DYNAMIC ASSESSMENT OF SPANISH-ENGLISH SPEAKING CHILDREN FOR DEVELOPMENTAL LANGUAGE DISORDER

by

Sarah Lynn Neiling

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF SPEECH, LANGUAGE, & HEARING SCIENCES

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2024

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by: Sarah Lynn Neiling

titled: Dynamic Assessment of Spanish-English Speaking Children for Developmental Language Disorder

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Mary Alt — Date: Jul 4, 2024

Mary Alt

Elena Plante

Elena Plante (Jul 8, 2024 08:32 PDT) — Date: Jul 8, 2024

Elena Plante

Leah L Kapa — Date: Jul 5, 2024

Leah Kapa

Elizabeth D Peña

Elizabeth D Peña (Jul 4, 2024 16:00 PDT) — Date: Jul 4, 2024

Elizabeth D Peña

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Mary Alt — Date: Jul 4, 2024

Mary Alt

Speech, Language, and Hearing Sciences

ACKNOWLEDGEMENTS

I express my sincerest gratitude to my mentor, Dr. Alt, and to my committee for their support in this project. I am deeply indebted to Dr. Alt for her unmatched optimism, kindness, wisdom, perseverance, patience, generosity, and cheerleading throughout my PhD journey. Through her belief in me, I began to believe in myself, and for that I could not be more grateful.

The invaluable wisdom from my other mentors has taught me how to be an academic. I am in awe of Dr. Plante's depth of knowledge and hope to one day be even a fraction of the force in the field that she is. Dr. Kapa's grounded guidance has kept me level-headed and clear-eyed. I am thankful for Dr. Peña's support on my dissertation as the titan who brought dynamic assessment into bilingual language assessment research. I am also deeply grateful to Dr. Arizmendi who has made me believe an academic career is possible.

I am grateful to the generous financial support from the University of Arizona Graduate & Professional Student Council Research and Projects Grant and Galileo Circle Scholarships. My doctoral training funding has made it possible for me to engage in and disseminate this work by funding my training and travel, so thank you to the Preparing Researchers In Early Intervention for Children with Disabilities from Multicultural Environments (PRIDE) training grant, from the U.S. Department of Education, Office of Special Education (OSERS-SOEP: H325-D190062).

Without the families who took the leap to participate and my dedicated research assistants, none of this work would have been possible, and I thank them for their commitment to this work.

Finally, I would be remiss not to thank my PhD cohort, family, and friends. They fed my mind, my spirits, and my body to sustain me through the PhD program.

LAND ACKNOWLEDGEMENT

We respectfully acknowledge the University of Arizona is on the land and territories of Indigenous peoples. Today, Arizona is home to 22 federally recognized tribes, with Tucson being home to the O'odham and the Yaqui. Committed to diversity and inclusion, the University strives to build sustainable relationships with sovereign Native Nations and Indigenous communities through education offerings, partnerships, and community service.

DEDICATION

To David, my rock, and to Mac, my world.

Table of Contents

List of Figures

List of Tables

List of Appendices

Abstract

Background and Objectives: Assessment is the gateway to speech-language pathology (SLP) services. For Spanish-English bilingual children, there are few assessments with high classification accuracy, and those that exist may have barriers to use for monolingual English-speaking clinicians. One under-explored route is dynamic assessment, any assessment into which teaching is incorporated with a focus is on the child's learning potential. The current project was a prospective feasibility study that updated a previously published dynamic assessment tool by Roseberry and Connell (1991) for the identification of Developmental Language Disorder (DLD), a lifelong difficulty with oral and written language whose origin is not attributable to another disorder or brain injury, in Spanish-English children, ages 4;0 – 5;11. The aims of this diagnostic accuracy study were to (1) create a replicable dynamic assessment, and (2) to understand which elements of scoring yielded the highest classification accuracy.

Methods: Children were eligible if they were Spanish-English bilingual, ages 4;0 – 5;11, had no history of neurologic or other impairments that would exclude DLD. Data was collected from a convenience sample from October 2023 – April 2024. For the dynamic assessment, children were taught a novel morpheme (/-əb/, partial object), as morphology is a hallmark deficit in DLD, through a scripted storybook that incorporated statistical learning principles. Following teaching, the child's modifiability was rated, and they were administered an immediate, one-hour, and one-day posttest. Children's language status was determined by the Bilingual English-Spanish Oral Screener (BESOS; Peña et al., 2008, in development).

Results: The assessment had high fidelity and scoring had high interrater reliability, demonstrating its replicability. To determine diagnostic accuracy, Receiver Operator Characteristic (ROC) curves with Bayesian logistic regression and Random Forest inputs determined the Area Under the Curve (AUC), which was used to compare scoring elements across 12 typically developing (TD) and 6 DLD children. Combining the modifiability rating scales and teaching phase score yielded the highest AUC (.83) with good sensitivity (83%) and specificity (100%).

Discussion: This was an improvement on the original dynamic assessment by Roseberry and Connell (1991) and replications of it. Conclusions are not broadly generalizable due to small sample size and because there was low interrater reliability of the modifiability scales. Future investigations should focus on increasing reliability and investigating other parameters of the dynamic assessment to optimize utility and classification accuracy. After future study, this dynamic assessment may be used to identify DLD in Spanish-English bilingual preschoolers for entry to SLP services.

Registration: Not registered.

This abstract was formatted in adherence to the STARD for abstracts guidelines (STARD #1 -2; Cohen et al., 2017).

Introduction

The current study examines the classification accuracy of a dynamic assessment for identifying Developmental Language Disorder (DLD) in Spanish-English bilingual children. DLD entails a marked difficulty acquiring language for all languages spoken in oral and written language) across a person's lifespan. The current study was motivated by the need for Speech-Language Pathologists (SLPs) to have accessible assessment materials with high classification accuracy, meaning they correctly identify children with DLD or with typically developing language (TD) with at least 80% and ideally 90% accuracy (Plante & Vance, 1994). While most language assessments focus on testing a child's acquired linguistic knowledge, dynamic assessments measure the child's learning potential as the indicator of language-learning difficulty. The current dynamic assessment evaluates the learning potential for the invented morpheme, /-əb/ in English, which refers to a partial object. In the introduction, I discuss bilingual language development in TD and DLD populations as well as variables to consider in assessment. Next, I provide a history of dynamic assessment and relate it to the current study. Finally, the theoretical framework and current study are explained. The Standards for Reporting of Diagnostic Accuracy Studies (STARD; Cohen et al., 2016) were followed throughout this document, indicated by (STARD #).

Bilingual Language Development

Children who are exposed to two languages regularly are engaged in bilingual language development. Past research has questioned whether bilinguals' languages would develop at a delayed rate compared to monolinguals given they have two sets of linguistic input to organize (Volterra & Taeschner, 1978). A bilingual child's language is expected to develop at a similar rate to a monolingual child's language with some skills developing at a slightly accelerated pace and other skills at a slightly decelerated pace, but on average they are on par (Hammer et al., 2014; Hoff et al., 2012; Paradis & Genesee, 1996 but see Castilla-Earls et al., 2023). Crucially, though, bilingual language development is not qualitatively the same as a monolingual child's language development, and it is not the same as two monolinguals in one brain (Grosjean, 1989). Bilinguals' languages interact (Hammer et al., 2014; Paradis & Navarro, 2003), providing for an experience unique to them, so monolingual language development is not an appropriate standard of comparison for bilingual language development. Although research traditionally has compared bilinguals to monolinguals, in this manuscript, I will compare bilinguals to themselves. In this subsection, I will detail the variables critical to consider within bilingual language development *writ large* before summarizing typical Spanish-English bilingual morphosyntactic development.

Characterizing Bilingualism

Bilingual language development has a trajectory that depends on multiple variables regarding the child's bilingual experience, often called their bilingual profile. The age of first exposure (Hammer et al., 2014; Kremin & Byers-Heinlein, 2021; Luk & Bialystok, 2013; Marian & Hayakawa, 2021), the amount of exposure and use of each language (Bedore et al., 2012; Hammer et al., 2014; Hoff et al., 2012; Peña et al., 2023), language proficiency and dominance (Peña et al., 2023), and the linguistic environment in the community (Escobar et al., 2022) are all aspects of the bilingual profile that influence how a child's two languages will develop. For instance, a child who hears primarily Spanish at home and then begins English at school will have stronger Spanish dominance at school onset. If the child were to also be exposed to Spanish in school as they began to learn English, they could use their linguistic knowledge of Spanish to boost their English language learning and develop proficiency in both languages. Other variables may also be included in a bilingual profile, such as manner of acquisition, immigration status/acculturation, and the language (and dialect) combinations and their typological distance (Escobar et al., 2022; Grosjean, 2014; Laketa et al., 2021; Lee, 2022; Luk & Bialystok, 2013; Marian & Hayakawa, 2021). Though these other variables may be important, I chose to focus on age of first exposure, language use, language proficiency and dominance, and community context because they were relevant to the age range in this study (4;0 – 5;11), have sufficient evidence to support their use, and are easily measured. Easy measurement was a criterion for selection because it meant that the values could be captured with high accuracy relative to other potential variables that did not have established instruments or measurement guidelines. Additionally, these variables may be measured by clinicians, enhancing the clinical utility of this work.

Age of onset, or age of first exposure, is the age at which a child first hears each language regularly. Traditionally, the question of age of onset has been centered around whether the child was exposed to both languages regularly before or after age three, resulting in the categorization of simultaneous or sequential bilinguals, respectively. These categories, though, may not be as useful as listing the year at which a child first hears each language because the categorical variable falsely assumes equivalent experiences for all children who heard two languages before age three or after. Age of onset is important to a child's bilingual profile because it helps gauge a child's relative practice with a language. A child who heard Spanish and English from birth would have more practice than a child who heard Spanish from birth and English from age four. At face value, this might seem predictive of a child's relative language proficiency in each language (i.e., language dominance). While age of first exposure may account for up to 35% of the variance in predicting language dominance, current language input and output have been more predictive (explaining 60% of the variance) of language proficiency and dominance in pre-kindergarteners and kindergarteners (Bedore et al., 2012).

Language use is the combination of input and output and may be measured by hour-by-hour reports of which language the child hears and uses throughout their typical week. Language use is predictive of a child's developmental pace for each of their languages (Hammer et al., 2014; Hoff et al., 2012). If a child uses one language substantially more than the other ("unbalanced use"), they may be more likely to develop that language more quickly than the other. In contrast, if a child uses both languages to a similar extent ("balanced use"), they may be more likely to have parallel development in their languages. Relatedly, language *use* has been shown to predict language proficiency and dominance across children's languages (Bedore et al., 2012; Peña et al., 2023).

