinstead et al., 2012). MLU alone yielded only 58% sensitivity and 74% specificity for Spanish-dominant 4- and 5-year-olds (Simón-Cereijido & Gutiérrez-Clellen, 2007). Better sensitivity was achieved for other participant samples also consisting of 4- to 5-year-old bilingual children (85%; Lazewnik et al., 2019) and of 3- to 6-year-old monolingual Spanish-speaking children (81%; Grinstead et al., 2012), but specificity remained inadequate at 57.1% and 76%, respectively.

Previous studies have relied on discriminant function analysis and linear regression, which examine diagnostic accuracy at a single cutoff score. Particularly when the lower metric is nearly acceptable, such as for percent ungrammatical utterances found by Simón-Cereijido & Gutiérrez-Clellen (2007) or perhaps even Restrepo's (1998) findings for errors per T-unit, an alternative cutoff may reveal more optimal diagnostic accuracy for clinical purposes. Alternatively, receiver operator characteristic (ROC) analysis and its multivariable variant (multiROC) evaluate diagnostic accuracy at all possible cutoffs, making it possible to identify the one that optimally balances sensitivity and specificity and, thus, the maximal potential of the measure.

For ecological validity and generalization of research findings to clinical practice, LSA measures must be validated across a range of bilingual experiences, and the influence of experience on the diagnostic performance of the measures must be understood. Studies to date have focused on monolingual (Grinstead et al., 2012) or Spanish-dominant speakers, though descriptors and criteria varied, making direct comparisons difficult. For example, the majority of participants in Guiberson et al. (2015) - described as "all to mostly Spanish-speaking" - had 90-100% exposure to Spanish and none less than 80%, while the participants in Simón-Cereijido & Gutiérrez-Clellen (2007) - described as "Spanish-speaking with limited English proficiency" - had a mean of 83% Spanish input (output was not reported). Others have used proficiency measures to determine Spanish dominance (e.g., Restrepo, 1998) or language history (i.e., first language was Spanish, years of exposure to English; Lazewnik et al., 2019; Kapantzoglou et al., 2017).

The challenge in including a heterogeneous sample of bilingual speakers lies in the influence of language experience on language production, which could affect the diagnostic performance of measures that use production to predict ability. Amount of exposure to a language strongly predicts acquisition and accuracy in producing different structures in that language (Bedore et al., 2012; Paradis, 2010). Examination of potential diagnostic indicators must, therefore, account for a child's experience with the language when identifying cutoff scores (Castilla-Earls et al., 2016; Coloma et al., 2016). If we estimate diagnostic accuracy of LSA measures for DLD without accounting for the variation that language exposure contributes to observations on those measures, our results will potentially underestimate their true diagnostic accuracy (Janes & Pepe, 2008). Extending previous findings on Spanish LSA to the continuum of bilingualism requires both a clear measure of participants' language exposure and its inclusion in the analyses, such as with covariate-adjusted ROC (Janes et al., 2009).

Purpose

LSA offers some unique advantages that could be leveraged for rapid change in clinical practice in the midst of the scarcity of home language assessments. Clinicians have expressed a need for better understanding of clinical indicators of DLD across different languages in order to conduct more valid and accurate assessments in the child's home language, whether administering them themselves or in collaboration with an interpreter. Research to validate LSA methods for diagnosis of DLD across different languages would provide the desired expertise through evidence-based guidelines for selecting and interpreting measures. To that end, the current study extends prior research by exploring the clinical usefulness of Spanish LSA measures across a broader range of bilingual experience using statistical methods that account for language exposure and that can identify the measure's optimal diagnostic performance. Focusing on LSA measures with the strongest evidence to date, the following research questions were addressed:

1) What is the optimal classification accuracy for identifying DLD in Spanish-English bilingual 5- and 6-year-olds using percent grammatical utterances, errors per C-unit, MLU in words, and subordination index calculated from Spanish narratives?

2) Is classification accuracy improved by adjusting for language exposure?

3) Is classification accuracy improved by using a combination of these measures?

Methods

Data Source

This study involved secondary analysis of existing language sample data drawn from two projects: Development of a Test for Hispanic Children in the U.S. (DTHC) and Diagnostic Markers of Language Impairment (DM). For a detailed description of the participants and data collection, see Gutiérrez-Clellen et al. (2006), Gutiérrez-Clellen & Simón-Cereijido (2007), and Peña et al., 2018 for DTHC and Gillam et al., (2013) and Peña et al. (2011) for DM.

The goal of the DTHC project was to develop and validate a new diagnostic language test for identifying DLD in Spanish-English speakers. 756 participants aged 4;0-6;11 were recruited via a one-gate design from school districts in California, Texas, and Pennsylvania serving primarily low-income students. The goal of the DM project was to examine clinical markers of DLD in bilingual children. Participants were recruited via a one-gate design across 12 elementary schools from three districts in Texas and Utah serving a high number of bilingual Latinx students, and 168 participants aged 5;0-6;5 completed follow-up testing as part of the longitudinal component of the project.

Current Study Participants

Participants were selected for the current study from the larger samples who were between age 5;0 and 6;11, were not missing data for determining language exposure or ability status, and produced a language sample in Spanish even if they did not in English.

Table 3.1 summarizes participant characteristics. Language ability was classified as TD or DLD using standard scores on the Bilingual English-Spanish Assessment (BESA; Peña et al., 2018) Language Index Composite score based on a cutoff of 85 (-1 *SD*), which has acceptable diagnostic accuracy for 5-year-olds (88% sensitivity/85% specificity) and good accuracy for 6-year-olds (96% sensitivity/92% specificity). Fifty-seven participants were classified as having DLD, of which six were excluded due to inadequate samples (i.e., no analyzable utterances of more than one word, no Spanish used). The remaining 51 were matched with TD participants from the same source project based on language exposure within 2% and age within 5 months. A TD match could not be identified for nine DLD participants. The final participant sample included 84 participants, with 28 participants from the DTHC study and 56 from the DM study. Spanish exposure ranged from 28 to 100% (*M*=60.1, *SD*=24.3).

Table 3.1

Study 3 Participant Characteristics

	Combined			DTHC			DM		
	Total	TD	DLD	Total	TD	DLD	Total	TD	DLD
N	84	42	42	28	14	14	56	28	28
Age in months (*SD*)	68.2 (5.2)	68.15 (5.22)	68.24 (5.30)	70.2 (6.6)	70.4 (7.0)	70.0 (6.4)	67.3 (4.2)	67.1 (3.9)	67.4 (4.6)
Spanish exposure									
Mean (%)	60.1	60.05	60.06	84.4	84.5	84.3	48.8	48.7	48.8
Range (%)	28-100	28-100	29-100	44-100	45-100	44-100	28-85	28-85	29-83

Note. DTHC = Development of a Test for Hispanic Children in the US participants. DM = Diagnostic Markers of Language Impairment participants. TD = typically developing. DLD = developmental language disorder.

Materials

BIOS. Relative exposure to and use of Spanish and English was measured using the Bilingual Input-Output Survey (BIOS; Peña et al., 2018). The BIOS is a structured interview completed with a parent and teacher. Interviewees provide an hour-by-hour report of which language(s) a child currently hears and speaks, which is calculated into an overall percentage of input-output in each language, as well a year-by-year history of exposure to the two languages at home and school.

Language Samples. Participants across both projects completed a story retell task in Spanish using wordless picture books. They listened to a story told by the examiner while looking at the book, then retold the same story while looking at the book. DTHC samples were elicited using *Frog on His Own* (Mayer, 1973). DM samples were elicited using either *Frog on His Own* (Mayer, 1973) or *A Boy, a Dog, and a Frog* (Mayer, 1967), which were counterbalanced by target language. All samples were audio-recorded and transcribed using Systematic Analysis of Language Transcripts (SALT; Miller & Iglesias, 2008) software by a trained research assistant. Utterances were segmented into C-units, and mazes (i.e., word repetitions, filled pauses, etc.) were marked. Verbs were linked to their word roots. Morphemes were marked, which in Spanish only includes plural -*s*. Transcripts were checked by a second research assistant, and any discrepancies were resolved by an independent transcriber (see Gutiérrez-Clellen et al., 2006 and Bedore et al., 2010 for detailed procedures).

Procedures

Language Sample Coding. For the current study, transcripts were additionally coded for errors and subordination. The grammaticality of each utterance was judged in isolation, without reference to the story details or the child's previous utterances. Utterances that were judged to be ungrammatical were coded with the least number of changes needed to make a grammatical utterance. Errors were coded according to SALT conventions, which include morpheme omissions (in Spanish, this only includes plural -s), word omissions including but not limited to prepositions and obligatory clitic pronouns, and word-level substitution errors including subject-verb agreement, tense, number, and gender substitutions (see Table 3.2). SALT conventions do not include coding of semantic errors. Each utterance was coded for the number of clauses it contains according to SALT conventions for subordination index (i.e., clauses per utterance). These coding procedures closely align with previous studies (e.g., Simón-Cereijido & Gutiérrez-Clellen, 2007).

Table 3.2

Coding Scheme for Spanish Grammatical Errors

Error Type	Code(s)	Example
Morpheme omission		
Plural –s	WORD/*s	“mucha ranas” = mucha/*s ranas
Word omission		
Prepositions	*WORD	“Fueron la casa” = fueron *a la casa
Obligatory clitic pronouns		“Dio la leche” = *Le dio la leche
Word-level substitution		
Subject-verb agreement	[EW:WORD]	“Las ranas estaba allí.” = Las ranas estaba[EW:estaban] allí.
Tense		“Vio que la rana no está allí” = Vio que la rana no está[EW:estaba] allí
Number		“Los perro se cayó” = Los[EW:el] perro se cayó
Gender		“la bosque” = la[EW:el] bosque
Utterance-level Errors		
Word Order	[WO]	“Había una grande rana” = Había una rana grande [WO]
Extraneous Words	[EW]	“El regalo era para del niño” - El regalo era para[EW] del niño

Training & Reliability. Coding for all transcripts was completed by the first author and trained research assistants (RA). RAs included an undergraduate student, two master's level graduate students in speech-language pathology, and one practicing SLP. All were bilingual Spanish-English speakers. RAs completed relevant self-paced online courses provided on the SALT website, which included coding of 3 practice transcripts. They met with the first author to review their practice transcripts and the coding guidelines for the current study. To establish initial inter-rater reliability, RAs coded at least 10 transcripts from a separate dataset and continued coding additional transcripts until reaching at least 85% agreement on 2 transcripts. To determine inter-rater reliability for coding of the current dataset, 20% of total samples were reviewed by a second coder using a consensus procedure modeled on Guo et al. (2019). Discrepancies were discussed to reach agreement. Interrater reliability was 82%.

Measures

The SALT software was used to generate the following measures: mean length of utterance in words (MLU-w), subordination index, total utterances containing error codes, total utterances, total morpheme omissions, total word omissions, total word codes, and total utterance codes. Percent grammatical utterances (PGU) was calculated as the inverse of the total utterances containing error codes. Errors per C-unit were calculated by dividing the sum of all error types (total morpheme omissions, total word omissions, total word codes, total utterance codes) by the total number of utterances. See Table 3.3 for a summary of these procedures. The following utterances were excluded from analyses: abandoned, unintelligible, single word, and utterances containing code-switching.

Table 3.3

Procedures for Calculating LSA Measures in Spanish

Measure	Calculation	Method/Source
Mean length of utterance in words	Total number of words in the sample divided by total number of utterances/C-units in the sample	SALT SMR: MLU-w
Errors per C-unit	Total number of omissions, substitutions, and word order errors coded in the sample divided by total number of utterances/C-units in the sample	Hand-calculated using • SALT SMR: Total Omitted Morphemes • SALT SMR: Total Omitted Words • SALT SMR: Total Word Codes • SALT SMR: Total Utterance Codes • SALT SMR: Total Utterances
Percent grammatical utterances	Total utterances with no coded errors divided by total utterances in the sample, converted to a percentage	Calculated as the inverse of: • SALT SMR: % Utterances Containing Error Codes
Subordination Index	Total clauses (independent and subordinate counted separately) in the sample divided by total utterances in the sample	SALT SMR: Subordination Index

Note. SALT SMR = Systematic Analysis of Language Transcripts Standard Measures Report.

Analytic Strategy

The first research question explored the optimal classification accuracy for identifying DLD in Spanish-English bilingual 5- and 6-year olds using four LSA measures (PGU, errors per C-unit, MLU in words, and subordination index) calculated from Spanish narratives when adjusting for language exposure. To address this question, descriptive statistics including group means and standard deviations were calculated to examine the distribution of each LSA measure. Correlations between Spanish exposure and LSA measures were calculated to verify appropriateness for inclusion as a covariate in subsequent analyses (Janes et al., 2009). Independent samples *t*-tests were run to verify DLD versus TD group differences on each LSA measure. Analyses were performed using jamovi (The jamovi project, 2024).

To examine classification accuracy, each LSA measure showing a statistically significant difference between TD and DLD participants was entered into a receiver operator characteristic (ROC) analysis, with the optimal cut point determined using the Youden index (Youden, 1950), which has been used in previous studies of diagnostic accuracy for identification of DLD (e.g., Oetting et al., 2021; Redmond et al. 2019). The following indicators of classification accuracy were generated and evaluated against established criteria for clinical usefulness: area under the curve (AUC; Youngstrom, 2014), sensitivity and specificity (Plante & Vance, 1994), positive and negative likelihood ratios (Dollaghan, 2004), and 95% confidence intervals. Analyses were performed using the ROCnReg package for R Studio, and curves were plotted using SPSS version 29.0.0.0.

To examine whether accounting for language exposure improves classification accuracy, LSA measures were then entered into covariate-adjusted ROC with Spanish exposure as a continuous variable included as a covariate (AROC; Janes & Pepe, 2008). Indicators of classification accuracy were again generated (i.e., AUC, sensitivity and specificity, positive and negative likelihood ratios, 95% confidence intervals) and evaluated against the pooled ROC and established criteria for clinical usefulness. Analyses were performed using the ROCnReg package for R Studio, and curves were plotted using SPSS version 29.0.0.0.

The third research question explores whether classification accuracy is improved by using a combination of the measures of interest. To evaluate whether combining one or more LSA measures improves on their individual classification accuracy, LSA measures were entered into multivariate receiver operator characteristic curves (multiROCs; Schultz, 1995) and run separately for each language exposure group if the covariate-adjusted curves yielded higher classification accuracy. Indicators of classification accuracy were again generated (i.e., AUC, sensitivity and specificity, positive and negative likelihood ratios, 95% confidence intervals) and evaluated against the pooled ROC and established criteria for clinical usefulness. Analyses were performed using the multipleROC package for R Studio, and curves were plotted using SPSS version 29.0.0.0.

Results

Descriptive Analyses

Table 3.4 summarizes the descriptive statistics for the four LSA measures of interest - percent grammatical utterances (PGU), errors per C-unit, mean length of utterance in words (MLUw), and subordination index - as well as additional characteristics of participants' language samples. Bivariate correlations are presented in Table 3.5, and scatterplots of each LSA measure with English exposure are shown in Figures 3.1 through 3.4. Results of independent *t*-tests are presented in Table 3.6.

Table 3.4

Descriptive Statistics for Spanish Language Sample-Derived Measures

				95% CI				
	Ability	*N*	Mean	Lower	Upper	SD	Min	Max
Total Utterances	DLD	41	25.78	22.31	29.25	11.01	6	61
	TD	41	30.66	26.52	34.79	13.87	14	73
Total Words	DLD	41	116.93	97.62	136.23	61.17	14	275
	TD	41	161.59	136.78	186.39	78.60	50	388
# Different Words	DLD	41	49.10	43.85	54.34	16.62	10	85
	TD	41	61.85	56.32	67.38	17.52	30	94
Type-Token Ratio	DLD	41	0.47	0.43	0.50	0.11	0.24	0.71
	TD	41	0.42	0.39	0.45	0.10	0.24	0.66
PGU	DLD	41	0.52	0.46	0.58	0.18	0.17	0.83
	TD	41	0.65	0.59	0.70	0.18	0.29	1.0
Errors per C-unit	DLD	41	0.71	0.60	0.82	0.35	0.25	1.71
	TD	41	0.48	0.38	0.57	0.29	0.00	1.27
MLUw	DLD	41	4.39	4.06	4.72	1.04	2.33	6.30
	TD	41	5.17	4.86	5.48	0.98	3.13	7.67
Subordination Index	DLD	41	1.02	0.99	1.04	.08	0.82	1.19
	TD	41	1.07	1.04	1.12	.11	0.81	1.32

Note. PGU = Percent Grammatical Utterances. MLU= Mean Length of Utterance in words. DLD = Developmental Language Disorder. TD = Typically Developing.

Table 3.5

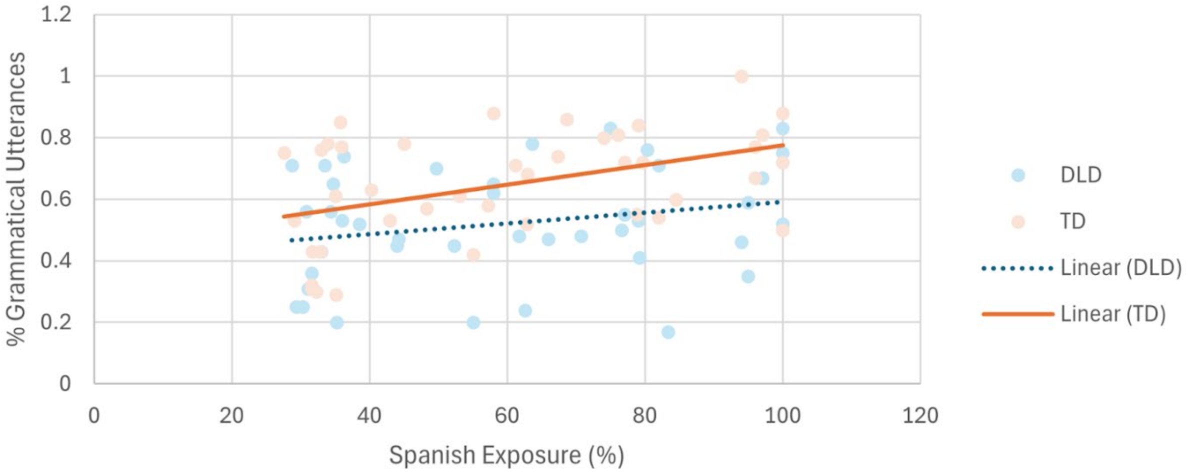
Bivariate Correlations between Study 3 Participant Characteristics and Spanish LSA Measures

		Ability	Exposure	Age	PGU	EPC	MLUw	SI
Ability	Pearson's *r*	—						
	p	—						
Exposure	Pearson's *r*	.00	—					
	p	1.00	—					
Age	Pearson's *r*	-.01	-.01	—				
	p	.93	.96	—				
PGU	Pearson's *r*	.33 **	.32 **	-.001	—			
	p	.002	.004	.99	—			
EPC	Pearson's *r*	-.35 **	-.23 *	-.03	-.94 ***	—		
	p	.001	.04	.81	< .001	—		
MLUw	Pearson's *r*	.36 ***	.31 **	-.08	.20	-.13	—	
	p	< .001	.004	.50	.07	.24	—	
SI	Pearson's *r*	.28 *	.23 *	-.08	.31 **	-.32 **	.57 ***	—
	p	.01	.04	.45	.004	.003	< .001	—

Note. PGU = Percent Grammatical Utterances. EPC = Errors per C-unit. MLUw = Mean Length of Utterance in words. SI = Subordination Index. * *p* < .05, ** *p* < .01, *** *p* < .001

Figure 3.1

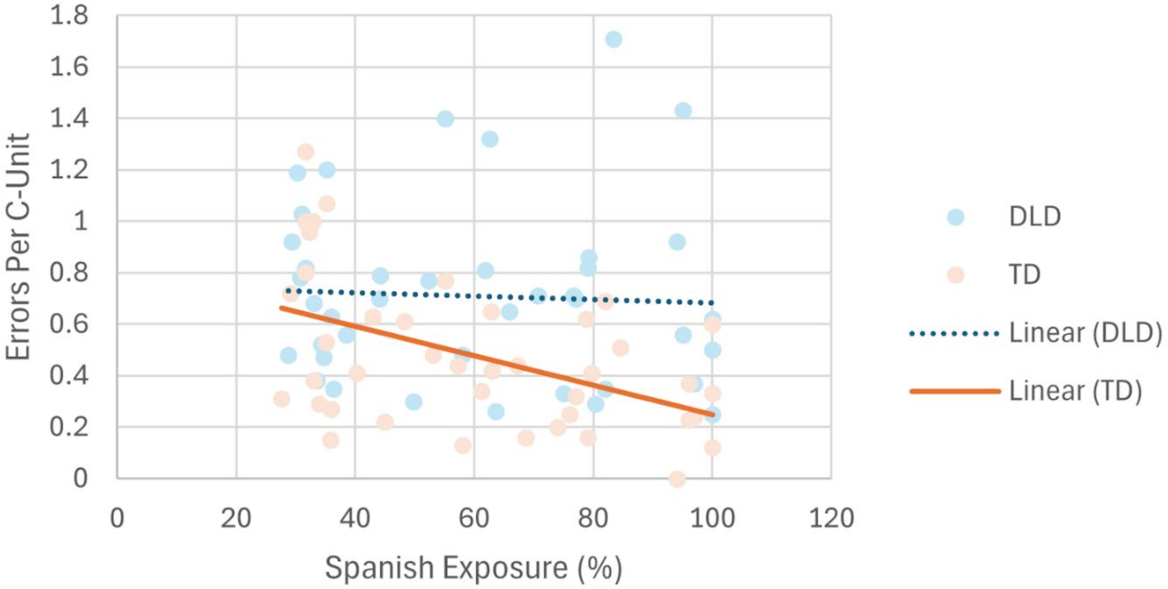
Percent Grammatical Utterances by Spanish Exposure



Note. DLD = Developmental Language Disorder. TD = Typically Developing

Figure 3.2

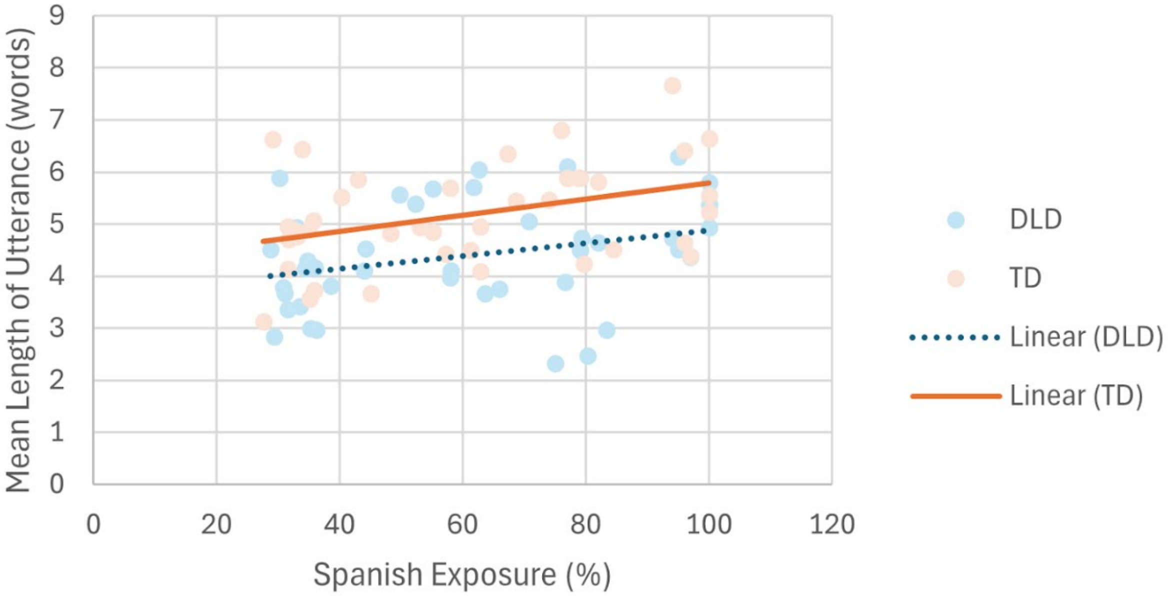
Errors Per C-unit by Spanish Exposure



Note. DLD = Developmental Language Disorder. TD = Typically Developing

Figure 3.3

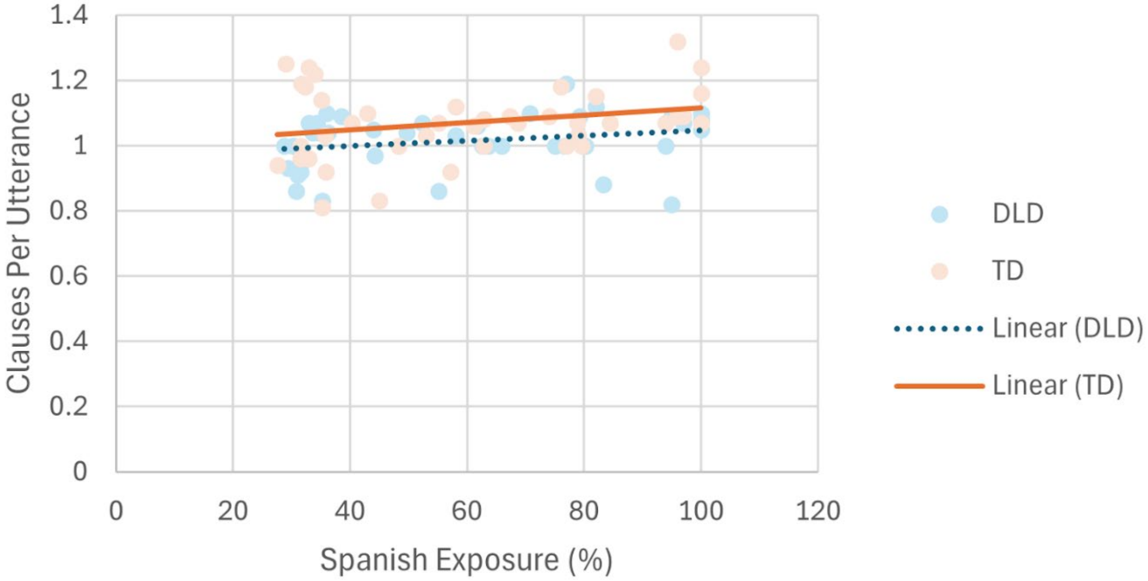
Mean Length of Utterance in Words by Spanish Exposure



Note. DLD = Developmental Language Disorder. TD = Typically Developing

Figure 3.4

Subordination Index by Spanish Exposure



Note. DLD = Developmental Language Disorder. TD = Typically Developing

Table 3.6

Independent Samples T-Tests for Age and Spanish LSA Measures

	t	**df**	***p***	***d***
Age (months)	0.84	80.0	.93	.02
Total Utterances	-1.18	77.7	.007	-.40
Total Words	-2.87	75.4	<.01	-.63
# Different Words	-3.38	79.8	<.01	-.75
Type-Token Ratio	1.93	78.1	.06	.43
PGU	-3.17	80.0	< .01	-.70
Errors per C-unit	3.29	77.9	< .01	.73
MLUw	-3.49	79.7	< .001	-.77
Subordination Index	-2.60	74.5	.01	-.57

Note. PGU = Percent Grammatical Utterances. MLUw = Mean Length of Utterance in words. H_a $\mu_0 \neq \mu_1$

Language exposure and ability status were moderately correlated with all four diagnostic LSA measures (positively with PGU, MLUw, and subordination and negatively with errors per C-unit). Language exposure and ability were uncorrelated (*p* = 1.00). Age in months was not correlated with ability, language exposure, or any of the LSA measures. Subordination index was significantly correlated with the other three LSA measures, with a large effect size for MLUw (*r* = .57, *p* = <.001) and medium effect sizes for the grammatical accuracy measures (*r* = -.32-.31, *p* = .003-.004). Length of the language samples measured in utterances was comparable between ability groups, but the TD group produced more total words (*p* = .005) and a greater number of different words (*p* = .001). The TD group outperformed the DLD group on all four diagnostic LSA measures, and differences were statistically significant with moderate to moderately large effect sizes.

Diagnostic Accuracy of LSA Measures With and Without Covariate Adjustment

Pooled and covariate-adjusted ROC results are presented in Table 3.7, and Figures 3.5 through 3.9 show the corresponding plotted curves.

Table 3.7

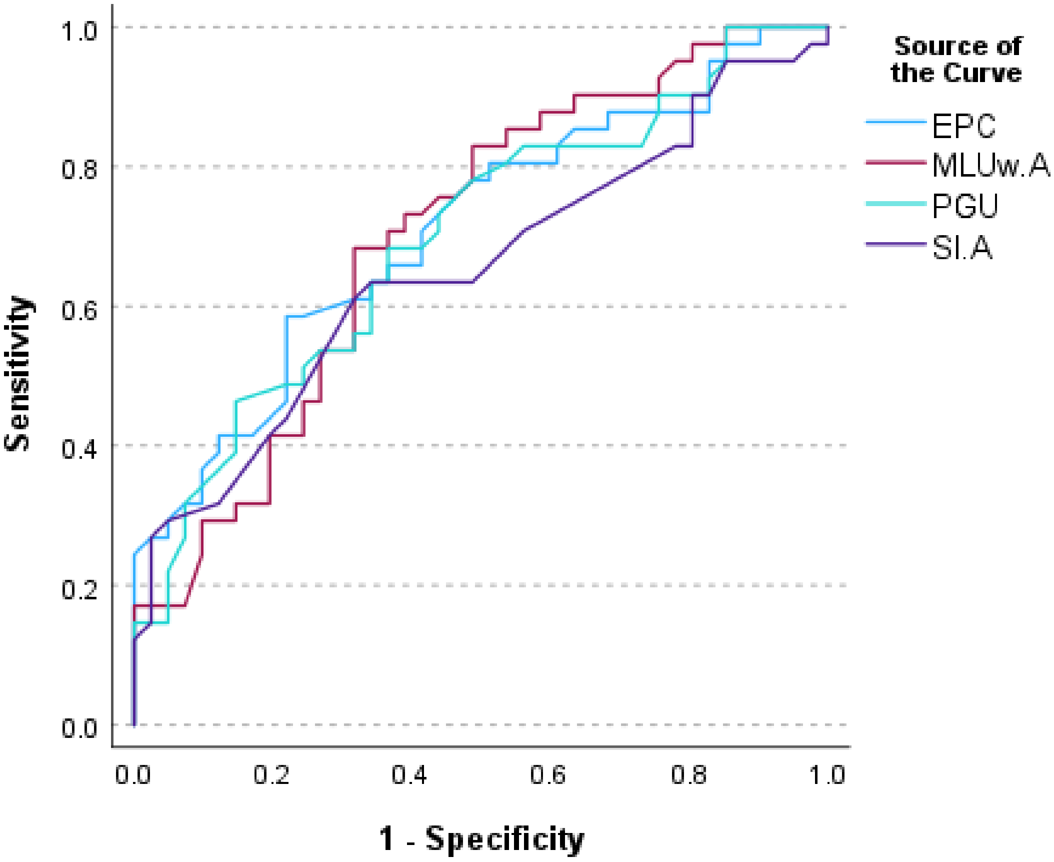
Pooled and Covariate-adjusted ROC Results in Spanish

	AUC	Cutoff	Sensitivity	Specificity	LR+	LR-
Pooled						
PGU	.70 [.59, .80]	56.00%	63%	68%	1.97	.54
EPC	.71 [.59, .82]	0.44	78%	59%	1.9	.37
MLUw	.70 [.59, .81]	4.73	68%	68%	2.13	.47
SI	.65 [.53, .76]	1.06	.66%	.63%	1.78	.54
AROC						
PGU	.70 [.59, .81]	-	-	-	-	-
EPC	.69 [.58, .81]	-	-	-	-	-
MLUw	.71 [.59, .82]	-	-	-	-	-
SI	.64 [.52, .76]	-	-	-	-	-

Note. AUC = area under the curve. LR+ = positive likelihood ratio. LR- = negative likelihood ratio. PGU = Percent Grammatical Utterances. EPC = errors per C-unit. MLUw = Mean length of utterance in words. SI = Subordination Index. Metrics that reached the 80% threshold for sensitivity and/or specificity are in bold text. Em dash = not applicable.

Figure 3.5

Pooled ROC curves for individual Spanish LSA measures



Note. EPC = errors per C-unit. MLUw.A= mean length of utterance in words. PGU = percent grammatical utterances. SI.A = subordination index.