Language proficiency is a child's linguistic ability in each language compared to a monolingual's performance in that respective language (Bedore et al., 2012). Language dominance is the relative proficiency across linguistic skills in each language (Bedore et al., 2012; Peña et al., 2023). A child may have congruent dominance, performing best in the same language across different types of language measures (e.g., semantic *and* morphosyntactic measures), or a child may exhibit mixed dominance, performing best in one language on a semantic measure and in the other language on a morphosyntactic measure. In a compilation of data from multiple studies of children aged 5 – 12 years (n = 495 TD, n = 100 DLD), Peña et al. (2023) evaluated the interaction between language ability (TD and DLD) and its impact on the likelihood of being characterized as exhibiting mixed dominance at varying levels of English exposure (0% – 100%). They found that language use influences language dominance on a continuum and by language ability. As English use increased, the likelihood of mixed dominance increased for TD children (Peña et al., 2023). Children with DLD displayed the opposite relationship, decreasing their likelihood of mixed dominance as English language use increased, and they demonstrated greater variability in their likelihood of mixed dominance across the spectrum of English exposure (Peña et al., 2023). The authors speculated the greater variability for children with DLD in their likelihood of mixed dominance was because children with DLD take longer to learn language. It is also important to separate language dominance, which is relative proficiency, from language ability, which refers to linguistic skill overall when considering both languages. As an example, a Spanish-dominant child with DLD might have similar Spanish language proficiency to an English-dominant child with TD. However, the child with TD will have greater language ability than the child with DLD because DLD, by definition, means that a child has low language ability across both languages whereas the TD child's English proficiency would be in the typical language ability range. I will use the term *language ability* when referring to the diagnostic categories of TD versus DLD rather than as an aspect of the child's bilingual profile.

The age of first exposure, current language use, language proficiency, and language dominance are all child-level variables, but the community context also plays an important role in shaping a child's bilingual development. Policies, cultural attitudes toward bilingualism, and demographics all comprise the community context, which spans the local, state, and federal levels. For the older children in this study, one of the most important community contextual factors is the Arizona state policy (Proposition 203) that only English may be the language of instruction in schools (Arizona Revised Statutes §15-752). This policy creates a subtractive bilingual environment, meaning children are learning English in public school but are not receiving any support for their home language of Spanish. Subtractive bilingual language environments put the children at risk of language loss or attrition, the gradual reduction in abilities or halt in development of the child's home language, respectively (Anderson, 2022; Fillmore, 1991; Hiebert & Rojas, 2021).

Other factors may counteract (but not fully negate the effects of) the monolingual school language policies in Arizona. The demographic make-up of the community itself in part determines which languages are frequently spoken, which influences opportunity for use. In Tucson, 44.8% of residents are Latine (United States Census Bureau, n.d.). Because about 75% of Latines in the US speak Spanish at least conversationally (Mora & Lopez, 2023), it can be roughly estimated that about 33.6% of Tucson residents speak Spanish. For comparison, in the rest of the United States only 18.7% of people are Hispanic or Latino (Jensen et al., 2021). Additionally, as the population of Spanish-speakers in the US continues to proliferate (Krogstad et al., 2023), the demand for Spanish use in all environments grows, increasing its relative prestige at the national level. The frequency and contexts of each language's use also reflect to the child the relative prominence or value of their languages. Hearing Spanish in stores, in media, and in the neighborhood can communicate a value of importance for the Spanish language. Similarly, cultural activities, events, and symbols that reflect the child's culture(s) provide opportunities for the child to link their linguistic and cultural identities, which could reinforce their language use.

A child's bilingual development is influenced by a complex interplay of community factors, but, in the literature, most attention has been given to how child-level variables impact bilingual language development. This is not surprising given the challenge of quantifying the influence of the many community factors and determining their causal relationship to bilingual language growth. It is important to keep the bilingual profile in mind when considering bilingual language development and the large variation that it may foster. In the next section, I will focus on morphological development in Spanish-English bilinguals as the grammatical skill of interest in the current study is utilizing a derivational morpheme.

Typical Development of Spanish-English Morphology

The pace of Spanish-English morphological development is influenced by the child's bilingual profile (e.g., language use, language in school) and the typology of each language. Spanish is a morphologically rich language, meaning it makes heavy use of verb conjugations, gender agreement, number agreement, and other morphological markers. In contrast, English is morphologically sparse, employing bound morphological markers infrequently, especially with verbs. Nor does English require gender agreement or have clitic pronouns in the way Spanish does. Frequency of grammatical constructions and their phonetic salience (Baron et al., 2018; Cooperson et al., 2013; Dunn Davison & Scheffner Hammer, 2012) are correlated with the developmental trajectory of Spanish and English morphological acquisition. For example, Spanish-English bilingual children acquire the imperfect and preterit verb tenses before subjunctive and direct object clitics, which are less frequent in the input and less phonetically salient, respectively (Baron et al., 2018; Cooperson et al., 2013). Similarly, phonetically salient forms in English (e.g., irregular past-tense verbs) may lead to quicker acquisition of those particular grammatical forms than those with less salience (e.g., regular past tense –ed) for Spanish-English bilinguals (Dunn Davison & Scheffner Hammer, 2012).

The specific combination of the two languages being learned will also influence the morphosyntactic developmental trajectory. Some skills are language-specific, such as gender agreement or clitic pronouns, which are only acquired in Spanish and are absent in English. Other skills have a similar conceptual basis but differing syntactic structures, leading to crosslinguistic competition that makes those structures likely to be acquired later than structures with morphosyntactic concordance across languages. Similar syntactic structures, though, may facilitate acquisition. For instance, the possessive form in English is a final /-s/ (e.g., “mom’s shoe”) while in Spanish, it is a prepositional phrase “OBJECT de PERSON” (e.g., “el zapato de mamá/the shoe of mom”). The possessive form tends to be acquired later in Spanish-English bilinguals than the plural noun marker, /-s/, which is the same in both languages (e.g., “cats/gatos”) and tends to be one of the earlier acquired forms (Baron et al., 2018; Dunn Davison & Scheffner Hammer, 2012). For the grammatical skills in Spanish that are most prone to developmental challenges (i.e., language attrition or DLD), TD Spanish-English bilingual children in the U.S. have acquired articles with 67 - 69% accuracy, verbs with 65 - 67% accuracy, clitics with 39 - 64% accuracy, and subjunctive verbs with 25 - 55% accuracy on average by five years, six months (Castilla-Earls et al., 2021; Castilla-Earls et al., 2023).

The focus of the current study is assessing how well Spanish-English speaking children learn a derivational morpheme to help identify whose language is TD and who has DLD. A derivational morpheme is typically a bound morpheme that carries a semantic meaning and that can change the meaning of the base word. While the morphemes may be language-specific, the linguistic process of deriving is common to both English and Spanish. For example, the derivational morpheme “re-” means to do again, so it adds meaning to the word “heat/calentar” when combined. “Reheat/recalentar” means to heat something again. Derivational morphemes can also change a word’s class. For instance, the derivational morpheme “-er” means an agent who does a particular thing. When “-er” binds to a verb (e.g., “paint”), the base word changes class to become a noun (e.g., “painter”). Another class of morphemes, and the class that is typically of the most focus in speech-language pathology, is inflectional morphemes, which are free or bound morphemes that add grammatical information to a word or syntactic structure (e.g., "is + -ing" to signal the ongoing action that an agent "is walking"). There are a limited number of inflection morphemes within each language. Inflectional morphemes tend to be of most interest to SLPs because they are the most critical to early grammatical development and may be markers of DLD. Derivational morphemes, by contrast, can be avoided (e.g., use "she" instead of "the painter") and so are less vital to grammaticality. Both types of morphemes are likely learned through a similar process of pattern recognition, storage, and integration (i.e., statistical learning; Erickson & Thiessen, 2015) given that they tend to follow linguistic "rules" (e.g., re+VERB to indicate to VERB again; NOUN+s to indicate more than one NOUN). In the following section, I will introduce how morphology may present in Spanish-English bilingual children with DLD before considering a driving mechanism behind morphological development—statistical learning.

DLD in Spanish-English Bilinguals

DLD is s a lifelong condition of unknown cause in which a person has difficulty understanding and producing both spoken and written language (Bishop et al., 2017). It is more prevalent (approximately 7% of the population) but less researched than autism (McGregor, 2020; Tomblin et al., 1997). Functionally, children with DLD may experience more difficulties in literacy, academic skills, forming social relationships, maintaining mental health, and retaining employment (Law et al., 2009; Nippold et al., 2008; Norbury et al., 2016; Whitehouse, Line, et al., 2009; Whitehouse, Watt, et al., 2009), and they may be more likely to have more encounters with the justice system (Snow & Powell, 2004). DLD is defined by difficulties with semantics, morphology, syntax, and/or phonology in the absence of intellectual disability or other biological etiology (e.g., autism, acquired brain injury, hearing loss; Bishop et al., 2017).

Although children with DLD may have impairment across a number of linguistic domains, the hallmark deficit is difficulty with morphosyntax (Bishop et al., 2017). Overall children with DLD exhibit more protracted language development than their TD peers in these skills, and there is large individual variation across both TD and DLD groups (Castilla-Earls et al., 2021; Castilla-Earls et al., 2023). Bilingual children with DLD exhibit linguistic difficulties in both languages, but the error patterns are language specific.

Depending on their languages, bilingual children with DLD will exhibit relative difficulties in particular grammatical skills that vary by language. In Spanish, Spanish-English bilingual children with DLD have more difficulty with noun morphology, such as articles and clitics, verbs, and the subjunctive than their TD peers (Castilla-Earls et al., 2021). In English, Spanish-English bilingual children with DLD have marked difficulty with verb morphology, including third person singular present tense, regular past-tense -ed, the copula BE, and the auxiliary DO (Gutierrez-Clellen et al., 2008). Notably, these are all inflectional morphemes, but not all inflectional morphemes are equally likely to be compromised. The most difficult morphemes to acquire tend to be those of lower frequency, lower phonetic salience, and those whose low phonetic salience co-occurs with abstract meaning (i.e,. "mapping theory"; Chiat, 2001).

Mapping theory is the proposal that the core difficulty with morphological forms in DLD stems from how a word or grammatical form's phonological salience maps onto its lexical characteristics (Chiat, 2001). Forms with higher phonological salience and stronger perceptual cues (e.g., concrete, imageable referent) are predicted to be less impaired than forms with lower phonological salience and weaker perceptual cues (e.g., past tense -ed). While inflectional morphology is typically explored when looking at the grammatical deficits in DLD, difficulty with derivational morphology could be predicted. Children with DLD would be predicted to have the most difficulty learning derivational morphemes of low phonetic salience and with relatively abstract meaning. This is one of many theories attempting to account for the difficulties in DLD.

An alternative explanation is that children with DLD are delayed at grammatical learning (Castilla-Earls et al., 2021; Castilla-Earls et al., 2023). The grammatical targets that Chiat (2001) predicted would be the most difficult for children with DLD are also structures that are relatively later acquired for TD children and are susceptible for bilingual TD children experiencing language loss (e.g., Castilla-Earls et al., 2019). In other words, mapping theory may explain which grammatical forms are acquired in relative order for all children, and children with DLD are simply slower at that acquisition. The prediction of *what* might be delayed is useful, but a full theory of DLD needs an explanation of *how* the delay is instantiated. Mapping theory is not incompatible with the findings that children with DLD perform poorly in statistical learning, which may be the underlying mechanism driving these difficulties.