Figure 3.6

Pooled & Covariate-Adjusted ROC Curves for Spanish Percent Grammatical Utterances (PGU)

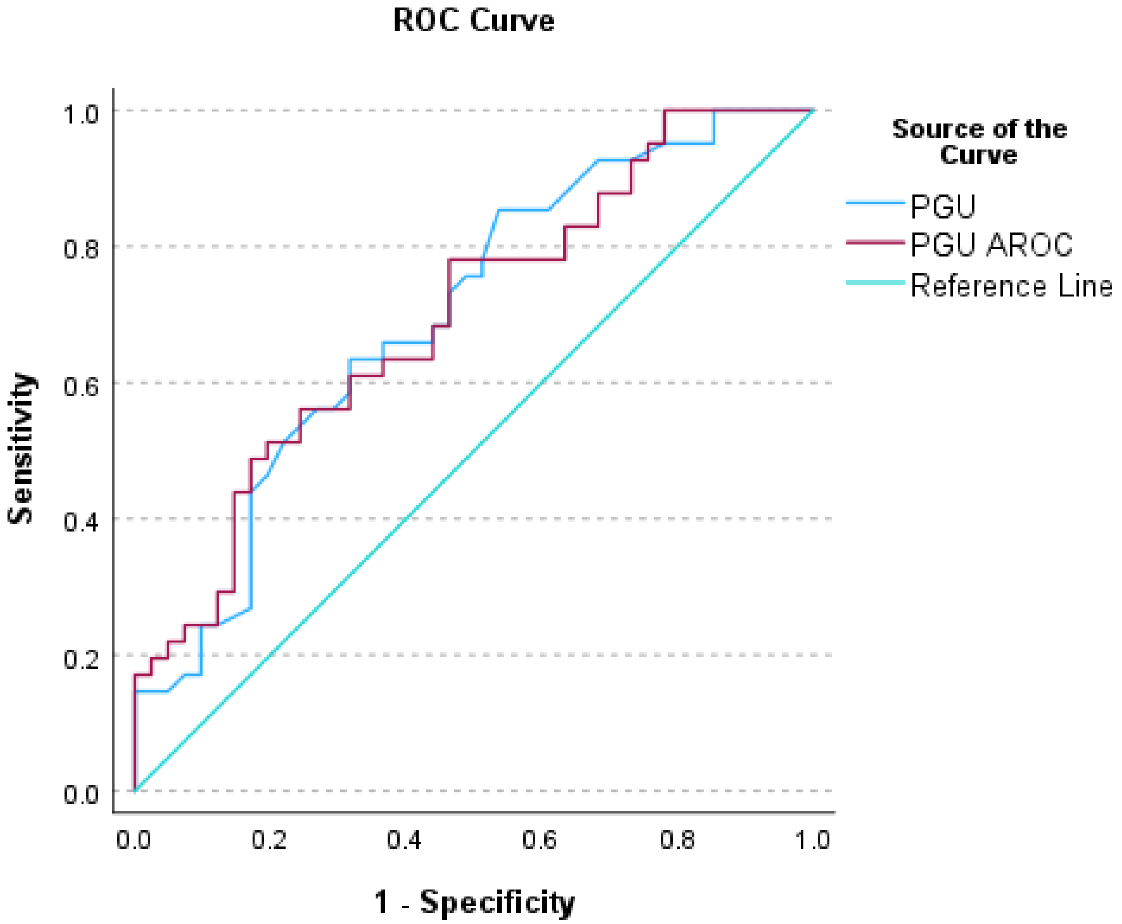


Figure 3.7

Pooled & Covariate-Adjusted ROC Curves for Spanish Errors per C-unit (EPC)

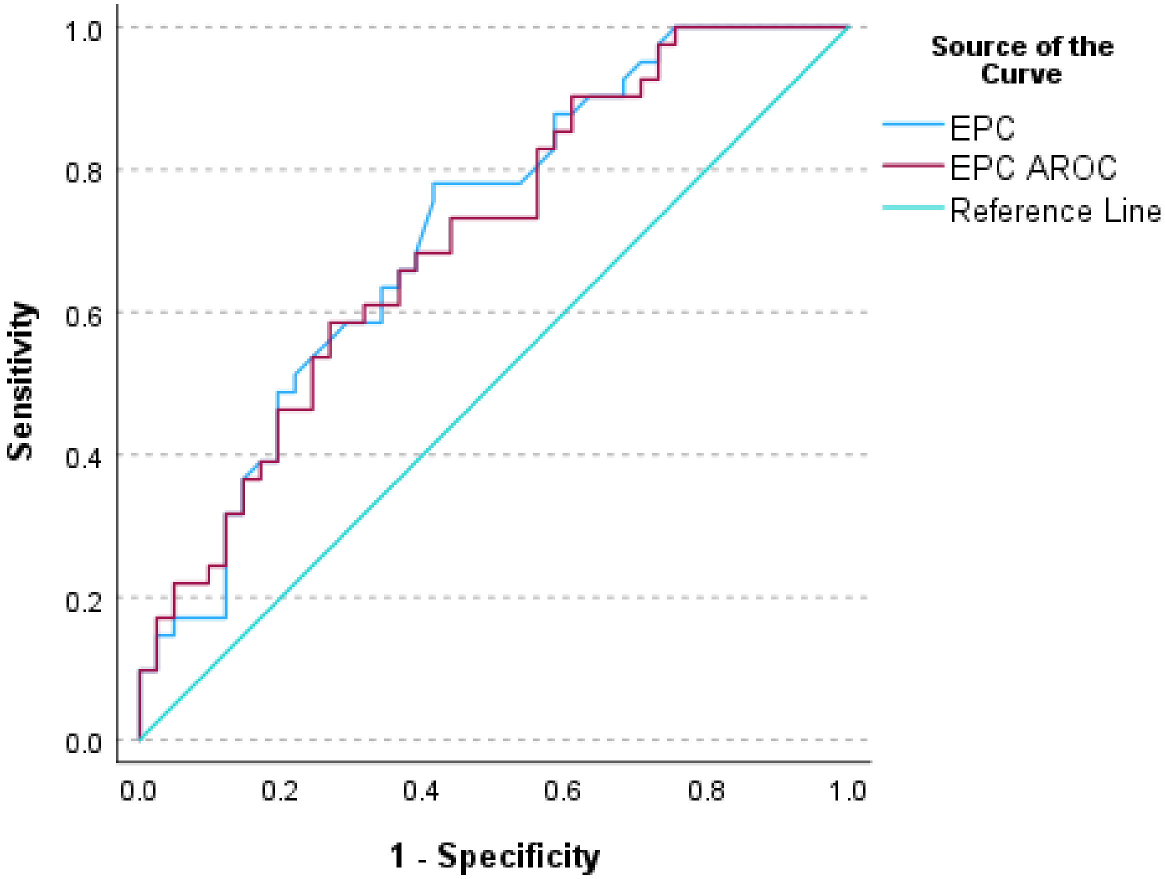


Figure 3.8

Pooled and Covariate-Adjusted ROC Curves for Spanish Mean Length of Utterance in Words (MLUw)

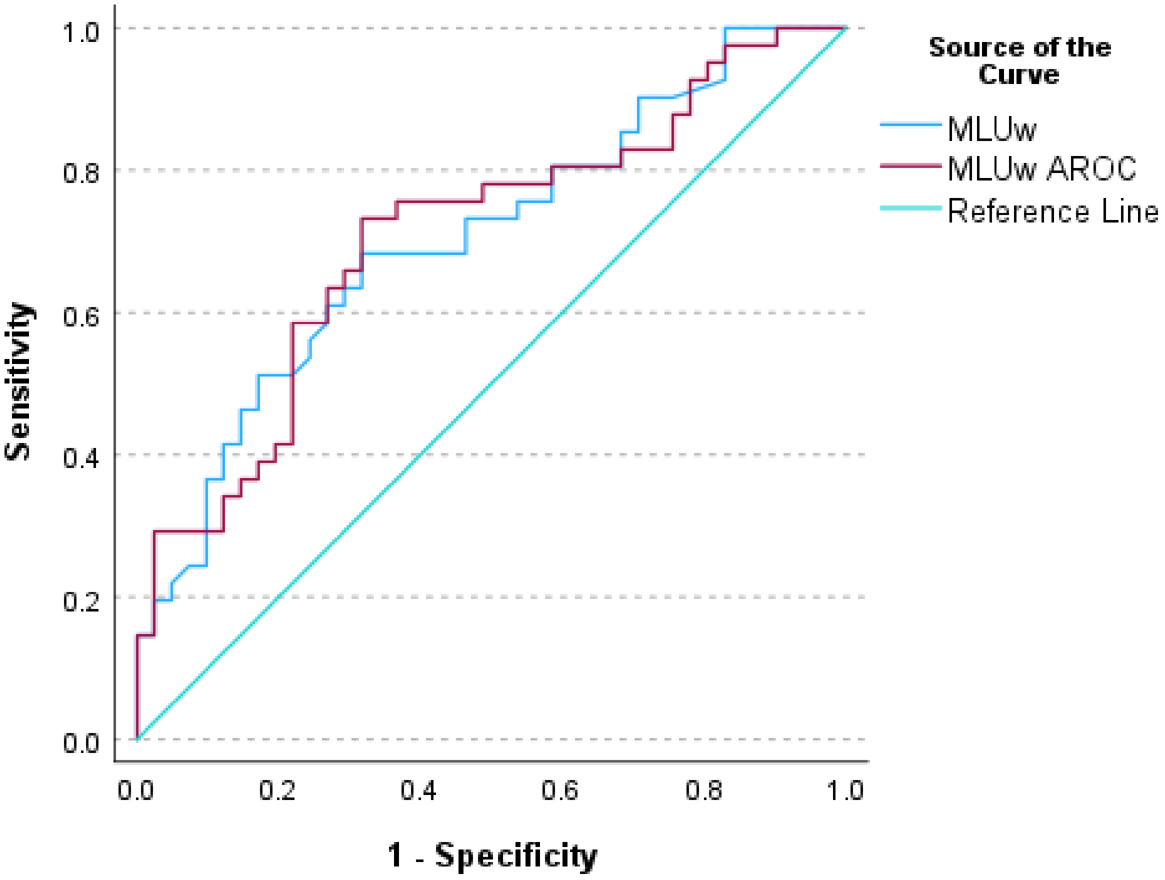
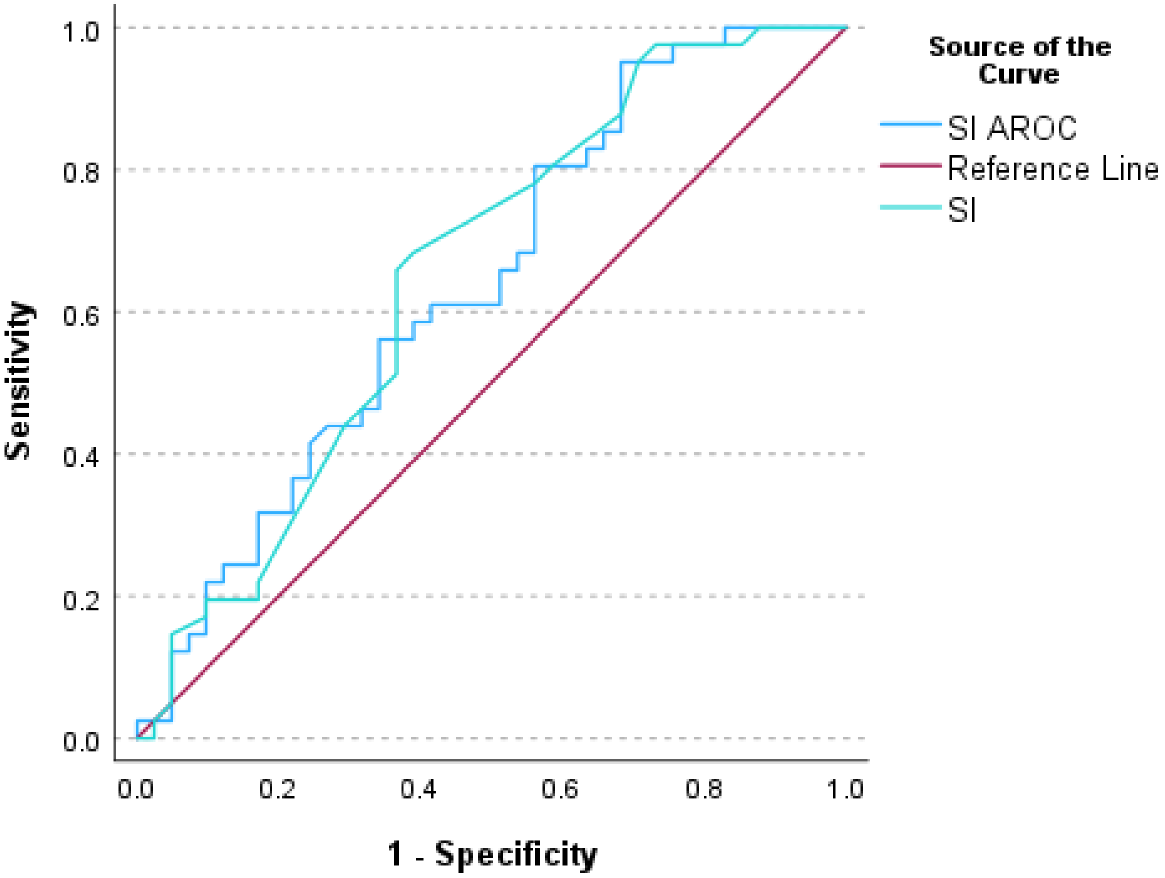


Figure 3.9

Pooled and Covariate-Adjusted ROC Curves for Spanish Subordination Index (SI)



Percent Grammatical Utterances (PGU)

Based on pooled ROC analysis, PGU yielded poor classification accuracy (AUC=.70, 95% CI [.59, .80]), with 63% sensitivity and 68% specificity at the optimal threshold of 56% grammatical utterances. AROC analysis with language exposure as a covariate yielded a comparable AUC (.70, 95% CI [.59, .81]) that was not statistically significant from that of the pooled ROC (*p* = .98).

Errors per C-unit

Pooled ROC analysis of EPC yielded poor classification accuracy (AUC=.71, 95% CI [.59, .82]), with 78% sensitivity and 59% specificity at the optimal threshold of .44 errors per C-unit. AROC analysis with language exposure as a covariate yielded a slightly smaller AUC (.69, 95% CI [.58, .81]). The difference between the pooled and covariate-adjusted curves was not statistically significant (*p* = .67).

Mean Length of Utterance

Pooled ROC analysis of MLU yielded poor classification accuracy (AUC=.70, 95% CI [.59, .81]), with 68% sensitivity and 68% specificity at the optimal threshold of 4.73 words per utterance. AROC analysis with language exposure as a covariate yielded a slightly higher AUC (.71, 95% CI [.59, .82]), but this difference was not statistically significant (*p* = .83).

Subordination Index

Pooled ROC analysis of subordination index yielded poor classification accuracy (AUC=.65, 95% CI [.53, .76]), with 66% sensitivity and 63% specificity at the optimal threshold of 1.06 clauses per utterance. AROC analysis with language exposure as a covariate yielded an AUC that was slightly smaller (.64, 95% CI [.52, .76]), but the difference was not statistically significant (*p* = .55).

Diagnostic Accuracy of Combined LSA Measures

Since covariate-adjusted ROCs yielded negligible differences from pooled ROCs, if any, multiROC analyses were run for the whole participant sample. Table 3.8 summarizes the results, and the multiROC curves are presented in Figure 3.10.

Table 3.8

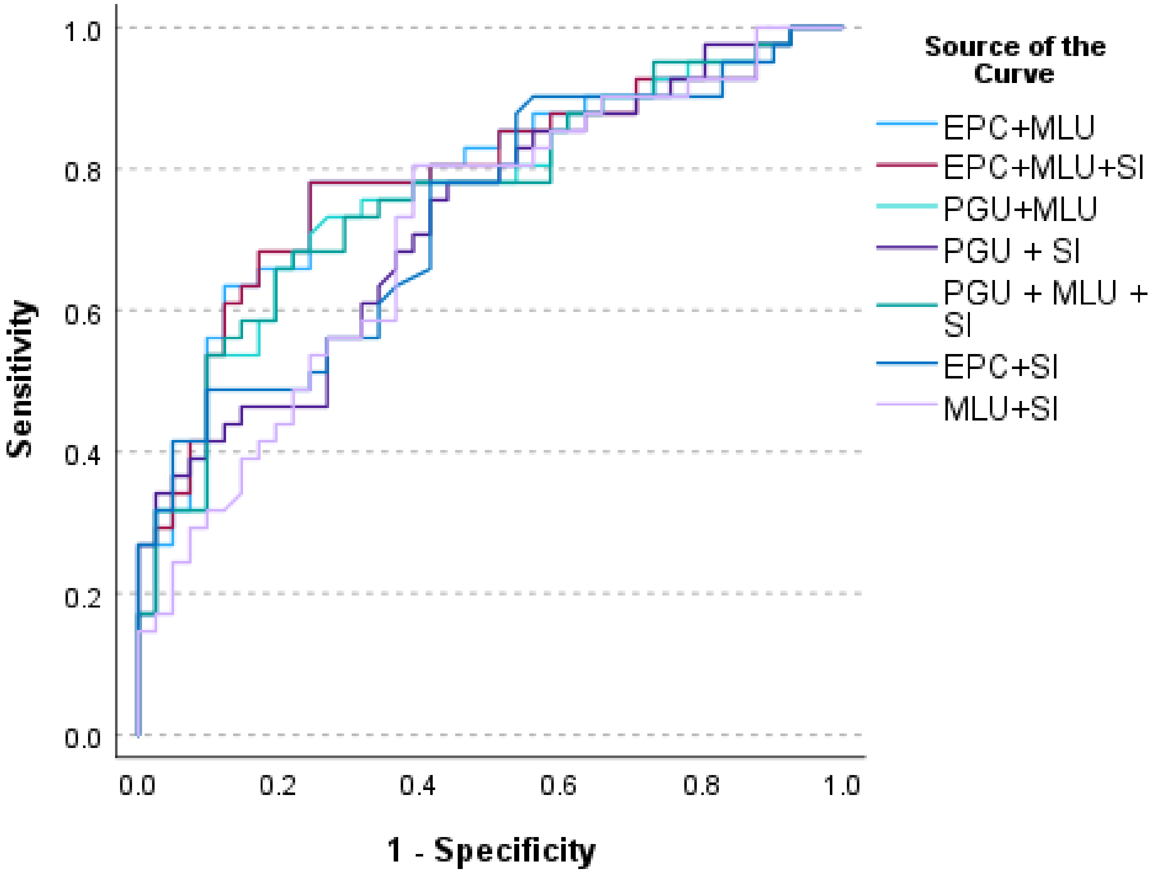
Multivariate ROC Results in Spanish

	AUC	Cutoffs	Sensitivity	Specificity	LR+	LR-
PGU + MLU	.76 [.66, .87]	78%, 3.67	73%	73%	2.7	0.37
PGU + SI	.72 [.61, .83]	52%, 1.08	78%	56%	1.77	0.39
PGU + MLU + SI	.76 [.66, .87]	57%, 4.82, 1.00	68%	78%	3.09	0.41
MLU + SI	.71 [.60, .82]	4.86, 0.96	**81%**	61%	2.08	0.31
EPC + MLU	.79 [.69, .89]	0.27, 3.73	78%	76%	3.25	0.29
EPC + SI	.73 [.62, .84]	0.41, 1.07	49%	**90%**	4.9	0.57
EPC + MLU + SI	.79 [.68, .89]	0.27, 3.73, 0.92	78%	76%	3.25	0.29

Note. AUC = area under the curve. LR+ = positive likelihood ratio. LR- = negative likelihood ratio. PGU = Percent Grammatical Utterances. EPC = errors per C-unit. MLUw = Mean length of utterance in words. SI = Subordination Index. Metrics that reached the 80% threshold for sensitivity and/or specificity are in bold text.

Figure 3.10

Spanish Multivariate ROC Curves



Note. EPC = errors per C-unit. MLUw= mean length of utterance in words. PGU = percent grammatical utterances. SI = subordination index.

All combinations of LSA measures tested outperformed the measures individually based on larger AUCs and improved sensitivity and specificity. The highest AUC for an individual LSA measure was .71, belonging to errors per C-unit, while the AUCs for the multi-measure models ranged from .71 for MLU with subordination index to .79 for errors per C-unit with MLU. It should be noted, however, that differences between AUCs for individual and multivariable models were not statistically significant in the majority of cases (the exceptions being between both pooled and covariate-adjusted subordination index and the errors per C-unit + MLU + subordination index model, *p* = .02). Though diagnostic accuracy improved over individual LSA measures, none of the multivariable models reached the desired threshold of 80% sensitivity and specificity, though errors per C-unit and MLU approached this threshold with 78% sensitivity and 76% specificity (as did adding subordination index to this combination). MLU with subordination index yielded acceptable sensitivity of 81% but inadequate specificity of 61%, while errors per C-unit with subordination index yielded excellent specificity at 90% but poor sensitivity at 49%.

Discussion

There is a critical need for empirical data on the diagnostic accuracy of LSA in Spanish and other languages to provide evidence-based guidance for clinicians, who report a lack of confidence in interpreting LSA results. The aims of the current study were to explore Spanish language sample measures for identifying DLD in bilinguals who represent the range of bilingual experience encountered in clinical practice by evaluating the bounds of their diagnostic accuracy. Our participant sample included 5- and 6- year-olds with 28 to 100% input and output in Spanish. We first used ROC analysis to test diagnostic accuracy across different cutoffs in order to identify the optimal cutoff that maximizes sensitivity and specificity. We then conducted covariate-adjusted analyses to account for the potential effect of Spanish exposure on diagnostic accuracy. Finally, we evaluated diagnostic accuracy for individual LSA measures as well as whether combining the measures would improve classification.

As for the optimal diagnostic accuracy of our four LSA measures of interest - PGU, errors per C-unit, MLUw, and subordination index - diagnostic accuracy was poor even at optimal cutoffs. Sensitivity ranged from 63% for PGU to 78% for errors per C-unit, indicating that children with DLD were frequently misclassified as having typical language. Specificity ranged from 59% for errors per C-unit to 68% for PGU and MLUw, indicating that typically developing children were also frequently misclassified.

While Spanish exposure was correlated with each of the LSA measures, covariate-adjusted models did not indicate that exposure-specific cutoffs would improve diagnostic accuracy. Visual inspection of the data confirms that there is no clear separation between the DLD and TD groups at any point along the exposure continuum on any of the LSA measures, suggesting equally poor diagnostic performance at all exposure levels. Given that correlations and *t*-tests were statistically significant for ability as well, our results reaffirm the importance of validating assessment measures for diagnostic accuracy, as group mean differences and strong correlations do not necessarily translate into good classification.

As with individual measures, combining LSA fell short of achieving both 80% sensitivity and specificity. The combination of errors per C-unit and MLUw approached this threshold though, with 78% sensitivity and 76% specificity for participants across the exposure continuum. MLUw and subordination index together had acceptable sensitivity at

81%, but only 61% specificity. Errors per C-unit and subordination index together had good specificity of 90% but very low sensitivity of 49%.

While previous studies were able to identify DLD in Spanish-English bilingual children using LSA measures, we were unable to replicate past findings of acceptable or near acceptable diagnostic accuracy, particularly for measures of grammatical accuracy (Kapantzoglou et al., 2017; Simón-Cereijido & Gutiérrez-Clellen, 2007). A likely explanation for differences would be our participants' range of Spanish exposure compared to the Spanish-dominant samples in other studies, but our results fell well below previous levels of diagnostic accuracy even for our participants with higher Spanish exposure. Previous studies included slightly younger participants than ours, so inconsistent findings may indicate an age effect on the diagnostic accuracy of the LSA measures examined. Though we adopted similar methods as those used in previous studies in terms of the reference measure, elicitation, and coding, subtle methodological differences may also have contributed to our disparate findings.

To assign our participants to ability groups, we used a single gold standard measure - the Total Language score from the BESA at its empirically-derived cutoff. Prior studies also used the BESA, but in combination with other measures. Simón-Cereijido and Gutiérrez-Clellen (2007) used only the Morphosyntax subtest of the BESA and included parent and teacher concern as well as SLP observation. By using the Total Language score, our sample may have included children whose deficits are primarily in semantics, who may not be identified as accurately by a grammaticality or even a complexity measure. Kapantzoglou et al. (2017) considered both the BESA Morphosyntax and BESA Semantics subtests and included a nonword repetition task; participants who scored below the cutoff on two of the three measures were identified as DLD. They also noted that most of the DLD participants scored below the cutoff on all three measures, while all TD participants scored above the cutoff on all three measures, potentially creating more differentiated groups than ours.

Differences in the length of the language samples across studies may also have impacted classification. Our participants had a mean of 25.78 and 30.66 utterances for the DLD and TD group, respectively, which is much lower than Simón-Cereijido and Gutiérrez-Clellen's (2007) participants (56.8 for DLD and 69.1 for TD), who combined participants' retell and tell samples for analysis. Kapantzoglou et al. (2017) required that samples be at least 50 words long to be included in their study. While we did not exclude any samples based on length, most of our participants produced at least 50 words, but 5 with DLD did not meet this criterion. Additionally, we excluded abandoned, interrupted, unintelligible, and single word utterances, as well as those that contained English words, which further limited sample length. Kapantzoglou et al. only reported excluding abandoned and unintelligible for the MLU calculation. Studies in English have recommended at least 25 and more often 50 utterances for clinical purposes with this age range (Heilmann, 2010; Pavelko et al., 2020), but analyses of diagnostic accuracy have been calculated with as few as 9 utterances (e.g., Guo et al., 2019) Though some LSA measures, such as PGU, appear to be stable relative to sample length (e.g., Eisenberg & Guo, 2015; Heilmann et al., 2010), including shorter samples may account for differences between our and previous findings.

The mean error rates for our sample were higher than those observed in other studies, while MLU and subordination index were more comparable. This may be due to our participants' wider range of language exposure, but it also may be indicative of differences in error coding. Our approach to identifying errors was very similar to that of Kapantzoglou et al. (2017), which involved first determining if the sentence was grammatical and then making the fewest possible changes to make it grammatical. The types of errors we coded for included all the types listed in Simón-Cereijido and Gutiérrez-Clellen (2007), but also additional errors, such as word order errors (Kapantzoglou et al. did not report on specific error types). Another common procedure was judging grammaticality independent of the context of the story or the child's previous utterances, though this may be too generous an interpretation of Simón-Cereijido and Gutiérrez-Clellen's description "independent of discourse context." Our interrater reliability rate was similar to Kapantzoglou et al., who had 86% agreement, but lower than Simón-Cereijido and Gutiérrez-Clellen, who had 93% agreement. Though our procedures were modeled on these studies, any gaps in their descriptions may have left enough room for differences that classification was impacted.

A number of utterances were judged to be grammatical that may have contained errors if context was considered. This was the case especially for sentences with a null subject and those with clitic pronouns that were not double marked with their referent. Absence of an overt subject or referent prevented agreement from being verified. Null subject frequency itself does not have good discrimination accuracy (Grinstead et al., 2018), but frequent use of null subjects could mask errors on more sensitive clinical markers. Judging grammatical accuracy of utterances in isolation may therefore not be a desirable approach when a language's clinical markers are referential in nature.

One major difference between Kapantzoglou et al. (2017) and the current study is the inclusion of lexical *D* in their analysis and best model, while we did not examine any semantics measures nor did we include semantic errors in our coding of PGU and errors per C-unit. Similar to their findings, we did find that combining measures improved diagnostic accuracy compared to individual measures. In particular, errors per C-unit with MLU was the best model, which may capture tradeoffs between accuracy and complexity in spontaneous language, as Simón-Cereijido and Gutiérrez-Clellen also discuss. Simón-Cereijido and Gutiérrez-Clellen also found grammaticality to be the measure with the highest diagnostic accuracy at 79% sensitivity and 100% specificity both with and without MLU. While the highest number of participants was accurately identified with this morphosyntactic model, the authors found that some participants were only identified accurately with a syntactic-semantic model that included MLU, theme arguments, and ditransitive verbs. Diagnostic accuracy studies of English LSA have generally found semantics measures to be poor (e.g., Charest et al., 2020) unless included with a grammaticality measure (e.g., Fletcher & Peters, 1984). Though the hallmark deficit of DLD is morphosyntax, all domains of language are affected, including semantics. Semantic measures such as lexical diversity may tap into clinically informative characteristics that enhance the effectiveness of grammaticality alone in Spanish. For example, the diversity of words used might uniquely identify children who achieve high grammaticality by frequently repeating utterances that are within their morphosyntactic repertoire throughout their narrative, as we observed anecdotally during coding.

Future Directions

Our null findings, which differed from similar previous studies, highlight the importance of replication studies. Exploratory analyses, such as ours, must be confirmed with another similar sample of children. Given the extent of variation in language performance that we are trying to account for in looking at typical and atypical bilingual development, a larger sample would help ensure enough statistical power at all intervals of the exposure range.

Since the current study was not an exact replication of previous studies, future research should investigate the extent to which procedural differences impact diagnostic accuracy. This could include validation of additional LSA measures with a diverse group of bilinguals, incorporating codeswitching in the analysis, and testing classification reliability across different sample lengths and coding variations. For example, we judged grammaticality of each utterance apart from the context of the story or the rest of the sample, but this may have prevented the detection of informative error patterns and may not align with typical clinical practice. Alternatives to this procedure could be tested empirically in future research, though it would require that language samples be carefully indexed to pages of the book or episodes of the story when they are collected.

Conclusion

Empirical evidence of the diagnostic accuracy of language sample analysis is critical for guiding clinical practice, especially given SLPs' feelings of inadequate expertise. Previous studies have found that measures of grammatical accuracy and lexical diversity have over 80% sensitivity and specificity, or very nearly meet this threshold, for Spanish-dominant 4- and 5-year-olds, but the current study found that 5- and 6-year-old bilinguals with a wide range of Spanish exposure, even the best model - errors per C-unit paired with MLU - fell short of the 80% threshold. Future research should explore whether procedural modifications, such as inclusion of different LSA measures or adaptation of coding rules, can achieve better diagnostic accuracy for this age range.

CONCLUSION

To address the persistent trend of misidentification of DLD in bilinguals, the three studies presented in this dissertation investigated the diagnostic accuracy of language sample analysis (LSA) in Spanish and English - a familiar but underutilized assessment tool that is often recommended as a gold standard. Guidance on interpreting LSA results has often been based on developmental norms and group means rather than on empirical evidence of accurate classification. The first study was a systematic review of research examining the diagnostic accuracy of LSA measures in English in order to compile the evidence and identify the best LSA measures for diagnosis of DLD. Several measures or combinations of measures were found to have at least 80% sensitivity and specificity for ages 3 to 10, and at least one measure had at least 90% sensitivity and specificity for each year within that range except for 6-year-olds. One critical limitation of the review for the broader purpose of this dissertation was that only two studies were identified that examined bilingual speakers of English. This motivated the second study, which evaluated the diagnostic accuracy of English LSA for Spanish-English bilingual 5- and 6-year-olds with specific attention to the influence of language exposure. The measures that were chosen - percent grammatical utterances (PGU), mean length of utterance in words (MLUw), errors per C-unit, and subordination index - had prior evidence of good diagnostic accuracy, either alone or in combination with other measures, as found in the systematic review. A parallel search of the literature on Spanish LSA also identified few studies that have analyzed diagnostic accuracy, and results were inconsistent. Study 3 was conducted to further explore whether accounting for language exposure would improve classification accuracy of the same set of LSA measures in Spanish - PGU, MLUw, errors per C-unit, and subordination - for use with children who represent a continuum of bilingualism.

In study 2, when analyses pooled participants at all language exposure levels together, none of the LSA measures reached the desired threshold for diagnostic accuracy individually or in combination. Adjusting for English exposure as a covariate resulted in slight improvement in classification using the grammatical accuracy measures, and follow-up analyses by exposure group revealed excellent diagnostic accuracy using these measures with participants who have at least 70% English exposure. Combining LSA measures for participants with 30 to 70% exposure improved diagnostic accuracy compared to individual measures, but while sensitivity was excellent, specificity was poor for all models. These results highlight two findings of the systematic review. First, diagnostic performance of a given measure cannot be assumed across varieties of English, much like with African American English (Oetting et al., 2021), especially if English exposure is not accounted for. However, the measures that are best for monolingual English also appear to be best for bilingual speakers relative to other measures. It seems worthwhile to continue exploring the generalizability of diagnostically accurate LSA measures identified in our systematic review to the bilingual population, being sure to account for English exposure when doing so.

In Spanish, on the other hand, the 80% threshold was not met by any of the LSA measures individually or in combination for any segment of the exposure range. Grammatical accuracy measures were effective in English for our participants with the most exposure, and previous studies of Spanish LSA also found the best diagnostic accuracy using such measures. The model that came closest to adequate sensitivity and specificity consisted of errors per T-unit with MLUw (78% and 76%, respectively), but we did not find any that could be considered adequate for clinical purposes. Furthermore, while factoring language exposure into English analyses revealed improved diagnostic performance of the grammatical accuracy measures, exposure did not have an effect on diagnostic accuracy in Spanish.

The results of all three studies support the clinical utility of LSA, which is considered the gold standard of language assessment and especially suited to culturally and linguistically diverse speakers. Though evidence of diagnostic accuracy of English and Spanish measures in the empirical studies was not as compelling as previous studies with monolinguals per the systematic review or with Spanish-dominant speakers (e.g., Kapantzoglou et al., 2017), we identified measures or models that had good diagnostic accuracy for a subset of participants or that approached a desirable threshold of diagnostic accuracy. Further exploration could reveal more promising modifications or alternatives. For example, a major limitation of the current work is that only a single-language approach was used, where a cross-linguistic or best language approach may have produced better results, particularly for participants with more balanced language exposure. Another consideration that should inform future studies is the amount of agency the speaker has in what they produce or, perhaps more importantly, avoid producing. Structured probes and tests, which can force a specific type of response (or at least an informative non-response), have achieved good diagnostic accuracy by directly targeting the clinical markers that differentiate DLD from typically developing language. In contrast, in language samples, children can choose the words and structures they are most comfortable with and strategize around more challenging ones, so diagnostic LSA models may need to be flexible enough to be able to capture the linguistic tradeoffs that inevitably result from those strategies (Simón-Cereijido & Gutiérrez-Clellen, 2007).

The primary motivation for the current studies was to provide clinical guidance for implementing LSA with bilingual children. Based on our findings, two recommendations can be made. For 5- to 6-year-old Spanish-English bilinguals who are exposed to English at least 70% of the time based on their current input and output, calculating either PGU or errors per C-unit from a narrative retell offers good diagnostic accuracy of over 90% sensitivity and specificity. Additional instruments, such as parent questionnaires, observations, and dynamic assessment, should of course be conducted to comply with IDEA and to address other non-diagnostic goals of the assessment (e.g., eliciting parent input, generating a profile of strengths and needs, developing a treatment plan). For children with less English exposure, our findings support that assessment in both languages is necessary. When analyzing Spanish language samples, error rate and MLU appear to be the most informative measures, but given their limited diagnostic accuracy, results should be corroborated with other Spanish assessment data.