Statistical Learning

There is mounting evidence that many of the core aspects of language are acquired through implicit, statistical learning (Erickson & Thiessen, 2015; Siegelman, 2019). Statistical learning "is defined as the mechanism by which cognitive systems extract regularities from such sensory stimulation, to discover its underlying structure" (Siegelman, 2019, p. 2). Statistics are taken on the co-occurrence or distribution of elements, which are then held in working memory, chunked into units, and then converted into long-term memory where they can be integrated with other chunks to form a template or integrated into existing templates (Erickson & Thiessen, 2015). Through this computational process, children acquire many linguistic patterns, such as phonological patterns (e.g., Maye et al., 2002; Thiessen & Saffran, 2003), morphosyntactic structures (e.g., Gómez, 2002; Thompson & Newport, 2007), and new words (e.g., Benitez et al., 2016; Crespo et al., 2023; Smith & Yu, 2008). The focus of this subsection will involve the learning of derivational morphemes. To my knowledge, there have not been statistical learning studies on derivational morphemes in oral language (though there are several on spelling and reading), so theory about how they are learned statistically must be extrapolated from the grammar and semantic statistical learning literature.

There are multiple types of statistics that can be computed during statistical learning: frequency (the number of occurrences of items), co-occurrence (the probability of X occurring given Y, but without a need for a particular order), and sequences (the probability of X occurring before/after Y, in adjacency or non-adjacency). Additionally, distributional statistical learning facilitates the acquisition of word classes by detecting regularities in the positions of words and morphemes in sentences (Braine, 1987; Hall et al., 2018; Reeder et al., 2013). It is likely that learners are attuned to multiple probabilities in the input to facilitate learning and integrate prior experience in real-world language learning (Siegelman, 2019). For derivational morphemes, it is likely that the patterned of WORD+MORPHEME or MORPHEME+WORD is learned by sequential statistical learning of adjacent information rather than by engaging in distributional statistical learning.

Statistical learning of adjacent sequences has been studied in the acquisition of word boundaries in artificial grammar learning studies (Aslin et al., 1998; Saffran et al., 1996). Young infants have been shown to use transitional probabilities within and between words to determine the location of a word boundary by listening to a string of nonsense syllables whose probability has been controlled by the researchers. The infants learn which sound sequences have the highest probability of adjacent, sequential co-occurrence (i.e., transitional probability) and which sound sequences have the lowest transitional probability (Aslin et al., 1998; Saffran et al., 1996). Sound sequences with the highest transitional probability are within word sounds (e.g., “ha-ppy”) while those with low transitional probability occur at word boundaries (e.g., “ppy#birth” in “happy birthday”). This is relevant to the current study because word boundary detection is one level of statistical learning that may happen when hearing a new derivational morpheme. However, beyond infancy, children have learned word boundaries for most words, and so it is likely that statistical learning cues at the phrasal level may play a stronger role in learning a new morpheme.

As a computational mechanism, statistical learning is not limited to computing statistics at the phonological level; evidence shows that it works at the phrasal level of input, as well. Thompson and Newport (2007) had adults listen to an artificial language that contained variable phrase structures and had control participants listen to an artificial language without variable phrase structure (i.e., transitional probability was always 100%) to determine if adults used transitional probabilities to learn phrase boundaries. No semantic information was provided. Adults in the experimental condition with the most complex (i.e., most variable) artificial language demonstrated the most learning of all conditions. Thus, syntactic phrases are likely learned by the same mechanism as word boundaries—high transitional probabilities indicate items within a phrase while low transitional probabilities indicate a phrase boundary. Children have also been shown to learn visually based artificial grammars by engaging in statistical learning at the phrase level (Crespo & Kaushanskaya, 2022; Gillis et al., 2022; Pavlidou et al., 2009; Witt & Vinter, 2012).

The current study differs from artificial grammar learning paradigms in that children will receive semantic information about the target derivational morpheme from the language input and from a visual referent. Cross-situational word-learning paradigms have been used in basic statistical learning research to examine how learners resolve ambiguity in the environment to learn words (Smith et al., 2011; Yu & Smith, 2007). With enough regularity of the target words and enough doses (i.e., presentations of the target), both adults and children map phonological and semantic information onto one lexical representation (Smith et al., 2011; Smith & Yu, 2008; Yu & Smith, 2007). Although cross-situational word-learning studies are focused on words, this literature may apply to derivational morpheme learning because derivational morphemes carry similar properties to words in that they pair a sound sequence with meaning. To learn a derivational morpheme via statistical learning, a child will need to hear it consistently paired with its visual referent or with semantic content conferred through linguistic input.

Bilinguals and Statistical Learning

The previously examined research has primarily included monolingual, English-speaking participants, but the current study is designed for Spanish-English bilingual children. How does statistical learning apply to bilingual morphological and semantic learning? Bilingual children take statistics from two different streams of linguistic input and must organize this information to make templates for the phonology, semantics, morphology, and syntax of their respective languages. If statistical learning is an innate computational mechanism that all humans are born with, then it stands to reason that bilinguals will utilize statistical learning mechanisms to learn language, just as been shown in monolinguals. Indeed, this has been borne out in the literature. Bilingual infants can parse word boundaries through statistical learning mechanisms in multiple languages (Antovich & Graf Estes, 2018). Bilingual children and adults can also learn grammars and words through statistical learning paradigms (Ahufinger et al., 2021; Antovich & Graf Estes, 2018; Bartolotti et al., 2011; Crespo & Kaushanskaya, 2021, 2022; Crespo et al., 2023; Escudero et al., 2016; Kovács & Mehler, 2009; Onnis et al., 2017; Poepsel & Weiss, 2016). There is some evidence that bilinguals may be better able to manage certain input variability than monolinguals (e.g., visual and speaker variability, 2:1 word:picture mappings; Antovich & Graf Estes, 2018; Crespo et al., 2023; Escudero et al., 2016; Poepsel & Weiss, 2016), but others have failed to show this effect (Benitez et al., 2016; Crespo & Kaushanskaya, 2021). Regardless, the current study is not predicated upon a monolingual—bilingual comparison but rather is centered on a TD—DLD comparison, capitalizing on the fact that bilingual children can learn new words and grammatical structures using statistical learning mechanisms.

Statistical Learning in Children with DLD

Although there is no currently agreed-upon underlying mechanism responsible for the language difficulties that children with DLD incur, there is evidence that children with DLD have difficulty with statistical learning (Ahufinger et al., 2021; Evans et al., 2009; Gillis et al., 2022; Goffman & Gerken, 2023; Lammertink et al., 2017; McGregor et al., 2022; Obeid et al., 2016; Torkildsen et al., 2013). Children with DLD have the most difficulty with sequential dependency statistical learning tasks both in auditory (linguistic and nonlinguistic) and visual mediums (Goffman & Gerken, 2023; Lammertink et al., 2017; Obeid et al., 2016) while their distributional statistical learning is comparable to their TD peers (Hall et al., 2018). In their seminal paper on statistical learning in children with DLD, Evans et al. (2009) found that monolingual, English-speaking children with DLD performed worse than TD peers and no better than chance at recognizing nonwords they had heard from a set of six trisyllabic (CVCVCV) nonwords. When they doubled the exposure dose and replicated the experiment first with nonwords then with musical tone triplets, Evans et al. (2009) found that children with DLD still did worse than their TD peers. However, there was an interaction by stimulus type with both groups performing above chance on the speech stimuli and the DLD group performing at chance in tones. Notably, there was variation as some children with DLD scored better than chance in each condition. The difficulties in sequential dependency statistical learning for children with DLD have been seen in visual artificial grammars, as well (Gillis et al., 2022). Thus, converging evidence supports the assertion that children with DLD perform poorly in statistical learning tasks.

As statistical learning in DLD is a relatively new field of exploration, certain nuances are beginning to surface. The first point is that statistical learning may be a broad, computational theory that functions differently by domain (Sandoval et al., 2017; Siegelman, 2019). Task or domain specificity would imply that children with DLD have difficulties with statistical learning as a broad, computational mechanism given that they have demonstrated difficulties with statistical learning tasks across domains. Goffman and Gerken (2023) used this evidence to support their claim that "DLD is an impaired ability to detect and use sequential dependencies across multiple domains" (p. 1). This theory begs the question of how the deficit functions across domains if indeed statistical learning is domain specific. More research is needed to understand the specific conditions in which statistical learning difficulties arise for children with DLD and how that relates to their language development.

The second point of nuance is the need to consider *how* children with DLD engage in statistical learning tasks because they can learn in optimal conditions. McGregor et al. (2022) attempted to answer that question by studying 44 TD and 28 DLD first graders who were monolingual, English-speakers in a cross-situational word learning task. They found that children with DLD started off worse than children with TD, potentially due to difficulties encoding the novel stimuli, but then grew at a similar trajectory to their TD peers. They also found that the DLD group were limited in their strategy use to learn the stimuli. A propose-and-verify strategy is when the learner makes a hypothesis and holds onto it until it is disproven. A statistical aggregation strategy is when children amass statistical information on the co-occurrence of stimuli. Children with DLD only used a propose-and-verify strategy in the latter cycles of training while TD peers started by aggregating co-occurring dependencies in the first three cycles and updated to use the propose-and-verify strategy in the latter three cycles. Additionally, attention was related to the statistical learning performance of children with DLD while vocabulary was related to TD children's performance. These qualitative differences suggest that children with DLD are less flexible and efficient with statistical learning, using a different set of resources than their TD peers to encode statistical learning.

Statistical learning for children with DLD may be enhanced under particular conditions, including sufficient regularity of the target structure with few counterexamples and sufficient variability of the nontarget stimuli (Plante & Gómez, 2018). High regularity requires a sufficient dose of the target stimuli presentations, which, for children with DLD, may be double what children with TD need to learn (e.g., Evans et al., 2009). Gómez (2002) demonstrated the need for high variability in finding that infants and adults learned better when they had 24 (versus 12 or fewer) exemplars to learn nonadjacent dependencies. TD learners tend to show a boost from higher variability, and people with DLD extract a larger benefit from high variability than TD learners (Torkildsen et al., 2013). A potentially contradictory finding from Broedelet et al. (2023) was that 7- to 9-year-old children with DLD performed similarly well on a cross-situational word-learning task with low variation and with high contextual variation with a slight but not statistically significant preference for high context variation. They operationalized context variation as the diversity of other objects presented alongside the target word-picture referent in the learning phase. In low context variation, some of the non-target items were repeated while in the high context variation condition, the target word was always presented alongside a different non-target item. It could be that there were not enough doses (eight per word), so the regularity principle was violated, and it could also be that there was not enough variability in the high variability condition as there were only eight contrasting exemplars. This study provides some evidence that variability may function differently in statistical learning tasks for word-learning compared to grammar learning, highlighting the need for more research to better understand how task parameters and learner characteristics interact in statistical learning (McGregor et al., 2022; Siegelman, 2019). In summary, while children with DLD have difficulties with statistical learning, increasing the dose and the variability of exemplars can improve their learning.