Just as clinical markers of DLD are language-specific, our results demonstrate that the same LSA measure, generalizable though it may seem, is not necessarily as effective in classifying DLD in different languages, as we found with PGU and errors per C-unit in English versus Spanish. Considering again a child's agency in producing their narrative, an important parameter to account for may be the frequency of obligatory use of the clinical markers in that language. It was not uncommon for children to produce an entire sample in Spanish without using clitic pronouns or subjunctive verbs, for example, while one would be hard pressed to produce an English narrative while entirely avoiding tense and agreement markers. Furthermore, coding procedures that are quite adequate for capturing performance on clinical markers in one language might miss critical information in another language, such as judging the grammaticality of sentences independent of context when clinical markers are referential, as in the case of Spanish clitic pronouns. Diagnostic indicators thus appear to be language-specific as well, and procedural modifications or alternative LSA measures, such as lexical diversity, may be needed to achieve comparable classification accuracy across languages.

Usage-based theory (UBT) posits that language input - the quantity of input as well as its structural properties - determines the pace of mastery (e.g., Tomasello, 2001), and the moderate correlations we found between all four LSA measures and exposure to the target language are consistent with this claim. In both English and Spanish, greater current exposure to the language was associated with higher grammatical accuracy and longer, more complex utterances. We did observe substantial variation in performance at similar levels of exposure, even within ability groups, which may be related to how we operationalized input. According to UBT, mastery of a form occurs when a child accumulates a critical mass of relevant exemplars from their input, which for bilinguals, implies two dimensions of input - relative amount in each language over time (Gathercole, 2007; Paradis, 2010). Our participants' cumulative exposure may have been informative for accounting for variation given similar current exposure (Bedore et al. 2016). As for ability, group differences between our DLD and TD participants generally support the prediction that children with DLD need a greater amount of input before reaching the same level of mastery (Jacobson & Yu, 2018). The DLD group's means indicated lower levels of mastery than TD matches on all LSA measures, with trendlines that either paralleled the TD group's improving performance with greater language exposure or showed less dramatic change. Individual-level data on each measure often deviated from this pattern, however, suggesting a more complex relationship between input, ability, and spontaneous language production.

Limitations & Future Directions

The studies in this dissertation involved exploratory analyses of a selection of LSA measures with a limited number of matched pairs. Though the sample size was larger than many previous LSA studies, the inclusion of such a range of bilingual experience likely requires greater statistical power evenly distributed along the continuum of language exposure than we achieved with our sample. Subsequent confirmatory studies could also provide converging evidence that would lend confidence in our findings.

Our results suggest that expanding the selection of LSA measures analyzed and/or making language-specific adaptations to procedures may reveal measures and models that have more optimal diagnostic accuracy. We made specific choices in terms of LSA measures and analysis, but alternatives are certainly possible and should be explored empirically to determine the robustness of our findings. Study design decisions such as these should be informed by properties of clinical markers that may differ across languages and that are uniquely relevant to analysis of spontaneous language samples, such as the level(s) of discourse at which they operate and to what extent they are optional at the utterance level. Given the clinical focus of such research, it is also advisable to design these studies with implementation in mind by ensuring that procedures are intuitive and do not require extensive training to override natural responses.

A single-language approach was used to evaluate diagnostic accuracy of measures in English and in Spanish separately with overlapping but different participant samples. Thus, we were not able to directly compare performance across languages on each LSA measure, nor were we able to test the diagnostic accuracy of cross-linguistic LSA models. Assessment in both of a bilingual child's language is considered best practice, typically required in most practice settings, and may yield the best diagnostic accuracy using LSA when relative language exposure is more balanced. An informative follow-up to these studies would therefore be to evaluate the best method of incorporating dual language LSA data for diagnosis of DLD. Furthermore, research should extend these studies to speakers of other home languages, especially those in which a valid standardized test is not currently available.

Summary

The three studies of this dissertation reviewed extant research on the diagnostic accuracy of language sample analysis (LSA) and explored the potential of four LSA measures - percent grammatical utterances, errors per C-unit, mean length of utterance in words, and subordination index - to identify developmental language disorder in Spanish-English bilingual with a wide range of language exposure. The measures of grammatical accuracy we tested yielded excellent diagnostic accuracy in English for participants with at least 70% exposure to the language, while Spanish errors per C-unit with MLU only approached the 80% threshold for clinical use. Results highlight the importance of considering the language of administration and a child's exposure to that language in selecting and interpreting LSA measures for diagnosis. While we were not able to identify a model in either language that was accurate across the range of language exposure, even when applying exposure-adjusted cutoffs, our findings demonstrate that LSA can be diagnostically informative for bilinguals when key parameters are considered. Future research should continue to explore diagnostic LSA for diverse bilingual children to identify valid and accurate procedures that can be applied in clinical practice and improve identification of DLD in this population.

References

Abbot-Smith, K., & Behrens, H. (2006). How known constructions influence the acquisition of other constructions: The german passive and future constructions. *Cognitive Science*, *30*(6), 995–1026. https://doi.org/10.1207/s15516709cog0000_61

Abedi, J. (2004). The no child left behind act and English language learners: assessment and accountability issues. *Educational Researcher*, *33*(1), 4–14. https://doi.org/10.3102/0013189X033001004

American Speech-Language-Hearing Association. (2022). *2021 Demographic profile of ASHA members providing multilingual services.* . http://www.asha.org.

Andreou, G., & Lemoni, G. (2020). Narrative skills of monolingual and bilingual preschool and primary school children with developmental language disorder (DLD): A systematic review. *Open Journal of Modern Linguistics*, *10*(05), 429–458. https://doi.org/10.4236/ojml.2020.105026

Aram, D. M., Morris, R., & Hall, N. E. (1993). Clinical and research congruence in identifying children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, *36*(3), 580–591. https://doi.org/10.1044/jshr.3603.580

Arias, G., & Friberg, J. (2017). Bilingual language assessment: Contemporary versus recommended practice in American schools. *Language, Speech, and Hearing Services in Schools*, *48*(1), 1–15. https://doi.org/10.1044/2016_LSHSS-15-0090

Arnold, B. R., & Matus, Y. E. (2000). Test translation and cultural equivalence methodologies for use with diverse populations. In *Handbook of multicultural mental health* (pp. 121–136). Elsevier. https://doi.org/10.1016/B978-012199370-2/50008-0

Artiles, A. J., Kozleski, E. B., Trent, S. C., Osher, D., & Ortiz, A. (2010). Justifying and explaining disproportionality, 1968–2008: A critique of underlying views of culture. *Exceptional Children*, *76*(3), 279–299. https://doi.org/10.1177/001440291007600303

Artiles, A J, Rueda, R., Salazar, J. J., & Higareda, I. (2002). English-language learner representation in special education in California urban school districts. *Racial Inequity in Special Education*, 117–136.

Ash, A. C., & Redmond, S. M. (2014). Using finiteness as a clinical marker to identify language impairment. *Perspectives on Language Learning and Education*, *21*(4), 148–158. https://doi.org/10.1044/lle21.4.148

Barrera, M. (2006). Roles of definitional and assessment models in the identification of new or second language learners of English for special education. *Journal of Learning Disabilities*, *39*(2), 142–156. https://doi.org/10.1177/00222194060390020301

Bedore, L. M., & Leonard, L. B. (1998). Specific language impairment and grammatical morphology: A discriminant function analysis. *Journal of Speech, Language, and Hearing Research*, *41*(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. https://doi.org/10.1044/1092-4388(2001/072)

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*, *15*(3), 616–629. https://doi.org/10.1017/S1366728912000090

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

Betz, S. K., Eickhoff, J. R., & Sullivan, S. F. (2013). Factors influencing the selection of standardized tests for the diagnosis of specific language impairment. *Language, Speech, and Hearing Services in Schools*, *44*(2), 133–146. https://doi.org/10.1044/0161-1461(2012/12-0093)

Bishop, D. V. M., Snowling, M. J., Thompson, P. A., Greenhalgh, T., & and the 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, and Allied Disciplines*, *58*(10), 1068–1080. https://doi.org/10.1111/jcpp.12721

Blom, E., Paradis, J., & Duncan, T. S. (2012). Effects of input properties, vocabulary size, and L1 on the development of third person singular -s in child L2 English. *Language Learning*, *62*(3), 965–994. https://doi.org/10.1111/j.1467-9922.2012.00715.x

Blom, E., & Paradis, J. (2015). Sources of individual differences in the acquisition of tense inflection by English second language learners with and without specific language impairment. *Applied Psycholinguistics*, *36*(04), 953–976. https://doi.org/10.1017/S014271641300057X

Blood, G. W., Thomas, E. A., Ridenour, J. S., & Qualls, C. D. (2002). Job stress in speech-language pathologists working in rural, suburban, and urban schools: Social support and frequency of interactions. *Contemporary Issues in Communication Science and Disorders*, *29*(Fall), 132–140. https://doi.org/10.1044/cicsd_29_F_132

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

Bracken, B. A., & Barona, A. (1991). State of the art procedures for translating, validating and using psychoeducational tests in cross-cultural assessment. *School Psychology International*, *12*(1–2), 119–132. https://doi.org/10.1177/0143034391121010

Brownlie, E. B., Beitchman, J. H., Escobar, M., Young, A., Atkinson, L., Johnson, C., Wilson, B., & Douglas, L. (2004). Early language impairment and young adult delinquent and aggressive behavior. *Journal of Abnormal Child Psychology*, *32*(4), 453–467. https://doi.org/10.1023/b:jacp.0000030297.91759.74

Bybee, J. (2008). Usage-based grammar and second language acquisition. In P. Robinson & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 226–246). Routledge. https://doi.org/10.4324/9780203938560-18

Bybee, J. L. (2006). From usage to grammar: The mind’s response to repetition. *Language*, *82*(4), 711–733. https://doi.org/10.1353/lan.2006.0186

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. https://doi.org/10.1044/0161-1461(2007/020)

Castilla-Earls, A., Bedore, L., Rojas, R., Fabiano-Smith, L., Pruitt-Lord, S., Restrepo, M. A., & Peña, E. (2020). Beyond scores: Using converging evidence to determine speech and language services eligibility for dual language learners. *American Journal of Speech-Language Pathology / American Speech-Language-Hearing Association*, *29*(3), 1116–1132. https://doi.org/10.1044/2020_AJSLP-19-00179

Castilla-Earls, A., Pérez-Leroux, A. T., Fulcher-Rood, K., & Barr, C. (2021). Morphological errors in Spanish-speaking bilingual children with and without developmental language disorders. *Language, Speech, and Hearing Services in Schools*, *52*(2), 497–511. https://doi.org/10.1044/2020_LSHSS-20-00017

Castilla-Earls, A. P., 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*(5), 1147–1173. https://doi.org/10.1017/S0142716415000521

Charest, M., Skoczylas, M. J., & Schneider, P. (2020). Properties of lexical diversity in the narratives of children with typical language development and developmental language disorder. *American Journal of Speech-Language Pathology / American Speech-Language-Hearing Association*, *29*(4), 1866–1882. https://doi.org/10.1044/2020_AJSLP-19-00176

Cioè-Peña, M. (2017). The intersectional gap: How bilingual students in the United States are excluded from inclusion. *International Journal of Inclusive Education*, *21*(9), 906–919. https://doi.org/10.1080/13603116.2017.1296032

Collier, V. P., & Thomas, W. P. (2017). Validating the power of bilingual schooling: Thirty-Two years of large-scale, longitudinal research. *Annual Review of Applied Linguistics*, *37*, 203–217. https://doi.org/10.1017/S0267190517000034

Collins, B. A., O'Connor, E. E., Suárez-Orozco, C., Nieto-Castañon, A., & Toppelberg, C. O. (2014). Dual language profiles of Latino children of immigrants: Stability and change over the early school years. *Applied Psycholinguistics*, *35*(3), 581–620. https://doi.org/10.1017/S0142716412000513

Coloma, C. J., Araya, C., Quezada, C., Pavez, M. M., & Maggiolo, M. (2016). Grammaticality and complexity of sentences in monolingual Spanish-speaking children with specific language impairment. *Clinical Linguistics & Phonetics*, *30*(9), 649–662. https://doi.org/10.3109/02699206.2016.1163420

Conti-Ramsden, G., & Botting, N. (2008). Emotional health in adolescents with and without a history of specific language impairment (SLI). *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *49*(5), 516–525. https://doi.org/10.1111/j.1469-7610.2007.01858.x

Conti-Ramsden, G., Durkin, K., Toseeb, U., Botting, N., & Pickles, A. (2018). Education and employment outcomes of young adults with a history of developmental language disorder. *International Journal of Language & Communication Disorders*, *53*(2), 237–255. https://doi.org/10.1111/1460-6984.12338

Conti-Ramsden, G. (2003). Processing and linguistic markers in young children with specific language impairment (SLI). *Journal of Speech, Language, and Hearing Research*, *46*(5), 1029–1037. https://doi.org/10.1044/1092-4388(2003/082)

Costanza-Smith, A. (2010). The clinical utility of language samples. *Perspectives on Language Learning and Education*, *17*(1), 9–15. https://doi.org/10.1044/lle17.1.9

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)

Dollaghan, C. A. (2004). Evidence-based practice in communication disorders: What do we know, and when do we know it? *Journal of Communication Disorders*, *37*(5), 391–400. https://doi.org/10.1016/j.jcomdis.2004.04.002

Drobatz, K. J. (2009). Measures of accuracy and performance of diagnostic tests. *Journal of Veterinary Cardiology : The Official Journal of the European Society of Veterinary Cardiology*, *11 Suppl 1*, S33-40. https://doi.org/10.1016/j.jvc.2009.03.004

Dubois, P., St-Pierre, M.-C., Desmarais, C., & Guay, F. (2020). Young adults with developmental language disorder: A systematic review of education, employment, and independent living outcomes. *Journal of Speech, Language, and Hearing Research*, *63*(11), 3786–3800. https://doi.org/10.1044/2020_JSLHR-20-00127

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. https://doi.org/10.1044/jshr.3903.643

Eisenberg, S., & Guo, L.-Y. (2016). Using language sample analysis in clinical practice: Measures of grammatical accuracy for identifying language impairment in preschool and school-aged children. *Seminars in Speech and Language*, *37*(2), 106–116. https://doi.org/10.1055/s-0036-1580740

Eisenberg, S. L., & Guo, L.-Y. (2015). Sample size for measuring grammaticality in preschool children from picture-elicited language samples. *Language, Speech, and Hearing Services in Schools*, *46*(2), 81–93. https://doi.org/10.1044/2015_LSHSS-14-0049

Ellis, N. (2008). Usage-based and form-focused language acquisition: The associative learning of constructions, learned attention, and the limited L2 endstate. In P. Robinson & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 382–415). Routledge. https://doi.org/10.4324/9780203938560-24

Evans, J. (1996). Plotting the complexities of language sample analysis: Linear and non-linear dynamical models of assessment. *Assessment of Communication and Language*, *6*, 207–256.

Ferney Harris, S., Prater, M. A., Dyches, T. T., & Allen Heath, M. (2009). Job stress of school-based speech-language pathologists. *Communication Disorders Quarterly*, *30*(2), 103–111. https://doi.org/10.1177/1525740108323856

Fletcher, P., & Peters, J. (1984). Characterizing language impairment in children. *Language Testing*, *1*(1), 33–49. https://doi.org/10.1177/026553228400100104

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

Fulcher-Rood, K., Castilla-Earls, A., & Higginbotham, J. (2019). Diagnostic decisions in child language assessment: Findings from a case review assessment task. *Language, Speech, and Hearing Services in Schools*, *50*(3), 385–398. https://doi.org/10.1044/2019_LSHSS-18-0044

Fulcher-Rood, K., Castilla-Earls, A. P., & Higginbotham, J. (2018). School-based speech-language pathologists' perspectives on diagnostic decision making. *American Journal of Speech-Language Pathology / American Speech-Language-Hearing Association*, *27*(2), 796–812. https://doi.org/10.1044/2018_AJSLP-16-0121

Gathercole, V. C. M. (2007). Miami and North Wales, so far and yet so near: A constructivist account of morphosyntactic development in bilingual children. *International Journal of Bilingual Education and Bilingualism*, *10*(3), 224–247. https://doi.org/10.2167/beb442.0

Gathercole, V. C. M., & Hoff, E. (2007). Input and the acquisition of language: Three questions. In E. Hoff & M. Shatz (Eds.), *Blackwell handbook of language development* (pp. 107–127). Blackwell Publishing Ltd. https://doi.org/10.1002/9780470757833.ch6

Genesee, F., Lindholm-Leary, K., Saunders, W., & Christian, D. (2005). English language learners in U.S. schools: An overview of research findings. *Journal of Education for Students Placed at Risk*, *10*(4), 363–385. https://doi.org/10.1207/s15327671espr1004_2

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)

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)

Grinstead, J., Baron, A., Vega-Mendoza, M., De la Mora, J., Cantu´-Sa´nchez, 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. https://doi.org/10.1044/1092-4388(2012/11-0289)

Grinstead, J., Lintz, P., Pratt, A., Vega-Mendoza, M., De la Mora, J., Cantú-Sánchez, M., & Flores-Avalos, B. (2018). Null subject occurrence in monolingual Spanish SLI: A discriminant function analysis. *Languages*, *3*(2), 17. https://doi.org/10.3390/languages3020017

Guiberson, M., & Atkins, J. (2012). Speech-language pathologists' preparation, practices, and perspectives on serving culturally and linguistically diverse children. *Communication Disorders Quarterly*, *33*(3), 169–180. https://doi.org/10.1177/1525740110384132

Guiberson, M., Rodríguez, B. L., & Zajacova, A. (2015). Accuracy of telehealth-administered measures to screen language in Spanish-speaking preschoolers. *Telemedicine Journal and E-Health*, *21*(9), 714–720. https://doi.org/10.1089/tmj.2014.0190

Guo, L.-Y., Eisenberg, S., Schneider, P., & Spencer, L. (2019). Percent grammatical utterances between 4 and 9 years of age for the Edmonton Narrative Norms Instrument:

Reference data and psychometric properties. *American Journal of Speech-Language Pathology*, *28*(4), 1448–1462. https://doi.org/10.1044/2019_AJSLP-18-0228

Guo, L.-Y., Eisenberg, S., Schneider, P., & Spencer, L. (2020). Finite verb morphology composite between age 4 and age 9 for the Edmonton Narrative Norms Instrument: Reference data and psychometric properties. *Language, Speech, and Hearing Services in Schools*, *51*(1), 128–143. https://doi.org/10.1044/2019_LSHSS-19-0028

Guo, L.-Y., & Schneider, P. (2016). Differentiating school-aged children with and without language impairment using tense and grammaticality measures from a narrative task. *Journal of Speech, Language, and Hearing Research*, *59*(2), 317–329. https://doi.org/10.1044/2015_JSLHR-L-15-0066

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. https://doi.org/10.1044/1092-4388(2006/087)

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. https://doi.org/10.1044/1092-4388(2007/068)

Gutiérrez-Clellen, V. F., & Simón-Cereijido, G. (2009). Using language sampling in clinical assessments with bilingual children: Challenges and future directions. *Seminars in Speech and Language*, *30*(4), 234–245. https://doi.org/10.1055/s-0029-1241722

Heilmann, J., Nockerts, A., & Miller, J. F. (2010). Language sampling: Does the length of the transcript matter? *Language, Speech, and Hearing Services in Schools*, *41*(4), 393–404. https://doi.org/10.1044/0161-1461(2009/09-0023)

Heilmann, J. J. (2010). Myths and realities of language sample analysis. *Perspectives on Language Learning and Education*, *17*(1), 4. https://doi.org/10.1044/lle17.1.4

Heilmann, J. J., & Westerveld, M. F. (2013). Bilingual language sample analysis: Considerations and technological advances. *Journal of Clinical Practice in Speech-Language Pathology*, *15*(2), 87–93.

Hernandez, A., Li, P., & MacWhinney, B. (2005). The emergence of competing modules in bilingualism. *Trends in Cognitive Sciences*, *9*(5), 220–225. https://doi.org/10.1016/j.tics.2005.03.003

Hewitt, L. E., Hammer, C. S., Yont, K. M., & Tomblin, J. B. (2005). Language sampling for kindergarten children with and without SLI: Mean length of utterance, IPSYN, and NDW. *Journal of Communication Disorders*, *38*(3), 197–213. https://doi.org/10.1016/j.jcomdis.2004.10.002

Hoffman, L. M. (2009). The utility of school-age narrative microstructure indices: INMIS and the proportion of restricted utterances. *Language, Speech, and Hearing Services in Schools*, *40*(4), 365–375. https://doi.org/10.1044/0161-1461(2009/08-0017)

Huang, R.-J., Hopkins, J., & Nippold, M. A. (1997). Satisfaction with standardized language testing. *Language, Speech, and Hearing Services in Schools*, *28*(1), 12–29. https://doi.org/10.1044/0161-1461.2801.12

Jacobson, P. F, & Yu, Y. H. (2018). Changes in English past tense use by bilingual school-age children with and without developmental language disorder. *Journal of Speech,* *Language, and Hearing Research*, *61*(10), 2532–2546. https://doi.org/10.1044/2018_JSLHR-L-17-0044

Jacobson, P. F., & Schwartz, R. G. (2002). Morphology in incipient bilingual Spanish-speaking preschool children with specific language impairment. *Applied Psycholinguistics*, *23*(01). https://doi.org/10.1017/S0142716402000024

Janes, H., Longton, G., & Pepe, M. S. (2009). Accommodating covariates in receiver operating characteristic analysis. *The Stata Journal: Promoting Communications on Statistics and Stata*, *9*(1), 17–39. https://doi.org/10.1177/1536867X0900900102

Janes, H., & Pepe, M. S. (2008). Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: An old concept in a new setting. *American Journal of Epidemiology*, *168*(1), 89–97. https://doi.org/10.1093/aje/kwn099

Jasso, J., McMillen, S., Anaya, J. B., Bedore, L. M., & Peña, E. D. (2020). The utility of an English semantics measure for identifying developmental language disorder in Spanish-English bilinguals. *American Journal of Speech-Language Pathology*, *29*(2), 776–788. https://doi.org/10.1044/2020_AJSLP-19-00202

Johnson, A. (2019). The effects of English learner classification on high school graduation and college attendance. *AERA Open*, *5*(2), 233285841985080. https://doi.org/10.1177/2332858419850801

Kangas, S. E. N. (2014). When special education trumps ESL: An investigation of service delivery for ELLs with disabilities. *Critical Inquiry in Language Studies*, *11*(4), 273–306. https://doi.org/10.1080/15427587.2014.968070

Kangas, S. E. N. (2018). Why working apart doesn't work at all: Special education and English learner teacher collaborations. *Intervention in School and Clinic*, *54*(1), 105345121876246. https://doi.org/10.1177/1053451218762469

Kapantzoglou, M., Fergadiotis, G., & Restrepo, M. A. (2017). Language sample analysis and elicitation technique effects in bilingual children with and without language impairment. *Journal of Speech, Language, and Hearing Research*, *60*(10), 2852–2864. https://doi.org/10.1044/2017_JSLHR-L-16-0335

Katz, L. A., Maag, A., Fallon, K. A., Blenkarn, K., & Smith, M. K. (2010). What makes a caseload (un)manageable? School-based speech-language pathologists speak. *Language, Speech, and Hearing Services in Schools*, *41*(2), 139–151. https://doi.org/10.1044/0161-1461(2009/08-0090)

Klatte, I. S., van Heugten, V., Zwitserlood, R., & Gerrits, E. (2022). Language sample analysis in clinical practice: Speech-language pathologists' barriers, facilitators, and needs. *Language, Speech, and Hearing Services in Schools*, *53*(1), 1–16. https://doi.org/10.1044/2021_LSHSS-21-00026

Klee, T., Gavin, W. J., & Stokes, S. F. (2017). Utterance length and lexical diversity in American- and British-English speaking children: What is the evidence for a clinical marker of SLI? In Rhea Paul (Ed.), *Language disorders from a developmental perspective: Essays in honor of Robin S. Chapman* (pp. 103–140). Psychology Press. https://doi.org/10.4324/9781315092041-4

Kraemer, R., & Fabiano-Smith, L. (2017). Language assessment of Latino English learning children: A records abstraction study. *Journal of Latinos and Education*, *16*(4), 1–10. 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 / American Speech-Language-Hearing Association*, *12*(1), 73–91. https://doi.org/10.1044/1058-0360(2003/054)

Langacker, R. W. (1987). *Foundations of Cognitive Grammar: Volume I: Theoretical Prerequisites* (illustrated, reprint ed.). Stanford University Press.

Law, J., Rush, R., Schoon, I., & Parsons, S. (2009). Modeling developmental language difficulties from school entry into adulthood: Literacy, mental health, and employment outcomes. *Journal of Speech, Language, and Hearing Research*, *52*(6), 1401–1416. https://doi.org/10.1044/1092-4388(2009/08-0142)

Lazewnik, R., Creaghead, N. A., Smith, A. B., Prendeville, J.-A., Raisor-Becker, L., & Silbert, N. (2019). Identifiers of language impairment for Spanish-English dual language learners. *Language, Speech, and Hearing Services in Schools*, *50*(1), 126–137. https://doi.org/10.1044/2018_LSHSS-17-0046

Leonard, L. B. (2014a). *Children with Specific Language Impairment*. The MIT Press. https://doi.org/10.7551/mitpress/9152.001.0001

Leonard, L. B. (2014b). Specific language impairment across languages. *Child Development Perspectives*, *8*(1), 1–5. https://doi.org/10.1111/cdep.12053

Liles, B. Z., Duffy, R. J., Merritt, D. D., & Purcell, S. L. (1995). Measurement of narrative discourse ability in children with language disorders. *Journal of Speech, Language, and Hearing Research*, *38*(2), 415–425. https://doi.org/10.1044/jshr.3802.415

Mayer, M. (1967). *A Boy, a Dog, and a Frog*. Dial Press.

Mayer, M. (1973). *Frog On His Own*. Dial Press.

Mayer, M. (1975). *One Frog Too Many*. Dial Press.

McGregor, K. K. (2020). How we fail children with developmental language disorder. *Language, Speech, and Hearing Services in Schools*, *51*(4), 981–992. https://doi.org/10.1044/2020_LSHSS-20-00003

Miller, J., & Chapman, R. (2008). Systematic analysis of language transcripts (SALT). *Research Version*.

Miller, J. F., Andriacchi, K., & Nockerts, A. (2016). Using language sample analysis to assess spoken language production in adolescents. *Language, Speech, and Hearing Services in Schools*, *47*(2), 99–112. https://doi.org/10.1044/2015_LSHSS-15-0051

Morgan, P. L., Farkas, G., Hillemeier, M. M., Li, H., Pun, W. H., & Cook, M. (2017). Cross-cohort evidence of disparities in service receipt for speech or language impairments. *Exceptional Children*, 001440291771834. https://doi.org/10.1177/0014402917718341

Norbury, C. F., Gooch, D., Wray, C., Baird, G., Charman, T., Simonoff, E., Vamvakas, G., & Pickles, A. (2016). The impact of nonverbal ability on prevalence and clinical presentation of language disorder: Evidence from a population study. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *57*(11), 1247–1257. https://doi.org/10.1111/jcpp.12573

Oetting, J. B., Rivière, A. M., Berry, J. R., Gregory, K. D., Villa, T. M., & McDonald, J. (2021). Marking of tense and agreement in language samples by children with and without specific language impairment in African American English and Southern White English: Evaluation of scoring approaches and cut scores across structures. *Journal of Speech,* *Language, and Hearing Research*, *64*(2), 491–509. https://doi.org/10.1044/2020_JSLHR-20-00243

Ooi, C. C.-W., & Wong, A. M.-Y. (2012). Assessing bilingual Chinese-English young children in Malaysia using language sample measures. *International Journal of Speech-Language Pathology*, *14*(6), 499–508. https://doi.org/10.3109/17549507.2012.712159

Paradis, J., & Blom, E. (2016). Do early successive bilinguals show the English L2 pattern of precocious BE acquisition? *Bilingualism: Language and Cognition*, *19*(03), 630–635. https://doi.org/10.1017/S1366728915000267

Paradis, J. (2005). Grammatical morphology in children learning English as a second language. *Language, Speech, and Hearing Services in Schools*, *36*(3), 172–187. https://doi.org/10.1044/0161-1461(2005/019)

Paradis, J. (2008). Tense as a clinical marker in English L2 acquisition with language delay/impairment. In B. Haznedar & E. Gavruseva (Eds.), *Current Trends in Child Second Language Acquisition: A generative perspective* (Vol. 46, pp. 337–356). John Benjamins Publishing Company. https://doi.org/10.1075/lald.46.17par

Paul, R. (1995). *Language Disorders from Infancy through Adolescence: Assessment and Intervention.* MOSBY.

Pavelko, S. L., Owens, R. E., Ireland, M., & Hahs-Vaughn, D. L. (2016). Use of language sample analysis by school-based SLPs: Results of a nationwide survey. *Language, Speech, and Hearing Services in Schools*, *47*(3), 246–258. https://doi.org/10.1044/2016_LSHSS-15-0044

Pavelko, S. L., Price, L. R., & Owens, R. E. (2020). Revisiting reliability: Using Sampling Utterances and Grammatical Analysis Revised (SUGAR) to compare 25- and 50-

ttterance language samples.*Language, Speech, and Hearing Services in Schools*, *51*(3), 778–794. https://doi.org/10.1044/2020_LSHSS-19-00026

Pawlowska, M. (2014). Evaluation of three proposed markers for language impairment in English: A meta-analysis of diagnostic accuracy studies. *Journal of Speech, Language, and Hearing Research*, *57*(6), 2261–2273. https://doi.org/10.1044/2014_JSLHR-L-13-0189

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

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, 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. (2007). Lost in translation: Methodological considerations in cross-cultural research. *Child Development*, *78*(4), 1255–1264. https://doi.org/10.1111/j.1467-8624.2007.01064.x

Pezold, M. J., Imgrund, C. M., & Storkel, H. L. (2020). Using computer programs for language sample analysis. *Language, Speech, and Hearing Services in Schools*, *51*(1), 103–114. https://doi.org/10.1044/2019_LSHSS-18-0148

Plante, E., & Vance, R. (1994). Selection of preschool language tests: A data-based approach. *Language, Speech, and Hearing Services in Schools*, *25*(1), 15. https://doi.org/10.1044/0161-1461.2501.15

Plante, E. (1998). Criteria for SLI: The Stark and Tallal legacy and beyond. *Journal of Speech, Language, and Hearing Research*, *41*(4), 951. https://doi.org/10.1044/jslhr.4104.951

Plante, E. (2004). Evidence based practice in communication sciences and disorders. *Journal of Communication Disorders*, *37*(5), 389–390. https://doi.org/10.1016/j.jcomdis.2004.04.001

Polat, N., Zarecky-Hodge, A., & Schreiber, J. B. (2016). Academic growth trajectories of ELLs in NAEP data: The case of fourth- and eighth-grade ELLs and non-ELLs on mathematics and reading tests. *The Journal of Educational Research*, *109*(5), 541–553. https://doi.org/10.1080/00220671.2014.993461

Prath, S. (2018). The how and why of collecting a language sample. *ASHA Leader Live*. https://leader.pubs.asha.org/do/10.1044/the-how-and-why-of-collecting-a-language-sample/full/

Price, L. H., Hendricks, S., & Cook, C. (2010). Incorporating computer-aided language sample analysis into clinical practice. *Language, Speech, and Hearing Services in Schools*, *41*(2), 206–222. https://doi.org/10.1044/0161-1461(2009/08-0054)

Ramos, M. N., Collins, P., & Peña, E. D. (2022). Sharpening Our Tools: A Systematic Review to Identify Diagnostically Accurate Language Sample Measures*. Journal of Speech, Language, and Hearing Research*, *65*(10), 3890-3907.

Redmond, S. M., Ash, A. C., Christopulos, T. T., & Pfaff, T. (2019). Diagnostic accuracy of sentence recall and past tense measures for identifying children's language impairments. *Journal of Speech, Language, and Hearing Research*, *62*(7), 2438–2454. https://doi.org/10.1044/2019_JSLHR-L-18-0388

Restrepo, M. A. (1998). Identifiers of predominantly Spanish-speaking children with language impairment. *Journal of Speech, Language, and Hearing Research*, *41*(6), 1398–1411. https://doi.org/10.1044/jslhr.4106.1398

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

Rice, M. L. (2003). A unified model of specific and general language delay: Grammatical tense as a clinical marker of unexpected variation. In Y. Levy & J. C. Schaeffer (Eds.), *Language competence across populations* (pp. 63-95). Psychology Press.

Rojas, R., & Iglesias, A. (2009). Making a case for language sampling. *ASHA Leader*, *14*(3), 10. https://doi.org/10.1044/leader.FTR1.14032009.10

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. https://doi.org/10.1177/0022219408326221

Santhanam, S. Priya, Gilbert, C. L., & Parveen, S. (2019). Speech-language pathologists' use of language interpreters with linguistically diverse clients: A nationwide survey study. *Communication Disorders Quarterly*, *40*(3), 131–141. https://doi.org/10.1177/1525740118779975

Schwob, S., Eddé, L., Jacquin, L., Leboulanger, M., Picard, M., Oliveira, P. R., & Skoruppa, K. (2021). Using nonword repetition to identify developmental language disorder in monolingual and bilingual children: A systematic review and meta-analysis. *Journal of*

Speech, Language, and Hearing Research, *64*(9), 3578–3593. https://doi.org/10.1044/2021_JSLHR-20-00552

Selin, C. M., Rice, M. L., Girolamo, T., & Wang, C. J. (2019). Speech-language pathologists' clinical decision making for children with specific language impairment. *Language, Speech, and Hearing Services in Schools*, *50*(2), 283–307. https://doi.org/10.1044/2018_LSHSS-18-0017

Shahmahmood, T. M., Jalaie, S., Soleymani, Z., Haresabadi, F., & Nemati, P. (2016). A systematic review on diagnostic procedures for specific language impairment: The sensitivity and specificity issues. *Journal of Research in Medical Sciences*, *21*, 67. https://doi.org/10.4103/1735-1995.189648

Shultz, E. K. (1995). Multivariate receiver-operating characteristic curve analysis: Prostate cancer screening as an example. *Clinical Chemistry*, *41*(8 Pt 2), 1248–1255. https://doi.org/10.1093/clinchem/41.8.1248

Silva, P. A., McGee, R., & Williams, S. M. (1983). Developmental language delay from three to seven years and its significance for low intelligence and reading difficulties at age seven. *Developmental Medicine and Child Neurology*, *25*(6), 783–793. https://doi.org/10.1111/j.1469-8749.1983.tb13847.x

Simon-Cereijido, G., & Gutiérrez-Clellen, V. F. (2007). Spontaneous language markers of Spanish language impairment. *Applied Psycholinguistics*, *28*(02). https://doi.org/10.1017/S0142716407070166

Skiba, R. J., Simmons, A. B., Ritter, S., Gibb, A. C., Rausch, M. K., Cuadrado, J., & Chung, C.-G. (2008). Achieving equity in special education: History, status, and current challenges. *Exceptional Children*, *74*(3), 264–288. https://doi.org/10.1177/001440290807400301

Smyk, E. (2012). *Second language proficiency in sequential bilingual children with and without primary language impairment* (Doctoral dissertation). https://https://keep.lib.asu.edu/items/151135..