Statistical Learning in Bilingual Children with DLD. It is important to note that most of the literature regarding DLD and statistical learning has studied monolingual, English speakers rather than bilinguals. Some of the conclusions about DLD and statistical learning may not hold or may differ by condition (e.g., sources of variability) with bilinguals. Given that bilinguals use the same computational mechanisms as monolinguals to acquire language, though, it is likely that there is some difficulty with statistical learning for bilingual children with DLD. The only study, to my knowledge, of statistical learning by bilinguals with DLD was a cross-situational word-learning study with Spanish-Catalan speaking bilingual children (ages 5.6 – 12.11 years) who had been exposed to both languages since birth (Ahufinger et al., 2021). Ahufinger et al. (2021) found that while the children with DLD did learn the words above chance (*M* = 37% accuracy, *SD* = 48%), they performed statistically significantly worse than their TD peers (*M* = 48% accuracy, *SD* = 5%). This finding of statistical learning being possible for children with

DLD but at lower levels of accuracy than their TD peers mirrors the results from the monolingual literature.

Other predictions about statistical learning abilities for bilingual children with DLD have to be extrapolated for now in the absence of evidence. Spanish-English bilinguals with DLD have difficulty with grammatical structures in both their languages (the same as many of the structures that monolinguals with DLD have difficulty with), which are typically learned by sequential statistical learning. They do not exhibit difficulties with marking word class, for example, confusing verbs and nouns (e.g., “Anda está el gato”/ “Walks is the cat”). Thus, like monolinguals with DLD, bilinguals with DLD likely do not have difficulties with distributional statistical learning. Research is needed to confirm this prediction.

Regarding optimal statistical learning conditions, in bilinguals, input variability has been examined across levels of attention but not yet by language ability. Spanish-English bilingual children with poorer attention benefited from greater speaker variability in a cross-situational word-learning paradigm (Crespo & Kaushanskaya, 2021), which is akin to higher variability improving performance for children with DLD. It may be that input variability improves the learner’s ability to hold onto the regular information because they can no longer hold the variable information in working memory (Plante & Gómez, 2018). While this may benefit TD bilingual children, it appears that bilingual children with attention difficulties receive a larger boost from highly variable input. As this variability boost in bilingual children with weaker attentional skills mirrors the variability boost for monolingual children with DLD, it can be hypothesized that bilingual children with DLD will also benefit from input variability. Again, though, more research is needed to understand the optimal statistical learning conditions for bilingual children with DLD, but for now we have some evidence that they have at least some difficulties with sequential statistical learning.

Memory and DLD

Consolidation refers to part of the memory-creation process. When people are first exposed to new stimuli, they encode it into working memory. In the case of language, the visual stimulus (if present) is encoded and paired with the phonetic sequence (i.e., the word). Encoded items are not yet in long-term memory and therefore have to be consolidated—shifted from working memory to long-term memory stores (Dudai, 2004). With effortful retrieval and with multiple exposures to a stimulus over time, the consolidated representation becomes stronger by being re-encoded and refined (e.g., Gordon, 2020; Gordon et al., 2021; McGregor et al., 2020). Sleep also plays a foundational role in facilitating consolidation into long-term memory (e.g., Axelsson et al., 2016; Born et al., 2006; Henderson et al., 2012).

Some studies show that children with DLD have difficulty with encoding but can consolidate their memories (Alt, 2011; Gordon et al., 2021; Jackson et al., 2021). Gordon et al. (2021) had 9 DLD and 9 TD monolingual, English-speaking preschoolers learn nine nonsense words across multiple training blocks each day for six consecutive days. They found that children with DLD had difficulty encoding the words (i.e., learning from each day), but the words they did encode were consolidated as evidenced by comparing the end of one day to the child’s free recall of words at the beginning of the next day. Jackson et al. (2021) conducted a similar experiment with a four-day learning trial of eight nonsense words, each day including training and testing of word knowledge. Their participants with DLD struggled with encoding, and those who had poor verbal working memory also performed poorly in naming the non-words. Thus, if children with DLD have poor encoding skills, they are predicted to form weaker phonological representations in phonologically based statistical learning tasks than TD groups at onset but would not differ in their level of retention across multiple sessions of testing.

Other studies show that consolidation may be the issue rather than encoding (Desmottes et al., 2017; Desmottes et al., 2016; Hedenius et al., 2011; Malins et al., 2021; Rice et al., 1994). These studies differed in task and methodology from the studies that found encoding but not consolidation issues. While two studies (Malins et al., 2021; Rice et al., 1994) examined word-learning, Malins et al. (2021) had children learn phonologically similar non-words and Rice et al. (1994) used incidental learning. Both of those learning tasks may have been too difficult for children with DLD to create robustly consolidated representations as Gordon et al. (2021) argued. The other three studies (Desmottes et al., 2017; Desmottes et al., 2016; Hedenius et al., 2011) examined consolidation in visual sequencing tasks (i.e., alternating serial reaction time tasks and classic serial reaction time tasks). Across these studies, the TD and DLD groups learned the pattern initially, but the DLD group did not improve across training blocks while the TD groups retained or improved (Desmottes et al., 2017; Desmottes et al., 2016; Hedenius et al., 2011). Critically, the serial reaction time tasks were non-verbal in nature, so taken together, these studies may signal that children with DLD have difficulty encoding phonological but not visual information and have difficulty consolidating visual but not phonological information. However, more information is needed about the conditions in which children with DLD encode but do not consolidate their learning given the differences across word-learning studies (Gordon et al., 2021; Jackson et al., 2021 versus Malins et al., 2021; Rice et al., 1994).

Assessing Bilingual Language

As the gateway to SLP services, assessment is high stakes. A child who does not need help but is told they do (over-identification) may have resulting negative self-attributions (Graham &

Tancredi, 2019). A child who needs help but is told they do not need it (under-identification) may receive the label of “lazy” from teachers and parents (Graham & Tancredi, 2019). Moreover, a child who is not identified but needs support can get caught in a negative feedback loop of struggling with academic work, losing motivation to try, and falling further behind making academic work more challenging. Unfortunately, both over- and under-identification are common for bilinguals in the US (Morgan, 2021; Morgan et al., 2017; Morgan et al., 2015; Sullivan & Bal, 2013). This problem stems from structural barriers to creating and obtaining accurate bilingual assessments and to applying best practices within bilingual assessment.

Bilingual language assessment has been riddled with challenges compared to the normative standard of monolingual language assessment. One of the difficulties is designing an assessment to capture the child’s ability in both languages. When determining eligibility for special education services, a lack of exposure cannot be the reason for a child’s admission to special education services per IDEA ("Individuals with Disabilities Education Improvement Act," 2004). Thus, the child’s full linguistic ability must be considered. Once a child has a minimum threshold of bilingual use (e.g., 31%; Peña et al., 2018), that child should be assessed in both their languages. The assessment, though, cannot simply be a translation of the English assessments because translations do not always apply culturally or linguistically (e.g., a possessive -s is present in English but not Spanish). Additionally, skills that develop later in English (e.g., past tense marking) are acquired earlier in Spanish and so may no longer be relevant to assess because children with DLD display a delayed growth pattern, meaning they may have acquired the past tense by the time they are assessed albeit slightly later than TD peers. Finally, different skills in each language are associated with different clinical markers, so while past and present tense verb conjugations are important to assess in English to yield high classification accuracy, in Spanish they are not diagnostically informative (Peña et al., 2023). Thus, assessments must be differentially attuned to the linguistic markers that are most discriminating by age.

Traditional monolingual English language assessment, at least in SLP practice, centers around administering a standardized, norm-referenced assessment to get a standard score that will determine the child's language status as TD or DLD (Fulcher-Rood et al., 2019; Fulcher-Rood et al., 2018). If using a standardized, norm-referenced test, it is crucial for the child to be compared to a normative sample that reflects their own experience for the comparison to be valid (e.g., same language background). Only if the items and the test overall do not demonstrate differential functioning on the basis of a child's characteristics (e.g., SES) may an assessment be deemed appropriate for the child. Because bilinguals exhibit heterogenous bilingual profiles and therefore variable growth patterns (Castilla-Earls et al., 2023; Peña et al., 2023), the task of finding an appropriately diverse normative sample is herculean. For instance, one of the most widely used standardized assessments, the Clinical Evaluation of Language Fundamentals—Fourth Edition, Spanish (CELF-4S; Semel et al., 2006), incorrectly identifies as many as a third of Spanish-English speakers from low SES backgrounds when abiding by the cutoff score from the test manual (Barragán et al., 2018). The CELF-4S is only in Spanish (i.e., not bilingual) and so is not appropriate for every bilingual profile (e.g., balanced language users). This underscores the fact that standardized, static assessments often evaluate language skills that would have been developed through exposure and practice (e.g., morphosyntax), which is a big assumption that might result in inaccurate identification.

One Spanish-English bilingual assessment battery with adequate classification accuracy has been developed for children aged 4;0 – 6;11 (years;months), the Bilingual English-Spanish

Assessment (BESA; Peña et al., 2018). It accounts for multiple bilingual profiles by recommending which language(s) to test in given the child's current use, and the final score is a composite of the child's best language scores for the morphosyntax and semantics subtests. This best language scoring effectively captures the child's linguistic ability for identification purposes (Peña et al., 2023). Even still, the BESA may not be a good match for some children who do not match the normative language sample (e.g., speakers of Argentinian Spanish), so other assessment options are needed. By comparison, there are at least five commonly used standardized omnibus tests of language for monolingual English speakers, and at least fifteen exist (Ogiela & Montzka, 2021). Moreover, when assessing a monolingual child, there are fewer factors to consider when matching a child to the normative sample than when assessing a bilingual child. This speaks to the difficulty of bilingual assessment development. Nonetheless, there is not necessarily the need for more than one norm-referenced, standardized assessment for identifying DLD if the one that exists meets the gold standard, which the BESA does. The reasons for looking at other avenues of assessment for Spanish-English speaking children at this juncture are primarily due to structural constraints.

The vast majority of SLPs, 91.7%, are monolingual English speakers (American Speech-Language-Hearing Association, 2022), and their training has typically focused on assessing monolingual, English-speaking children from middle-class backgrounds. This mismatch of monolingual clinician with monolingual training and a bilingual child who needs to be assessed gives rise to feelings of low self-efficacy and low competency for monolingual SLPs who have not received training in multilingualism (Narayanan & Ramsdell, 2022; Parveen & Santhanam, 2021; Santhanam & Parveen, 2018; Suswaram et al., 2023 but see international providers; Scharff Rethfeldt et al., 2023). Some of the structural challenges to bilingual assessment have included a lack of knowledge of the child's language and culture, a lack of appropriate assessment materials, lack of well-trained interpreters, and a lack of knowledge or guidelines for when to distinguish between DLD and TD for bilingual children (Santhanam & Parveen, 2018; Suswaram et al., 2023). Over-identification can occur when tests that were normed on monolingual populations are applied to bilingual children. This can result in an underestimate of the child's ability because it only assesses one of their languages, typically English, which is only a partial assessment of the child's full linguistic system. Under-identification can occur when professionals decide to "wait-and-see" for the child to develop English sufficiently for them to feel comfortable assessing the child's language ability. This may result in delayed identification, or it may completely miss the child because a child with DLD will have more difficulty learning English. They may never develop enough English skills for the professional to decide to assess them, or more likely, they move on to another school and slip through the cracks. These examples illustrate some of the systemic barriers to services for bilingual children, and one way to address those issues is to develop appropriate, accurate assessment measures to support timely identification.