Souto, S. M., Leonard, L. B., & Deevy, P. (2014). Identifying risk for specific language impairment with narrow and global measures of grammar. *Clinical Linguistics & Phonetics*, *28*(10), 741–756. https://doi.org/10.3109/02699206.2014.893372

Spaulding, T. J., Plante, E., & Farinella, K. A. (2006). Eligibility criteria for language impairment: Is the low end of normal always appropriate? *Language, Speech, and Hearing Services in Schools*, *37*(1), 61–72. https://doi.org/10.1044/0161-1461(2006/007)

Stark, R. E., & Tallal, P. (1981). Selection of children with specific language deficits. *The Journal of Speech and Hearing Disorders*, *46*(2), 114. https://doi.org/10.1044/jshd.4602.114

Stockman, I. J. (1996). The promises and pitfalls of language sample analysis as an assessment tool for linguistic minority children. *Language, Speech, and Hearing Services in Schools*, *27*(4), 355–366. https://doi.org/10.1044/0161-1461.2704.355

St Clair, M. C., Forrest, C. L., Yew, S. G. K., & Gibson, J. L. (2019). Early risk factors and emotional difficulties in children at risk of developmental language disorder: A population cohort study. *Journal of Speech, Language, and Hearing Research*, *62*(8), 2750–2771. https://doi.org/10.1044/2018_JSLHR-L-18-0061

Sullivan, A. L. (2011). Disproportionality in special education identification and placement of English language learners. *Exceptional Children*, *77*(3), 317–334. https://doi.org/10.1177/001440291107700304

Sylvan, L. (2014). Speech-language services in public schools: How policy ambiguity regarding eligibility criteria impacts speech-language pathologists in a litigious and resource constrained environment. *Journal of the American Academy of Special Education Professionals*.

The jamovi project (2024). jamovi (Version 2.5) [Computer Software]. Retrieved from https://www.jamovi.org

Thordardottir, E. (2015). The relationship between bilingual exposure and morphosyntactic development. *International Journal of Speech-Language Pathology*, *17*(2), 97–114. https://doi.org/10.3109/17549507.2014.923509

Tomasello, M. (2001). First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics*, *11*(1–2). https://doi.org/10.1515/cogl.2001.012

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

Tomblin, J. B., Records, N. L., & Zhang, X. (1996). A system for the diagnosis of specific language impairment in kindergarten children. *Journal of Speech and Hearing Research*, *39*(6), 1284–1294. https://doi.org/10.1044/jshr.3906.1284

U.S. Department of Education. EDFacts Data Warehouse (EDW), 2014–15. SEA File C141, LEP Enrolled. Extracted April 3, 2017.

Waitoller, F. R., Artiles, A. J., & Cheney, D. A. (2010). The miner's canary. *The Journal of Special Education*, *44*(1), 29–49. https://doi.org/10.1177/0022466908329226

Winstanley, M., Webb, R. T., & Conti-Ramsden, G. (2021). Developmental language disorders and risk of recidivism among young offenders. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *62*(4), 396–403. https://doi.org/10.1111/jcpp.13299

Wu, X., Li, J., Ayutyanont, N., Protas, H., Jagust, W., Fleisher, A., Reiman, E., Yao, L., Chen, K., & Alzheimer's Disease Neuroimaging Initiative. (2013). The receiver operational characteristic for binary classification with multiple indices and its application to the neuroimaging study of Alzheimer's disease. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *10*(1), 173–180. https://doi.org/10.1109/TCBB.2012.141

Wulff, S., & Ellis, N. C. (2018). Usage-based approaches to second language acquisition. In D. Miller, F. Bayram, J. Rothman, & L. Serratrice (Eds.), *Bilingual cognition and language: The state of the science across its subfields* (Vol. 54, pp. 37–56). John Benjamins Publishing Company. https://doi.org/10.1075/sibil.54.03wul

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32–35.

Youngstrom, E. A. (2014). A primer on receiver operating characteristic analysis and diagnostic efficiency statistics for pediatric psychology: We are ready to ROC. *Journal of Pediatric Psychology*, *39*(2), 204–221. https://doi.org/10.1093/jpepsy/jst062

Young, A. R., Beitchman, J. H., Johnson, C., Douglas, L., Atkinson, L., Escobar, M., & Wilson, B. (2002). Young adult academic outcomes in a longitudinal sample of early identified language impaired and control children. *Journal of Child Psychology and Psychiatry*, *43*(5), 635–645. https://doi.org/10.1111/1469-7610.00052

APPENDIX

Supplemental Material S1

Table S1

Diagnostic Accuracy of LSA Measures

Measure	Age	Task	Sensitivity	Specificity	Overall
Mainstream English Speakers					
Morphosyntax: Accuracy					
DSS Sentence Point [a]	4;0-4;11	Play	93%	94%	--
	5;0-5;10	Play	100%	100%	--
Errors per C-unit [b]	6	Narrative tell	91%	82%	85%
	8	Narrative tell	94%	80%	84%
Finite Verb Morphology Composite [a,b,c,d,e,f,g,h]	3;0-3;11[c]	Play (50 utterances)	67%	100	83%
	3;0-3;11[c]	Play (100 utterances)	83%	89%	86%
	3;0-3;11 [c] (SPELT<87)	Play (100 utterances)	100%	89%	94%
	3;0-3;6[c]	Play (100 utterances)	100%	75%	88%
	3;6-3;11[c]	Play (100 utterances)	70%	100%	85%
	3;7-5;9[d]	Conversation Picture description	84%	100%	--
	4;0-4;11[a]	Play	93%	94%	--
	5;0-5;10[a]	Play	91%	93%	--
	4;0-4;6 [e]	Play	100%	100%	--
	5;0-5;6 [e]	Play	92%	93%	--
	5;5-9:9 [f]	Conversation Expository	50%	86%	68%
	5;11-6;3[g] (MLU<-1SD)	Play	26%	90%	--

Measure	Age	Task	Sensitivity	Specificity	Overall
	5;11-6;3 [g] (PPVT-R<-1SD)	Play	35%	92%	--
	5;11-6;3 [g] (NWR<-1SD)	Play	30%	91%	--
	4;0-4;11[h]	Narrative tell	92%	94%	94%
	5;0-5;11[h]	Narrative tell	100%	90%	92%
	6:0-6;11 [b,h]	Narrative tell	82%	90%	89%
	7;0-7;11[h]	Narrative tell	85%	86%	86%
	8;0-8;11[b,h]	Narrative tell	76%	80%	79%
	9;0-9;11[h]	Narrative tell	80%	76%	77%
Grammaticality & Utterance Length Instrument [i]	4;0-6;11	Narrative retell	83%	92%	--
Noun Morphology Composite [d,f]	3;7-5;9[d]	Conversation Picture description	79%	100%	--
	5;5-9:9 [f]	Conversation Expository	54%	86%	70%
Percent Grammatical Utterances/C-units [b,j,k]	3;0-3;11[j]	Picture description	100%	88%	--
	4;0-4;11[k]	Narrative tell	83%	96%	94%
	5;0-5;11[k]	Narrative tell	100%	82%	86%
	6;0-6;11[b,k]	Narrative tell	82%	90%	89%
	7;0-7;11[k]	Narrative tell	92%	88%	89%
	8;0-8;11[b,k]	Narrative tell	88%	84%	85%
	9;0-9;11[k]	Narrative tell	90%	90%	90%

Measure	Age	Task	Sensitivity	Specificity	Overall
Percent Sentence Point [j]	3;0-3;11	Picture description	100%	82%	--
Proportion ‘restricted’ utterances [l]	8-10	Narrative tell	83%	88%	--
Percent Verb Tense Usage [j]	3;0-3;11	Picture description	100%	82%	--
Morphosyntax: Proficiency					
DSS Total [a]	4;0-4;11	Play	79%	94%	--
	5;0-5;10	Play	72%	87%	--
Mean tense/agreement [a]	4;0-4;11	Play	79%	81%	--
	5;0-5;10	Play	64%	80%	--
Mean Top 5 tense/agreement [a]	4;0-4;11	Play	71%	69%	--
	5;0-5;10	Play	73%	87%	--
Tense Marker Total [e]	4;0-4;6	Play	83%	87%	--
	5;0 to 5;6	Play	77%	80%	--
Tense/Agreement Productivity Score [c,e]	3;0-3;5[c]	Play (100 utterances)	88%	88%	88%
	3;6-3;11[c]	Play (100 utterances)	90%	80%	85%
	3;0-3;11[c]	Play (50 utterances)	94%	50%	72%
	3;0-3;11[c]	Play (100 utterances)	89%	78%	83%
	3;0-3;11[c] SPELT<87	Play (100 utterances)	100%	100%	100%
	4;0-4;6 [e]	Play	67%	87%	--

Measure	Age	Task	Sensitivity	Specificity	Overall
	5;0-5;6 [e]	Play	80%	80%	--
Morphosyntax: Length					
Clauses per Sentence [m]	3;0-7;11	Conversation	83%	91%	--
MLU (morphemes) [d,f]	3;7-5;9 [d]	Conversation Picture description	95%	89%	--
	5;5-9;9 [f]	Conversation Expository	72%	80%	76%
MLU (SUGAR) [m]	3;0-7;11	Conversation	86%	86%	--
Words per Sentence [m]	3;0-7;11	Conversation	83%	84%	--
Semantics					
Moving Average Type-Token Ratio [n]	4-9	Narrative tell	26%	88%	--
Number of Different Words (200w) [n]	4-9	Narrative tell	20%	90%	--
Number of Different Words (41u) [n]	4-9	Narrative tell	42%	91%	--
Number of Different Words (nar) [n]	4-9	Narrative tell	34%	92%	--
Total Number of Words [m]	3;0-7;0	Conversation	86%	85%	--
Pragmatics/Discourse					
Story Grammar [o]	4;0-9;11	Narrative retell	70%	84%	81%
TNW+Turns+References+Expansions[p]	8;4-13;2	Expository (Map task)	75%	60%	--
Composite Models					
10 SALT measures [q]	3;0-13;6	Conversation	69%	84%	--
	3;0-5;11	Conversation	87%	87%	--
	6;0-9;11	Conversation	80%	85%	--
	10;0-13;6	Conversation	77%	82%	--

Measure	Age	Task	Sensitivity	Specificity	Overall
Cohesive ties + % grammatical T-units + subordinate clauses/T-unit +words/subordinate clause [r]	7;6-10;6	Narrative retell	–	–	98%
	8;6-12;6	Narrative retell	–	–	79%
	9;0-11;4	Narrative retell	–	–	83%
MLU + Clauses Per Sentence [m]	3;0-7;11	Conversation	97%	82%	--
MLU + lexical diversity *D* + age [s]	2;0-4;0	Play	86%	91%	–
MLU-m + NDW + IPSyn total [t]	5;5-6;7	Conversation	–	–	74%
MLU + % structural errors + age [u]	2;6-6;11	Play	81%	83%	–
Noun Composite + MLU [d,f]	5;5-9;9 [f]	Conversation Expository	72%	84%	78%
	3;7-5;9 [d]	Conversation Picture description	89%	100%	--
Noun Composite + Verb Composite [d,f]	5;5-9;9 [f]	Conversation Expository	62%	86%	74%
	3;7-5;9 [d]	Conversation Picture description	84%	100%	--
Noun Composite + Verb Composite + MLU [d,f]	5;5-9;9 [f]	Conversation Expository	72%	88%	80%
	3;7-5;9 [d]	Conversation Picture description	89%	95%	--
Unmarked Verbs +Verb Types [v]	3;4-6;11	Play Picture description Narrative retell	89%	90%	--
Verb Composite + MLU [d,f]	5;5 – 9;9 [f]	Conversation Expository	74%	84%	79%
	3;7-5;9 [d]	Conversation Picture description	95%	95%	--
VP errors + Stage 1 Utterances + Age + 3-element NP [w]	2;0-4;2	Play	91%	92%	92%

African American English (AAE) and Southern White English (SWE) Speakers

Morphosyntax: Accuracy

Measure	Age	Task	Sensitivity	Specificity	Overall
Nonmainstream patterns [x]					
Full model (35 patterns)	4-6	Play	87%	94%	90%
Reduced model (4 patterns)	4-6	Play	74%	90%	84%
SWE-specific model (5 patterns)	4-6	Play	87%	95%	--
SAAE-specific model (3 patterns)	4-6	Play	75%	82%	--
Tense & Agreement Forms [y]					
Unmodified Scoring	5	Play	70%	64%	67%
Modified Scoring	5	Play	72%	74%	73%
Strategic Scoring	5	Play	43%	64%	54%
Past Tense (Strategic)	5	Play	70%	85%	77%
Past Tense (Strategic): SWE	5	Play	89%	89%	89%
Past Tense (Strategic): AAE	5	Play	83%	77%	80%
Morphosyntax: Proficiency					
DSS Total [z]	<6	Play	63%	100%	--
IPSyn Total [z]	<6	Play	45%	NR	--
Bilingual English Speakers					
Spanish/English Bilinguals					
MLU+Grammaticality+Number of Different Words +% mazes [aa]	5;3-8	Narrative retell	-	-	83%
Cantonese/English Bilinguals					

Measure	Age	Task	Sensitivity	Specificity	Overall
MLU+IPSyn+Lexical diversity [ab]	3;8-5;11	Conversation Play	78%	78%	78%

Note. DSS = Developmental Sentence Scoring. IPSyn = Index of Productive Syntax. MLU = Mean Length of Utterance. SUGAR = Sampling Utterances and Grammatical Analysis Revised. TNW = Total Number of Words. SALT = Systematic Analysis of Language Transcripts. IPSyn = Index of Productive Syntax. NDW = Number of Different Words. VP = Verb Phrase. NP = Noun Phrase. SWE = Southern White English. SAAE = Southern African American English. NR = Not reported. [a] Souto et al., 2014. [b] Guo & Schneider, 2016. [c] Guo & Eisenberg, 2014. [d] Bedore & Leonard, 1998. [e] Gladfelter & Leonard, 2013. [f] Moyle et al., 2011. [g] Rudolph et al., 2019. [h] Guo et al., 2020. [i] Castilla-Earls & Fulcher-Rood, 2018. [j] Eisenberg & Guo, 2013. [k] Guo et al., 2019. [l] Hoffman, 2009. [m] Pavelko & Owens, 2019. [n] Charest et al., 2020. [o] Schneider et al., 2006. [p] Scheffel, 1997. [q] Heilmann et al., 2010. [r] Liles et al., 1995. [s] Klee et al., 2007. [t] Hewitt et al., 2005. [u] Dunn et al., 1996. [v] Fletcher & Peters, 1984. [w] Gavin et al., 1993. [x] Oetting & McDonald, 2001. [y] Oetting et al., 2021. [z] Overton et al., 2021. [aa] Smyk, 2012. [ab] Ooi & Wong, 2012.