Dynamic Assessment

Dynamic assessment addresses some of the issues with bilingual language assessment. Dynamic assessment has been variably defined and may be most clearly understood in contrast to what it is not—static assessment. Static assessment involves having the test-taker do a task or answer questions and deriving a score. Other than providing instructions, the examiner does not interact with the test-taker. In contrast, dynamic assessment, at its core, is an evaluation that involves intervention through a mediated learning experience, and the focus is on the learner's modifiability or their response to teaching (Bamford et al., 2022; Lidz, 1991; Lidz, 2014). Other common characteristics include shifting focus from the product to the process of learning and a purpose of generating information to create an intervention plan (Lidz, 2014). In contrast, static, standardized norm-referenced assessments are solely focused on the child's final score, which is often used to identify a disorder, determine severity of a disorder, or compare the child's skills to the general population. As such, dynamic assessment marries intervention and assessment.

The first challenge of bilingual assessment that dynamic assessment resolves is that it is flexible. A dynamic assessment compares the child to themselves rather than to a normative sample (though there may be a criterion), so the child is assessed considering their own bilingual experience, making the dynamic assessment highly culturally and linguistically valid. In other words, the results from a dynamic assessment for a bilingual child are expected to reduce bias and to affirm the child's ability (Lidz & Peña, 1996; Peña, 2000). The additional benefit of not comparing a bilingual child to a normative sample is that the self-comparison is likely to be more ecologically, culturally, and linguistically valid. It is ecologically valid because the child's own experiences and learning potential are unique to them, so a comparison to a general group of other children who may or may not have a similar lived experience would be less realistic than allowing the child to demonstrate their learning profile. Dynamic assessment may be culturally and linguistically valid because it does not compare bilingual children to a normative sample. It is challenging to directly match a bilingual to a normative sample given the heterogeneity of the population, and a mismatch may result in an inaccurate classification. A standardized bilingual assessment is more likely to reflect the bilingual child's cultural and linguistic experiences, but the SLP must properly apply its procedures and interpretations, which may be a challenge given SLPs' low self-efficacy and competency with bilingual assessment (Parveen & Santhanam, 2021; Santhanam & Parveen, 2018). Dynamic assessment can avoid the quagmire of the normative sample for bilinguals.

Dynamic assessment allows practitioners to feel more confident that their results reflect the child's language learning ability than some static assessments, which capture only the skills that the child has already learned. Static assessments rest on the assumption that the child has been sufficiently exposed to the linguistic skills being assessed, which may not be the case for bilinguals with variable language exposure. Because dynamic assessment allows for a wide range of language abilities to be tested, researchers have examined language learning abilities from fundamental (e.g., morpheme-learning) to higher-level language tasks (e.g., narrative retelling). This allows for bilingual children from many different bilingual profiles to be evaluated via these dynamic assessments without sacrificing the test's validity, reliability, or diagnostic power. The caveat is that, when being developed in research, the assessment should be designed for a similarly diverse range of children to confidently state that it can generalize across the myriads of contexts.

The third issue that dynamic assessment addresses is that of language mismatch between the monolingual SLPs and bilingual clients. Some dynamic assessments have demonstrated classification accuracy even when only conducted in English or the child's home language (e.g., Kapantzoglou et al., 2012). The reasoning is that the test captures the child's learning potential rather than their experience with a specific language. Boyer and Martin (2012) tested their participants in a monolingual English condition and a bilingual Spanish and English dynamic assessment condition and found no difference in performance on the dynamic assessment. It is important to note, though, that they had a poor reference standard, resulting in unclear conclusions about the classification accuracy of their measure. Nonetheless, for dynamic assessments that are shown to be valid in English only, monolingual SLPs can work without an interpreter, can appropriately administer the test, and may even be able to use the same assessment for children from different language backgrounds. To use a dynamic assessment for children from a variety of language backgrounds, though, the research would need to support its diagnostic accuracy for the language pairs of interest. Although dynamic assessments themselves cannot solve the challenge of SLPs' training, there have been efforts recently to increase SLPs' knowledge of and experience with using dynamic assessments (e.g., American Speech-Language-Hearing Association, n.d.-b; Bamford et al., 2022; Castilla-Earls et al., 2020). Finally, since dynamic assessments are often accessed as supplemental materials or appendices in academic publications, they may be relatively cheaper and more accessible than standardized tests, reducing some of the structural barriers in bilingual assessment.

Dynamic assessment has the potential to support better assessment of bilinguals. Knowing how dynamic assessment was developed and how it has evolved will provide the historical and theoretical basis to support these claims of dynamic assessment's potential. In the following subsection, I will delve into a historical review of dynamic assessment. Then, I chronical how it has been adapted for use in language assessment before reviewing modern dynamic assessments of language.

Conceptual Roots of Dynamic Assessment

Vygotsky and Dynamic Assessment. Vygotsky could be deemed the godfather of dynamic assessment as he emphasized the socially mediated nature of learning and the need to measure a child's ability relative to what they could achieve with supports (Vygotsky, 1978, p. 86). This was the basis for his Zones of Proximal Development (ZPD) theory, which was defined as, "the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers" (Vygotsky, 1978, p. 86). The ZPD was a landmark theory that contrasted with Vygotsky's early 20th century contemporaries' conceptualizations of how children learned, creating the foundation for today's study of development in psychological and educational sciences, including the introduction of dynamic assessment.

To understand the theories of Vygotsky and his contemporaries, it must be noted that the use of the terms "development" and "learning" were not used synonymously as they often are today. Rather "development" referred to a growth or maturation process in cognition while "learning" was the process by which a child internalized external teaching (e.g., a child internalizing the problem-solving method after being taught multiplication). The first major theory Vygotsky refuted was that development must happen before learning, which led to a practice of waiting for a child to develop to a certain level before teaching them a new skill so they would not be taught something too early (Vygotsky, 1978). The second predominant theory of the day came from Piaget who conceptualized that learning co-occurred with development and even that they were the same (Vygotsky, 1978). The third theory that Vygotsky contested was an amalgam of the first two views. In this theory, learning and development were seen as in a continual loop with each other with development making learning possible which would then further development (Vygotsky, 1978). Koffka made the claim that specific learning changed specific developmental processes while Thorndike argued that specific learning could influence the child's entire cognitive landscape (e.g., increased attention to Latin grammar could increase a student's ability to focus attention on any task; Vygotsky, 1978).

Vygotsky, though, rejected all those theories, proposing that learning preceded development. He claimed that it was this sequence that led to the ZPD, which is the zone of ability that a child can achieve with mediated adult support (i.e., learning). Learning furthered the child's development by influencing their cognition (Vygotsky, 1978). It is worth highlighting that the "learning" to which Vygotsky referred was higher level cognitive processes (e.g., problem-solving), and was typically taught by explicit learning mechanisms, including prompts or metacognitive strategies, such as starting a problem and seeing if the child can finish it, explaining how to solve a certain type of problem, and asking questions that might lead the child to the answer (Vygotsky, 1998).

Does the ZPD align with modern understanding of cognitive development and learning? In modern theories, development and learning are not inherently separated as distinct processes but are often seen as the same thing—new synapses and neural pathways are formed and strengthened in response to environmental stimuli. Learning is also understood today to occur from implicit or explicit stimuli. However, the ZPD is not incompatible with the current understanding of learning. Many children do perform better with adult or environmental support than independently, which may demonstrate their learning potential, so the ZPD remains a relevant theoretical backdrop for dynamic assessment.

Two major aspects of the ZPD have been integrated into dynamic assessment. At the core of the ZPD was Vygotsky's claim that a child's true potential could not be fully measured by their past learning, but a truer measure of their abilities would be to examine how they learn with support (Vygotsky, 1978, 1998). Thus, dynamic assessments have focused on the child's modifiability in response to instructional support and the process of the child's learning (Lidz, 2014). The other critical component of the ZPD was that Vygotsky (1978) was adamant that learning occurred via socially mediated experiences, so all dynamic assessment involves some kind of intervention or mediated learning experience. Mediated learning experiences have taken many forms in dynamic assessments including but not limited to graduated prompting, explicit teaching, providing feedback, and engaging metacognitive strategies (Bamford et al., 2022; Lidz, 1987). Thus, the core of the ZPD is reflected in the core of dynamic assessment. The function of a dynamic assessment is to measure the examinee's learning potential, which is accomplished by the method of a mediated learning experience.

Dynamic Assessment: Two Camps. Dynamic assessment greatly advanced in the 1970's, and two schools of thought began to coalesce. One camp created relatively structured assessments, so they could have dynamic assessments with strong, measurable psychometric properties (Lidz, 1987). Its primary researchers, who were trained by Vygotsky's former mentee, Alexander Romanovich Luria, included Milton Budoff and the team of Joseph C. Campione and Ann L. Brown (Lidz, 1987). In the absence of a term from the literature, I will call this the "psychometric" camp. The contrasting camp included Feuerstein and followers who were more concerned with describing the child's developmental stages than with creating a tightly controlled protocol (Feuerstein, 1979; Lidz, 1987). Thus, I will refer to this camp as Feuerstenian, or, to save the reader from vowel overload and to capture their alignment with Vygotsky, "purists." To clarify, Vygotsky's theories informed both camps' work (Brown & French, 1979; Feuerstein, 1979; Lidz, 1987). Although the psychometricians claimed to borrow from Vygotsky's ZPD model, Minick (1987) asserted that only Feuerstein truly encapsulated Vygotsky's aim, which was to describe a child's prospective abilities, not to create a replicable, psychometrically sound assessment. Campione and Brown likely would have agreed with this characterization, stating that their approach "was in no way meant to represent Vygotsky's original views unchanged" (Lidz, 1987, p. 83). In the following paragraphs, I will outline both camps' contributions to dynamic assessment, starting with the purists.

Feuerstein (1979) introduced the Learning Potential Assessment Device (LPAD), a model of dynamic assessment that entailed having a learner work through progressively more difficult versions of a task after mastering each level (e.g., Raven's progressive matrices) with support from the adult (i.e., the mediated learning experience). Feuerstein added that the mediated learning experience provides a measure of the child's learning potential, "his capacity to become modified by a learning process" (Feuerstein, 1979, p. 56) and became a chief proponent of examining the child's structural cognitive modifiability, "the differences in the capability of individuals to benefit from both formal and informal opportunities to learn" (Lidz, 1987, p. 380), a concept which harkens back to the ZPD. Mediated learning in the LPAD included the following five areas:

"1. regulation of behavior through inhibition and control of impulsivity

2. improvement of deficient cognitive functions

3. enrichment of the repertoire of mental operations

4. enrichment of the task-related contentual repertoire (e.g., the labeling of relationships such as "up," "down," "equal to," etc.)

5. creation of reflective, insightful processes" (Feuerstein, 1979, p. 105)

The primary analytic outcome of the LPAD entailed the "cognitive map" and the quantitative product information (i.e., raw scores or adjusted raw scores to allow for partially correct answers). The cognitive map was the manifestation of Vygotsky's theories as it was a detailed description of the child's learning performance across seven dimensions:

1. Content (e.g., math, reading, etc.)

2. Modality (e.g., verbal, visual, numeric, etc.)
3. Phase (i.e., input, elaboration, or output; the three phases of a mental act per (Feuerstein, 1979))
4. Operations ("a strategy or a set of rules, in terms of which information derived from internal and external sources is organized, transformed, manipulated, and acted upon." (Feuerstein, 1979, p. 124))
5. Level of complexity (quantity as number of items and quality as degree of novelty)
6. Level of abstraction
7. Level of efficiency (measured quality and quantitatively)

Feuerstein also quantified modifiability from the LPAD by examining the difference between the predicted and actual posttest scores divided by the standard error of estimate. Predicted posttest scores were based on a regression equation of the pretest score and a constant. This was proposed as a starting point for creating modifiability indices, but it was notable that even with his own data, Feuerstein (1979) showed that using the predicted and actual posttest scores did not necessarily reflect the child's true learning potential. One child had a negative change score, which was attributed to "some interfering variable [that] had impeded the adequate performance of the subject on the posttest" (Feuerstein, 1979, p. 311).

The greatest critiques against the Feuerstein's LPAD are the high time cost (5 – 8 hours), the extensive training required for an assessor to become proficient (Lidz, 1991), and the challenge of making it reliable and valid. Embretson (1987) proposed methods of quantifying psychometrics for this approach. It was also ambiguous how Feuerstein's vision of dynamic assessment fit within the education system. He vehemently argued against the biased IQ tests that penalized children from "culturally deprived" and low socioeconomic status (SES) backgrounds, but it was unclear how he wanted his assessment to be used within the educational system to address this problem. Was it meant to test whole classes or schools to describe each child's developmental level and needs? Did he envision a future beyond special education, in which every child received individualized instruction based on their dynamic assessment results? This remains unanswered. Lidz (1991) also raised the critique that Feuerstein's (and Budoff's) tasks were poorly linked to daily classroom activities. In parallel to Feuerstein, the psychometricians were researching dynamic assessment with a more replicable, structured approach that would potentially identify students for special education more accurately.

The psychometric camp was primarily comprised of two major research groups—Budoff and colleagues and Brown and Campione and colleagues. Both camps preferred relatively standardized procedures and were less theory-driven than the purist approach. Budoff used similar assessments to Feuerstein (e.g., Raven's progressive matrices) in what he called the "Learning Potential Assessment," which, unlike Feuerstein's purpose, was intended to identify children for special education (Babad & Budoff, 1974; Budoff & Pagell, 1968; Lidz, 1987). In the Learning Potential Assessment, children were guided through a mediated learning experience with structured coaching to familiarize the child with the test procedures, the most important feature of which was "a verbal explanation of the problem's critical features" (Lidz, 1991, p. 25). The primary critiques of Budoff's work were that there was no substantive reference point to classify a child into Budoff's categories of "nongainer," "gainer," or a "high scorer," and that it was not clear how to link the information found on the Learning Potential Assessment to classroom instruction (Lidz, 1991). It is challenging to assess whether Budoff's work met its goal of accurately identifying children for special education because he did not report sensitivity and specificity nor how students were classified (i.e., as needing special education or not) (Babad &

Budoff, 1974; Budoff & Pagell, 1968). Additionally, there is an ethical concern of teaching a child to improve performance on a standardized test that is commonly used for identification of a disability as the child's performance could be enhanced in future attempts at the assessment, potentially unbeknownst to the assessor.

Campione and Brown used academically-based tasks, including math and reading skills, which they trained children on through a series of graduated prompts from most general to most specific (Campione et al., 1985; Ferrara et al., 1986; Lidz, 1991). As mentioned previously, they grounded their work in Vygotsky's ZPD (Brown & French, 1979; Campione et al., 1984), but purists critiqued their prompts as not being informative enough to adhere to Vygotsky's mandate for a descriptive profile of the child's learning ability (Lidz, 1987, 1991). Campione and Brown specified that it was the amount of instruction (i.e., prompts) that could be measured as learning transfer rather than the amount of improvement (Bryant, 1982; Campione et al., 1985; Ferrara et al., 1986). The prompts did not tap into a variety of metacognitive processes or problem-solving strategies, in contrast with Feuerstein's LPAD, so Campione and Brown's dynamic assessments did not generate sufficient information for intervention planning (Lidz, 1991). This critique could be characterized as primarily a difference of purposes for dynamic assessment, but it does hold merit that it is unclear whether the pattern of moving from general to specific prompts were the most appropriate for coaxing out learning potential. Campione and Brown's research group reported that transfer (i.e., generalization) from the teaching phase to posttest phase reliably distinguished between high and low ability groups on a matrix task and a fill-in-the-blank pattern task, but sensitivity, specificity, or classification as needing special education were not reported (Bryant, 1982; Campione et al., 1985). Thus, their claims are difficult to evaluate. Nonetheless, the psychometric camp advanced dynamic assessment research, demonstrating that it was possible to yield informative, replicable assessments when teaching was integrated.

Updating Dynamic Assessment: What to Keep and What to Leave. The psychometricians and purists paved the way for modern dynamic assessment research, and by evaluating their work against previous critiques and current knowledge, we can select the most critical elements to bring forward into dynamic assessment research today. In the final part of this section, I will detail the still-applicable aspects of and limitations of the psychometric and purist approaches to establish which theories are critical to modern dynamic assessment and which elements must be revised with our updated understanding of language and cognition. The beauty of dynamic assessment lies in its flexibility to take on a myriad of forms as long as its core principles are retained.

The underlying tenet to retain is not specific to dynamic assessment's administration, but rather to the need for dynamic assessment. At its core is the idea that people with disabilities deserve to be met with high educational expectations and quality educational instruction (Feuerstein, 1979; Vygotsky, 1978). This idea is intertwined with the assumption underlying dynamic assessment that all children can learn, and it is important to provide intervention (even within assessment) for children to be able to show their learning potential. Thus, the first element of dynamic assessment to keep is its cardinal emphasis on inducing a change in the learner to understand their modifiability and learning potential.

Another key to dynamic assessment is the stress placed on chronicling the process of the child's learning. Typically, product information, meaning the child's answers and the final outcomes, are the focus of static assessments, but the addition of process information allows the examiner to glean how the child is learning. This can be done through marrying both purist and psychometric approaches by allowing for quantitative scoring and some level of scriptedness or pre-determined structure while also collecting qualitative data about the child's change, including modifiability ratings. Although purists might argue against any level of scriptedness, its necessity is practical. In the United States' highly litigious context with many written and unspoken policies that influence clinical decision-making (Fulcher-Rood & Castilla-Earls, 2023; Fulcher-Rood et al., 2018; Ogiela & Montzka, 2021), having a reliable assessment helps protect practitioners. Moreover, some level of scriptedness and structure is almost a prerequisite for SLPs Pathologists (SLPs) to be open to using a new instrument that they can implement with minimal time and resources (Denman et al., 2021; Fulcher-Rood & Castilla-Earls, 2023; Fulcher-Rood et al., 2018; Ogiela & Montzka, 2021). It is important to retain the core values of dynamic assessment while borrowing from both purist and psychometric traditions.

Assessment Population and Purpose. Other elements of dynamic assessment practice and theory should be updated, discarded, or viewed with nuance given our current understanding of cognition and language. To start, both the purists' and psychometrician's target populations were underspecified per the modern standard and were not primarily children with disabilities. Rather, they focused on children who were educationally behind due to being from minoritized backgrounds[1] as their experiences did not reflect the curriculum and structure of the school system, which had been designed for middle- and upper-class White children (Feuerstein, 1979).

[1] Feuerstein used the term "retarded performer," which included children with "obvious" cognitive impairments and children who were "socially and culturally deprived," which he defined as, "an individual or a group that has become alienated from its own culture. This alienation may have been produced by a variety of sociological, economic, geopolitical, psychophysical, and cultural determinants affecting the individual and/or the group... It is by this lack of such an adherence, because of poor mediating transmissional processes, that an individual, or a group, becomes culturally deprived" (Feuerstein, 1979, p. 39). He also was adamant that most children who were seen as educationally behind were so due to too few quality adult-child teaching interactions (mediated learning experiences) in the child's life, so their seemingly low cognitive performance could be enhanced through quality mediated learning experiences. The term "culturally deprived" is problematic in its assumption that a person does not have culture, and so I avoid it as much as possible within this document.

The goal was usually to prove that these children could learn if given the proper experiences and to show that they did not need special education. This population of interest is important to note because it differs from most modern dynamic assessment in Speech, Language, and Hearing Sciences, which is centered on children with disabilities. Modern dynamic assessment for DLD is also concerned with ensuring that children from minoritized backgrounds are not misidentified, particularly those who have the intersectional identities of DLD and bilingualism. The primary difference is that modern dynamic assessment for identifying DLD tends to be situated as an assessment that will be used in the special education evaluation process, rather than as a general aptitude assessment writ large.

After clarifying the population, there is the need to clarify the purpose. The function of the dynamic assessment is to ascertain the child's learning potential. Each dynamic assessment can have this shared function but may diverge on their purposes (Bamford et al., 2022). Early dynamic assessments spanned a wide range of purposes with an "aim to engineer maximal levels of performance; others seek to measure the magnitude of response to instruction; still others focus on the efficiency of operation of specific cognitive processes" (Lidz, 1987, p. 83). In fact, this divergence of purposes was the one of the key differences between the psychometric and purist camps. Psychometricians used dynamic assessment for the purpose of identifying children who needed educational support (Campione et al., 1984). Purists used dynamic assessment for the purpose of planning an instructional program (Feuerstein, 1979). However, this characterization is overly simple and generous to both camps.

Both stated and implied multiple purposes for their singular dynamic assessments, which, in modern test design, is unsanctimonious. Psychometricians also included the purposes of understanding the etiology of the difficulty, which would help formulate plans for instruction

(Lidz, 1987), and using dynamic assessment “as an alternative to IQ tests” (Lidz, 1991, p. 21), and Feuerstein (1979) included the following purposes: 1) assessing modifiability, 2) determining the level of thinking (separating elementary and higher mental processes) that can be achieved with support, 3) determining the amount of teaching necessary, 4) determining whether the participant can generalize the taught skill(s), and 5) obtaining a profile of strengths and weaknesses to aid in instructional planning (Feuerstein, 1979, pp. 91-92). Moving forward, assessments should be designed with one clearly stated purpose (Messick, 1989; Plante & Vance, 1994). Otherwise, they risk introducing sources of “construct-irrelevant variance” (Messick, 1989), meaning sources of error that reduce the overall construct validity of the assessment.