Table S2

Reference Measures Used in Reviewed Studies

Reference Measure	Study
Clinical criterion (diagnosis by an SLP/currently receiving treatment)	Charest et al., 2020 [b]
	Dunn et al., 1996
	Gladfelter & Leonard, 2013 [a]
	Fletcher & Peters, 1984 [b] Gavin et al., 1993 [b]
	Guo & Eisenberg, 2014 [b]
	Guo & Schneider, 2016 [b]
	Guo et al., 2019 [b]
	Guo et al., 2020 [b]
	Hoffman, 2009 [b]
	Klee et al., 2007 [b]
	Pavelko & Owens, 2019 [b]
	Heilmann et al., 2010
	Liles et al., 1995
	Moyle et al., 2011 [a]
	Oetting & McDonald, 2001
	Ooi & Wong, 2012
	Overton et al, 2021
	Schneider et al., 2006 [a]
	Souto et al., 2014 [a]
Clinical Evaluation of Language Fundamentals, 3rd Edition (CELF-3)	Charest et al., 2020
	Guo & Schneider, 2016
	Guo et al., 2019
	Guo et al., 2020
	Hoffman, 2009
	Schneider et al., 2006 [a]
Clinical Evaluation of Language Fundamentals, 4th Edition Spanish (CELF-4 Spanish)	Smyk, 2012

Clinical Evaluation of Language Fundamentals, Preschool 2nd Edition (CELF-P2)	Charest et al., 2020 Guo et al., 2019 Guo et al., 2020 Schneider et al., 2006
Diagnostic Evaluation of Language Variation - Norm Referenced (DELV)	Oetting et al., 2021
EpiSLI System (Test of Language Development, Primary + narrative task)	Hewitt et al., 2005
Preschool Language Scales, 3rd Edition (PLS-3)	Bedore & Leonard, 1998
Peabody Picture Vocabulary Test, Revised (PPVT-R)	Moyle et al., 2011 Bedore & Leonard, 1998 Gavin et al., 1993 Oetting & Mcdonald, 2001 Rudolph et al., 2019
Reynell Developmental Language Scales (Expressive)	Fletcher & Peters, 1984
Spanish-English Language Proficiency Scales (SELPS)	Smyk, 2012
Sequenced Inventory of Communication Development, Revised (SICD-R)	Gavin et al., 1993 Klee et al., 2007
Structured Photographic Expressive Language Test, 3rd Edition (SPELT-3)	Castilla-Earls & Fulcher-Rood, 2018 Smyk & Restrepo, 2012
Structured Photographic Expressive Language Test, Preschool 2nd Edition (SPELT-P2)	Eisenberg & Guo, 2013 [a] Gladfelter et al., 2013 Souto et al., 2014 Guo & Eisenberg, 2014
Stephens Oral Language Screening Test	Fletcher & Peters, 1984
Test for Auditory Comprehension of Language, Revised (TACL-R)	Moyle et al., 2011 Bedore & Leonard, 1998

Test for Examining Expressive Morphology (TEEM)	Pavelko & Owens, 2019
Test of Language Development, Primary (TOLD-P)	Bedore & Leonard, 1998
Test of Language Development, Primary 2nd Edition (TOLD-P2)	Oetting & Mcdonald, 2001
Grammaticality	Smyk, 2012
Mean Length of Utterance (MLU)	Rudolph et al., 2019 Oetting & Mcdonald, 2001
Non-Word Repetition (NWR)	Rudolph et al., 2019

Note. [a] Reference measure included clinical criterion in addition to other measures. [b] Clinical criterion was confirmed by standardized measures.

Supplemental Material S2

Clinical Guide to LSA Measures with Best Accuracy

Age Group	English Variety	Measure(s) / Model
3 yo	ME	LARSP Model (VP Errors + Stage 1 Utterances + Age + 3-element NP)
4 yo	ME	Finite Verb Morphology Composite (FVMC) Version 1-3 Developmental Sentence Scoring (DSS) Sentence Point
5 yo	ME SWE	Finite Verb Morphology Composite (FVMC) Version 1-3 Developmental Sentence Scoring (DSS) Sentence Point Past Tense: Strategic Scoring
6 yo	ME	Finite Verb Morphology Composite (FVMC) Version 3 SUGAR Model (MLU + Clauses per Sentence) Percent Grammatical C-units (PGCU) Errors per C-Unit Unmarked Verbs + Verb Types
7 yo	ME	Finite Verb Morphology Composite (FVMC) Version 3 SUGAR Model (MLU + Clauses per Sentence) Percent Grammatical Utterances/C-Units
8 yo	ME	Percent Grammatical Utterances (PGU) Errors per C-Unit Proportion “Restricted” Utterances
9 yo	ME	Percent Grammatical Utterances (PGU)
10 yo	ME	Proportion “Restricted” Utterances

Note. ME = Mainstream English. SWE = Southern White English.

LARSP Model

Source: Gavin, W. J., Klee, T., & Membrino, I. (1993). Differentiating specific language impairment from normal language development using grammatical analysis. *Clinical Linguistics & Phonetics*, *7*(3), 191–206.

Ages: 2;0-4;2
Accuracy: Good
Elicitation: Conversation/Play (20-minute sample interaction between child and caregiver playing with a set of toys)
Materials: Age-appropriate toys
Sample Length: average 198 utterances (61-377)
Transcription: modified SALT conventions
Coding:
-Total Major Utterances (exclude single word yes/no utterances and 'unanalyzed' or 'problematic' utterances)
- 3-element Noun Phrases (count number of occurrences, divide by total major utterances)
    Determiner + Adjective + Noun (e.g., the big train)
    Adj + Adjective + Noun (e.g., big red truck)
    Preposition + Determiner + Noun (e.g., in my pocket)
- Verb Phrase Errors (count number of occurrences, divide by total major utterances)
- Stage 1 Major Utterances (count number of occurrences, divide by total major utterances)
    'V' (Command) (.e.g, Stop!)
    'Q' (Question) (e.g., What?)
    'V' (Statement).
    'N' (Statement).
    Other (Statement)
- Input these values into the following formula:

-7.58 + .14(Age in months) + 5.87(Stage 1 Major Utterances) + 12.96(VP Errors) - 16.58(3-element NP)

Cutoff: <0.025 classified as typical language, >0.025 classified as impaired

Verb Morphology Composite / Finite Verb Morphology Composite

Version 1

Source: 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.

Age: 4;0-4;6, 5;0-5;6
Accuracy: Good
Elicitation: Play (interactions between child and experimenter)
Materials: Age-appropriate toys
Sample length: 152 utterances or more
Transcription Conventions: SALT

Coding:

-Identify obligatory contexts for the morphemes of interest: regular past tense inflections, regular third person singular present inflections, copula and auxiliary BE forms (i.e., am, is, are, was, were in contracted or uncontracted form), and auxiliary DO forms (i.e., do, does, did)

-Mark instances of correct and incorrect (omissions and substitutions) usage (NOTE: overregularized past tense forms (e.g., throwed instead of threw) should be scored as an additional obligatory context and credited with an additional instance of past tense –ed)

-Calculate percentage of correct usage: the number of correct productions in the composite divided by the total number of obligatory contexts and multiplied by 100

Cutoff: 4yo = 76%, 5yo = 82.5%

Version 2

Source(s): Souto, S. M., Leonard, L. B., & Deevy, P. (2014). Identifying risk for specific language impairment with narrow and global measures of grammar. *Clinical Linguistics & Phonetics*, *28*(10), 741–756.

Ages: 4;0-5;10
Accuracy: Good
Elicitation: Play (interactions between child and experimenter)
Materials: Age-appropriate toys
Sample length: first 50 utterances containing a subject plus verb (100+ elicited)
Transcription Conventions: SALT
Coding:

-NOTE: Code all utterances beginning with the first utterance in the sample to the point at which the 50th utterance containing a subject plus verb

-Identify obligatory contexts for the morphemes of interest: regular past tense inflections, regular third person singular present inflections, and copula and auxiliary BE forms (i.e., am, is, are)

-Mark instances of correct and incorrect (omissions and substitutions) usage

-Calculate percentage of correct usage: the number of correct productions in the composite divided by the total number of obligatory contexts and multiplied by 100

Cutoff: 4yo = 76.95%, 5yo = 83.73%

Version 3

Source(s): Guo, L. Y., & Schneider, P. (2016). Differentiating school-aged children with and without language impairment using tense and grammaticality measures from a narrative task. *Journal of Speech, Language, and Hearing Research*, *59*(2), 317–329.

Source: Guo, L. Y., Eisenberg, S., Schneider, P., & Spencer, L. (2020). Finite verb morphology composite between age 4 and age 9 for the Edmonton Narrative Norms Instrument: Reference data and psychometric properties. *Language, Speech, and Hearing Services in Schools*, *51*(1), 128-143.

Ages: 4-9yrs

Accuracy: Adequate (6-7yrs) to Good (4-5yrs)
Elicitation: ENNI story generation task
Materials: ENNI picture sequences[1]
Sample length: average 58-81utterances (33-181)
Transcription Conventions: SALT
Coding:

-NOTE: Exclude C-units that contained verb forms but no subjects (e.g., Getting the airplane out of the swimming pool)

-Identify obligatory contexts for the morphemes of interest: regular past tense inflections, regular third person singular present inflections, and contracted and uncontracted copula and auxiliary BE forms (i.e., am, is, are, was, were). NOTE: Do not include the infinitive form of be (e.g., The rabbit will be sick), present participle form of be (e.g., The rabbit is being funny), past participle form of be (e.g., He has been trying to get the ball), or gerund form of be (e.g., Being happy is easy) in this calculation.

-Mark instances of correct and incorrect (omissions and substitutions) usage (NOTE: excluded overgeneralization of 3SG –s (e.g., The elephant haves an airplane) or regular past tense –ed (e.g., The elephant just standed there).

-Calculate percentage of correct usage: the number of correct productions in the composite divided by the total number of obligatory contexts and multiplied by 100

Cutoffs: 4yo = 83.77%, 5yo = 93.46%, 6yo = 93.50%, 7yo = 96.64%

Developmental Sentence Scoring (DSS) Sentence Point

Source(s): Souto, S. M., Leonard, L. B., & Deevy, P. (2014). Identifying risk for specific language impairment with narrow and global measures of grammar. *Clinical Linguistics & Phonetics*, *28*(10), 741–756.

Ages: 4:0-5:10
Accuracy: Good
Elicitation: Play (interactions between child and experimenter)
Materials: Age-appropriate toys
Sample length: first 50 utterances containing a subject plus verb (100+ elicited)
Transcription Conventions: SALT
Coding:

-NOTE: Code all utterances beginning with the first utterance in the sample to the point at which the 50th utterance containing a subject plus verb

- Score each utterance: give one sentence point if and only if the sentence was fully grammatical, regardless of whether it uses simple or complex morphosyntax, and give zero points for any grammatical error (e.g., a sentence point should be withheld for sentences such as “Her broke the window” (personal pronoun error), “Dad built new birdhouse and Mom ate two apple” (grammatical errors on articles and noun plural inflections)

-Calculate Sentence Point Score: add the total number of sentence points earned and divide by 50 (i.e., total number of utterances)

Cutoffs: 4yo = .755, 5yo = .815

SUGAR Model

Source: Pavelko, S. L., & Owens, R. E. (2019). Diagnostic accuracy of the sampling utterances and grammatical analysis revised (SUGAR) measures for identifying children with language impairment. *Language, Speech, and Hearing Services in Schools*, *50*(2), 211–223

Ages: 3;0-7;0
Accuracy: Adequate
Elicitation: SUGAR Conversation protocol
Sample length: 50 utterances
Transcription Conventions: SUGAR
Coding:

- Mean Length of Utterance (SUGAR): the total number of morphemes divided by 50. Per the rules in Pavelko and Owens (2017), count all free morphemes, five grammatical morphemes, 18 derivational morphemes, and each word in a proper name as one morpheme; all contractions and the words *hafta*, *wanna*, and *gotta* as two morphemes; and the word *gonna* as three morphemes.
- Clauses Per Sentence (CPS): the total number of clauses divided by the number of sentences

Cutoffs: Both measures below the cutoff indicates impairment

Measure	**3:0-3:5**	**3;6-3;11**	**4;0-4;5**	**4;6-4;11**	**5;0-5;11**	**6;0-6;11**	**7;0-7;11**
MLU	2.87	4.13	4.26	4.86	5.31	6.00	6.87
CPS	0.90	0.99	1.00	1.05	1.10	1.15	1.18

Percent Grammatical Utterances/C-Units

Source(s): Guo, L. Y., & Schneider, P. (2016). Differentiating school-aged children with and without language impairment using tense and grammaticality measures from a narrative task. *Journal of Speech, Language, and Hearing Research*, *59*(2), 317–329.

Guo, L. Y., Eisenberg, S., Schneider, P., & Spencer, L. (2019). Percent grammatical utterances between 4 and 9 years of age for the Edmonton Narrative Norms Instrument: Reference data and psychometric properties. *American Journal of Speech-Language Pathology*, *28*(4), 1448-1462.

Ages: 4-9yrs
Accuracy: Acceptable (4-8) to Good (9yrs)
Elicitation: ENNI story generation task
Sample length: average 58-81utterances (33-181)
Transcription Conventions: SALT
Coding:

- Identify errors: errors in tense marking, incorrect pronoun use, omission or incorrect use of grammatical morphemes, inconsistent argument structure (i.e., omission of a required constituent, other syntactic errors that were not included in the previous categories (e.g., semantic irregularities).
- Percent grammatical utterances/C-units (PGU/PGCU): 1) calculate the total number of utterances/C-units containing at least 1 error, then subtract from the total number of utterances/C-units. Divide by the total number of C-units.

Cutoff: 4 yrs = 54.04%, 5 yrs = 79.10%, 6 = 83.00%, 7 yrs = 85.40%, 8 yrs = 91.50%, 9 yrs = 88.42%

Errors per C-unit

Source: Guo, L. Y., & Schneider, P. (2016). Differentiating school-aged children with and without language impairment using tense and grammaticality measures from a narrative task. *Journal of Speech, Language, and Hearing Research*, *59*(2), 317–329.

Ages: 6yrs and 8yrs
Accuracy: Acceptable
Elicitation: ENNI story generation task[3]
Materials: ENNI picture sequences[3]
Sample length: average 58-81utterances (33-181)
Transcription Conventions: SALT
Coding:

- Identify errors: errors in tense marking, incorrect pronoun use, omission or incorrect use of grammatical morphemes, inconsistent argument structure (i.e., omission of a required constituent, other syntactic errors that were not included in the previous categories (e.g., semantic irregularities).
- Number of errors per C-unit (Errors/CU): total number of errors divided by total number of C-units that were included for analysis

Cutoff: 6yo = .14, 8yo = .09

Proportion “Restricted” Utterances

Source: Hoffman, L. M. (2009). The utility of school-age narrative microstructure indices: INMIS and the proportion of restricted utterances. Language, Speech, and Hearing Services in Schools, 40(4), 365-375.

Ages: 8-10 yrs
Accuracy: Acceptable
Elicitation: Narrative generation
Materials: *Frog Where Are You?* by Mercer Mayer
Sample length: average 38 utterances (22-72)
Transcription Conventions: SALT
Coding:

- Segment utterances into T-units

- Code restricted utterances: mark T-units as "restricted" if they have 1) a complete clause with a subject and predicate, and 2) contain any number of grammatical errors (including verb inflections or clausal structure) and/or semantic errors (i.e., inaccurate references or meanings, such as pronoun reversals or substituting "door" for *window*)
- Proportion "restricted" utterances: total number of utterances marked as "restricted" divided by total number of complete & intelligible utterances

Cutoff: 14% or higher indicates impairment

Unmarked Verbs + Verb Types

Source: Fletcher, P., & Peters, J. (1984). Characterizing language impairment in children. Language Testing, 1(1), 33–49.

Ages: 3;4-6;11
Accuracy: Acceptable
Elicitation: 1 hour session of 4 activities: free play with a familiar adult, narrative generation, board game play, narrative retell
Materials: Toys, board game, wordless picture book (generation), picture sequence (retell)
Sample length: 200 or more (50 per task)
Transcription Conventions: N/A
Coding:

- Code unmarked verbs: total number of verbs that are not marked by an auxiliary or inflection
- Code verb types: total number of unique verbs
- Input these values into the following formula:

$$-0.87710 + -0.2770(\textit{Unmarked Verbs}) + 0.10354(\textit{Verb Types})$$

Cutoff: .19 or below indicates impairment

Past Tense (Strategic Scoring)

Source: Oetting, J. B., Rivière, A. M., Berry, J. R., Gregory, K. D., Villa, T. M., & McDonald, J. (2021). Marking of tense and agreement in language samples by children with and without specific language impairment in African American English and Southern White English: Evaluation of scoring approaches and cut scores across structures. Journal of Speech, Language, and Hearing Research, 64(2), 491–509.

Ages: 5 yrs
Accuracy: Acceptable
Elicitation: Play (20-30 minutes), narrative generation
Materials: Toys (gas station set, picnic/park set, baby doll set), 3 action pictures (a child at a doctor's office getting a shot and a family fishing, grocery shopping, or washing a car)

Sample length: average 237 utterances
Transcription Conventions: SALT, except for utterance segmentation rules
Coding:

- Segment utterances into C-units, but allow two conjoined independent clauses to remain in the same utterance
- Code past tense on main verbs only (not participles, auxiliaries, or non-changing forms such as *cut*)
- Code mainstream overt (MO): past tense marked with forms that are consistent with standard English (e.g., *jumped*, *ate*)
- Code nonmainstream overt (NMO): past tense marked with dialect-specific patterns (e.g., *drunk*)
- Code nonmainstream zero form (NMZ): no acoustically perceptible marking
- Code other forms (O): more than one tense/agreement form marked within a predicate (e.g., *where did this went?*)
- Calculate percentage of overt marking: dividing the total mainstream and nonmainstream overt forms divided by total overt and zero forms (MO+NMO/MO+NMO+NMZ). Do not include forms coded as "other."

Cutoff: 91-93% or lower indicates impairment