We must consider how well the dynamic assessments from the psychometricians and Feuerstein fit the singular purpose of each dynamic assessment and discard or modify those that do not. Overly ambitious purposes may be a fault of adhering to Vygotsky too tightly. Vygotsky wanted to “diagnose” but did not use the term as we do in modern literature, meaning to identify a disability, but rather Vygotsky’s “diagnosing” referred to describing a child’s potential ability. While describing a child’s cognitive-linguistic profile is an important part of the evaluation process, it is not the same as identifying a child with a disability. Many modern dynamic assessments have explicitly stated their purposes (e.g., Hasson et al., 2012; Peña et al., 2006). However, potentially in part due to this historical difference in terminology, other researchers have left the door open to interpretation of their purpose or the intended use of the dynamic assessment. For example, Hasson et al. (2012) stated that their purpose was to “elicit significantly different performances” (p. 62) between TD bilingual children and children referred to SLPs for language concerns. In future research, though, Camilleri et al. (2014) used many of the same tasks from Hasson et al. (2012) to “derive profiles of strengths and weakness across language areas” (p. 57). The other potential reason researchers may have ambiguous dynamic assessment purposes may be due to borrowing from Feuerstein’s and others’ work without critical consideration of how each design element serves the purpose of the dynamic assessment. In light of this, I will lay out how the elements I incorporated into my dynamic assessment serve my purpose in my “Theoretical Framework.”

Language not Cognition. One other major area of difference between early dynamic assessment and the current dynamic assessment is the theories that influenced the design of early dynamic assessments. Most early dynamic assessment researchers focused on cognitive skills and considered learning to be Level I—associative—or Level II—learning in which the input is transformed by mental operations before responding (Jensen, 1969 as cited by Feuerstein, 1979). The majority designed their dynamic assessments for higher level (i.e., Level II) thinking (e.g., Babad & Budoff, 1974; Campione et al., 1985; Feuerstein, 1979). Crucially, though, the early dynamic assessment researchers were studying cognition, not language, and language is the focus of the current project.

Language has fundamental (e.g., semantics, morphosyntax) and higher-order dimensions (e.g., discourse) that could approximate the Level I and Level II designations by Feuerstein and his contemporaries, though with some differences as semantics and morphosyntax tend to separate into distinct factors as children mature (Language & Reading Research, 2017; Language and Reading Research Consortium et al., 2018; Language Reading Research Consortium, 2015). The purpose of the dynamic assessment in the current study is to identify DLD, which can be identified with good prediction of later outcomes by four years old (Bishop et al., 2016; Bishop et al., 2017). Thus, a focus on “Level I” or more fundamental linguistic processes, such as morphology, are a good fit for a diagnostic assessment because many morphological skills are expected to have been mastered by TD 4- and 5-year-olds (Brown, 1976; Dunn Davison & Scheffner Hammer, 2012; Paul & Alforde, 1993). Children with DLD, though, demonstrate a prolonged period toward mastery, so they are expected to demonstrate weaker morphological learning skills compared to their TD peers (Castilla-Earls et al., 2021; Castilla-Earls et al., 2023). This difference in expected developmental timeline results in a bimodal distribution that facilitates the delineation of DLD and TD skills. In contrast, discourse abilities (like Level II) are still emerging (Bitetti & Hammer, 2021; Lucero, 2018) in children with TD and DLD and may not result in an easily distinguishable distribution.

Implicit not Explicit. Another prominent, influential theory on dynamic assessment was socially mediated learning, which Vygotsky (1978) championed. Socially mediated learning entails providing feedback, definitions, and questioning to lead the child to higher-order thinking (i.e., Level II). These socially mediated learning practices are all methods of explicit teaching. As dynamic assessment is used for young children and is focused on language, the emphasis on explicit, socially mediated learning may be supplanted by other forms of learning to better match the assessment's purpose.

While language is learned in a social context in the natural environment, it is learned primarily implicitly rather than explicitly (Erickson & Thiessen, 2015), especially in the pre-school years. Therefore, even though traditional dynamic assessments were based on theories of explicit learning, the implicit mechanism might be a better target for this preschool population. Another reason for choosing implicit learning in the current context is that children with DLD tend to have more difficulty with statistical learning and sequential dependency learning than TD children (Evans et al., 2009; Goffman & Gerken, 2023; Lammertink et al., 2017), again serving to create the bimodal distribution that is ideal for identification purposes. I will expand upon implicit, statistical learning in the "Statistical Learning" section after presenting how researchers have used dynamic assessment to identify language-learning difficulties (including DLD) from the late 20th century through today.

Modern Dynamic Assessment of Language

Modern dynamic assessment of language has spanned multiple assessment purposes across differing populations. In this section, I will exclusively discuss dynamic assessments created with the purpose of identifying DLD for bi- or multilingual children. Most dynamic assessments, though, have been focused on vocabulary, narrative retells, and morphology. For dynamic assessments focused on vocabulary or word-learning, some helped children learn real-word labels (e.g., Peña, 2000) while others used non-words (e.g., Kapantzoglou et al., 2012), and at least three also worked on receptive vocabulary knowledge in addition to expressive vocabulary (Camilleri & Law, 2007; Hasson et al., 2012; Maragkaki & Hessels, 2016). Others practiced strategies of inferential word learning (Petersen et al., 2020) and categorizing words (Ukrainetz et al., 2000). Several dynamic assessment studies have examined narrative retells, focusing on story grammar and some also on complex language (e.g., Frahm, 2021; Petersen et al., 2017). Five dynamic assessments used an invented morpheme suffix to signify "half of" either in English (/-ə/: Boyer & Martin, 2012; Kohnert & Danahy, 2007; Roseberry & Connell, 1991), Spanish (/-be(i)/: Boyer & Martin, 2012; Kohnert & Danahy, 2007), (/-ono/: Restrepo, 1998), or Vietnamese (/-po/: Hwa-Froelich & Matsuo, 2005). Additionally, Jacobs (1998) examined global language ability on a computerized dynamic assessment across receptive and expressive vocabulary, morphemes, and auditory verbal sequencing. For children referred for a language evaluation, Camilleri et al. (2014) explored bilingual children's English language abilities across multiple domains—vocabulary, syntax, and phonology—as a pre-diagnostic tool.

The current state of dynamic assessment research for diagnosing bilinguals with language disorders is still emergent (Hunt et al., 2022), so while the early evidence is promising regarding dynamic assessment as a viable practice that could result in diagnostically accurate findings, most studies have had methodological flaws. A systematic review and a meta-analysis (Hunt et al., 2022; Orellana et al., 2019), respectively, reported on the methodological quality and the level of evidence from dynamic language assessment for 13 studies. Orellana et al. (2019) reviewed six articles, seven studies (Kapantzoglou et al., 2012; Kramer et al., 2009; Peña, 2000; Peña et al., 2014; Petersen et al., 2017; Roseberry & Connell, 1991) that involved the dynamic assessment for the diagnosis of language impairment of children who spoke English and another language. They used the Critical Appraisal of Diagnostic Evidence (CADE; Dollaghan, 2007) to rate the quality of the studies, awarding points, up to 9, for each quality category that had been met. The outcomes ranged from 1 out of 9 (Roseberry & Connell, 1991) to 7 out of 9 (Petersen et al., 2017). All studies used the same index measure assessments (i.e., the dynamic assessments) for every child in their respective samples, which was a positive quality indicator. Most studies (6/7) described the procedures to allow for a high level of replicability. About half had blinded testers (3/7), included fidelity information (3/7), and used a valid, reasonable, and/or reliable reference standard (4/7). A reference standard is the diagnostic tool to which the index standard (i.e., the instrument of study) is compared to establish the index standard's classification accuracy. Because there is not a singular ideal reference standard for bilingual children, the reference standards varied greatly across studies, and only two studies employed independent testing on the reference standard.

Another consistent issue was the use of case-control designs and two-gate designs (i.e., assessor knows the diagnostic status prior to assessment; Rutjes et al., 2005). Both of these designs contribute to spectrum bias, which entails having a sample consisting of only extreme ends of the spectrum (Mullerhein & Miller, 2002). This can lead to an overestimation of the diagnostic accuracy of the assessment because borderline cases, children with comorbidities, or complex cases may have been omitted (Dollaghan, 2007; Mullerhein & Miller, 2002). Two studies did use the preferred one-gate design in which the diagnostic status was unknown to the examiners prior to testing.

All studies but Roseberry and Connell (1991) included some form of modifiability rating, typically on a researcher-created checklist or one borrowed from previous research. Of those studies, all but one found a significant difference between groups based on modifiability ratings. In contrast, in Analysis of Variance (ANOVA) or *t* test analyses of the TD vs DLD groups on the change across pre- and posttests, only one study (Kramer et al., 2009) found a statistically significant group difference, but overall, the TD children performed statistically significantly better than their peers with DLD on the pre- and posttests alone. The ideal diagnostic accuracy would be indicated by sensitivity and specificity values greater than 80% for a fair rating or greater than 90% for a good rating (Plante & Vance, 1994), and they would have positive likelihood ratios (LR+) greater than or equal to 10 and negative likelihood ratios (LR-) less than or equal to 0.10 (Dollaghan & Horner, 2011). As indicated by the overall value ranges for sensitivity (76% - 100%), specificity (57% - 96%), LR+ (3.2 - 16.129), and LR- (0 - 0.288) the dynamic assessments ranged greatly in their diagnostic accuracy, but overall, the studies met the acceptable criteria of being 80% or greater for specificity (Plante & Vance, 1994) with the one exception of the cross-validation by Peña et al. (2014). Thus, Orellana et al. (2019) recommended the use of dynamic assessment for bilingual children in conjunction with other assessment procedures. They also called for studies with improved methodological quality.

Hunt et al. (2022) conducted a systematic review for children under 12 years who spoke a language other than the majority language at home and who underwent a dynamic assessment for a diagnosis of DLD. To assess the quality of their ten identified studies, four of which were also in the previously described meta-analysis, (Boyer & Martin, 2012; Camilleri & Law, 2007; Hasson et al., 2012; Jacobs, 1998; Kapantzoglou et al., 2012; Lazewnik et al., 2019; Maragkaki & Hessels, 2016; Peña et al., 2014; Petersen et al., 2017; Roseberry & Connell, 1991), they used The Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2; Whiting et al., 2011). The QUADAS-2 primarily identified the studies' risk of bias and the applicability of the results to the dynamic assessment of the whole population of multilingual children. Most studies had a high risk of patient selection bias given their use of case-control designs (all but one: Boyer & Martin, 2012) and given their sampling methods that selected for certain participant characteristics like SES (all but one: Hasson et al., 2012). Most studies had an unclear rating of bias regarding the index test (i.e., dynamic assessment) with a few scoring as low risk and one as high risk (Lazewnik et al., 2019). Six of the studies scored in the low-risk range for the reference standard due to using a convergence of evidence as their reference standard (Castilla-Earls et al., 2020) while three of the remaining studies did not independently confirm diagnoses, and one did not provide sufficient detail. Finally, two studies used different reference standards across participants, and one introduced selection bias by removing participants who exhibited a basement effect on their tasks, all three receiving high risk of bias for the flow and timing of their studies.

Additionally, both Hunt et al. (2022) and Orellana et al. (2019) acknowledged the small sample sizes in the studies they reviewed. Orellana et al. (2019) compared studies to a standard of having greater than 30 participants (the basis for this number was not provided), which Peña

(2000), Peña et al. (2014), and Petersen et al. (2017) met. Hunt et al. (2022) held studies to a standard of a minimum of 48 participants per group for a medium effect size to be detected with 80% power (Gaeta & Brydges, 2020). Per those criteria, all studies in both reviews were underpowered.

Despite the methodological shortcomings, Hunt et al. (2022) and Orellana et al. (2019) viewed dynamic assessment as a viable practice for accurately identifying DLD in bilingual children. Pre-test scores were not a significant indicator of DLD status, likely due to too much variance as the pre-test reflected the child's static language ability (Peña et al., 2001). Posttest scores were a significant indicator in many assessments likely because they reflected the learning that had occurred. Interestingly, though, the change between pre- and posttests did not contribute substantially to diagnostic accuracy. That may have been due to the large variance in the pre-test scores, as well. For the studies that included modifiability ratings, modifiability almost always classified the child with appropriate sensitivity and specificity (> 80% per Plante & Vance, 1994). Moreover, combining the modifiability scores with the child's posttest yielded highly accurate classification of DLD or TD. Studies not included in those reviews also converged on those same conclusions about the most diagnostically accurate features (e.g., Henderson et al., 2018; Peña et al., 2006; Peña et al., 2001; Peña et al., 2007; Petersen et al., 2020; Ukrainetz et al., 2000).

The reviews noted a few other points that might impact classification accuracy. Hunt et al. (2022) noted that not an immediate, but a delayed posttest may yield greater differences across groups than purely a comparison of immediate posttests (Hasson et al., 2012; Maragkaki & Hessels, 2016). They also found that diagnostic accuracy may increase with the age of the children tested. Hunt et al. (2022) suggested that task familiarity might also improve diagnostic accuracy as studies of invented vocabulary or morphemes demonstrated poor sensitivity, and learning invented language components are relatively uncommon tasks. Additionally, they found that all the included studies reported high reliability of their index measures. Hunt et al. (2022) also noted that the variety of languages used in the European studies demonstrated the generalizability of dynamic assessment across language pairings. In summary, dynamic assessment for identifying DLD in bilingual children is still an emerging area of research informed primarily by early-stage studies.

In the next section, I will describe the study upon which the current dynamic assessment is based.

Roseberry and Connell (1991)

I propose to follow a similar dynamic assessment procedure to Roseberry and Connell (1991) and those who replicated their procedures (Boyer & Martin, 2012; Hwa-Froelich & Matsuo, 2005; Kohnert & Danahy, 2007; Restrepo, 1998). Roseberry and Connell (1991) created an invented morpheme suffix /-ə/ that meant "part of" to assessing morpheme learning. With 26 children of Mexican descent in California (ages 4;5 – 6;4), they followed the test—teach—re-test model. The participants all spoke Spanish at home and had at least one year of English exposure in school and "had been designated as limited English proficient by a bilingual specialist or by their classroom teacher" (p. 598). There was no other information on the children's bilingual profiles.

Participants were designated as TD (they used "normal", *n* = 13) or DLD (they used "Specific Language Impairment (SLI)", *n* = 13). The TD group was selected based on (1) teacher identification as making expected linguistic and academic progress, (2) no social or emotional developmental problems, and (3) never referred for special education. Their scores on the Test of

Auditory Comprehension of Language—Spanish (TACL; Carrow, 1973 as cited by Roseberry & Connell, 1991) were reported as age-equivalents, *M* = 68.15, *SD* = 4.5. The IQ measure was the test of nonverbal intelligence (TONI; Brown et al., 1986 as cited by Roseberry & Connell, 1991), and the scores were a mean of 96.4 (*SD* = 6.4, range = 89.5 – 109.5). Additionally, a bilingual SLP judged their conversational language samples to be commensurate with TD.

Children in the DLD group were selected on the basis of a diagnosis by a bilingual SLP. All participants were tested on a standardized receptive and an expressive measure of language, but those measures varied across children. Expressive measures included the Expressive One-Word Picture Vocabulary Test (EOWPVT; Gardner, 1979 as cited by Roseberry & Connell, 1991), the Preschool Language Scales (PLS; Zimmerman et al., 1979 as cited by Roseberry & Connell, 1991), and the Del Rio (Toronto et al., 1975 as cited by Roseberry & Connell, 1991). It was not stated whether the EOWPVT was in English or Spanish. The receptive measures included the TACL, the PLS, and the Del Rio. Bias may have been introduced when children were tested on different reference standards as each of these assessments may have had differing classification accuracies, resulting in differing confidence levels of the child's language ability status as DLD. Additionally, the classification accuracy of these reference standards was not reported. In line with the California state guidelines for diagnosing DLD at the time, participants had either (1) age-equivalency scores of at least 15 months below their chronological age (CA), or (2) percentile scores were lower than the 10^{th} percentile. Age equivalency scores have been shown to be inaccurate interpretations of children's performance on standardized tests (Bracken; 1988; Reynolds, 1981). The authors stated that they confirmed their decisions by having (1) a bilingual SLP judge their conversational language samples to be commensurate with DLD, and (2) having their teacher state that the child was learning English at a rate slower than their peers.

Roseberry and Connell (1991) did not state whether the SLP and teachers were blinded to the child's diagnostic status nor how many children were potentially included based on test scores but then removed based on these judgments.

The dynamic assessment was administered across two 15 – 20-minute sessions in the exact same way both sessions. All instructions about the tasks were provided in English and Spanish, but the language of teaching and the child's expected responses was English. First, the pre-test was a check to ensure the child knew the vocabulary that they would be shown in the posttest. If a child did not know a label, they were told the name and asked to label it again after seeing other items. During the teaching phase, the examiner showed each child 20 different black-and-white line drawings of objects that were wholes and their partial counterparts, pointing to and labeling first the whole object then the partial object all with the same carrier phase, "This is a NOUN(/-ə/), a NOUN(/-ə/)." Children were instructed to listen and not speak. Because the child was not required to speak during the teaching phase, the teaching phase was not interactionist, and no modifiability rating (i.e., process information) was included. The images were black-and-white line drawings, and the partial images had a diagonal slice removed from the top right corner. In the posttest phase, the examiner labeled the same 20 pictures from the pre-test in their whole forms, "This is a NOUN," and immediately after each whole, asked the child "What is this?" to label the item in its partial form with the invented morpheme. A non-response was prompted by a repeat of the question. The authors did not state whether feedback was given, but it is assumed not because they stated, "When a response was given, the next pair was presented." (Roseberry & Connell, 1991, p. 600). It is also assumed the posttest phase occurred immediately after the pretest because no indication was given about a time delay. They had inter-rater agreement on the pre-test and posttest probe responses for all sessions, and five of the responses out of 1,040 resulted in disagreement. The authors did not state how disagreements were resolved.

To analyze results, Roseberry and Connell (1991) ran a repeated measures ANOVA for the posttest scores from both sessions across the TD and DLD groups. They reported that there was different variance across the groups but that the principle of homogeneity of variance was not violated per results from F-max, Bartlett's, and Cochran's tests (specific results from those were not reported). The ANOVA resulted in a statistically significant group difference with higher accuracy scores for the TD group (M = 17.2, SD = 6.7, $F(2, 25) = 18.24$, $p < .0002$) than the DLD group (M = 4.7, SD = 8.4). There was no statistically significant difference in performance across sessions; only four children increased scores across sessions. For classification accuracy, 10/13 children were correctly identified as DLD in the DLD group and 12/13 were correctly identified as TD in the TD group (sensitivity = 76.92%, specificity = 92.31%).

There were multiple methodological flaws with this study by Roseberry and Connell (1991) that make its conclusions difficult to interpret. The primary issue is that the identification of the TD and DLD groups was replete with potential sources of error and bias, which reduces confidence in the final classification accuracy numbers. Additionally, not all TD children learned the invented morpheme. The authors did not provide a theoretical basis for their explicit teaching style other than the results from Connell (1987) which showed that children with DLD performed worse compared to TD children with a modeling procedure but had no difficulty learning after imitation. Thus, they targeted the assessment at learning style to see if that could accurately classify bilingual children with DLD. However, they did not specify their reasoning behind choosing the parameters of the modeling task, including the dose in the teaching phase, the platform sentence to present the whole or target word, cutting out the top right part of the image via a diagonal, the morpheme choice, the selection of the words, randomization of the words (if any), teaching the morpheme over two days, and using the same stimuli both days. Roseberry and Connell (1991) did not classify this procedure as a dynamic assessment originally, but both Orellana et al. (2019) and Hunt et al. (2022) found that this study met their criteria for dynamic assessment in their respective meta-analysis and systematic review. As such, Roseberry and Connell (1991) did not include some of the components that are considered critical to dynamic assessments, such as interaction during the teaching phase and scoring modifiability. This dynamic assessment was replicated in several other studies, which I will detail in the next subsection.

Studies Inspired by Roseberry and Connell (1991). Four studies have modified the procedures from Roseberry and Connell (1991), two of which examined classification accuracy of the task (Boyer & Martin, 2012; Restrepo, 1998) and two of which examined its feasibility only with a TD population (Hwa-Froelich & Matsuo, 2005; Kohnert & Danahy, 2007). All but Hwa-Froelich and Matsuo (2005) worked with Spanish-English bilingual children. See Table 1 for an overview comparison of each replication.

Table 1

Summary of Roseberry and Connell (1991) and Studies with Dynamic Assessments Based on Roseberry and Connell (1991)

Variable	Roseberry & Connell (1991)	Restrepo (1998)	Boyer & Martin (2012)	Kohnert & Danahy (2007)	Hwa-Froelich & Matsuo (2005)
Sample Size	26	62	29	20	21
Participant Ages	4;5 – 6;4 DLD: 5;5(0;6) TD: 5;8(0;4)	5;0 – 7;1 DLD: 6;1(0;8) TD: 6;2(0;7)	3;1 – 5;10 4;6(0;10)	3;6 – 5;8 4;7	3;6 – 5;6 4;9
Language of Instructions	Spanish & English	Spanish	Spanish & English[a]	**English group**: English + Spanish **Spanish group**: Spanish	English & Vietnamese
Language of Test	English	Spanish	English & Spanish or English[a]	English or Spanish	English & Vietnamese
Morpheme	-ə	-ono	-ə, -bei	-ə, -be	-po
Stimuli	Black-and-white, drawing	Black-and-white, drawing	Colored, drawing	Black-and-white, drawing	Not reported
Sensitivity	77%	65%	71%	N/A	N/A
Specificity	92%	61%	57%	N/A	N/A
Reference Standard	Identified by bilingual SLP: age-equivalency scores 15 months behind or <10th percentile	Identified by bilingual SLP based on multiple sources of evidence	Teacher report & class observation	N/A	N/A
Learned by TDs?	12/13	Not reported	**90% accuracy** 10/21 4- or 5-year-olds **≥ 1 correct** 16/29 all participants	**90% accuracy** English: 1/10 Spanish: 7/10 **≥ 1 correct** 13/20	18/21 receptive 12/21 expressive

Variable	Roseberry & Connell (1991)	Restrepo (1998)	Boyer & Martin (2012)	Kohnert & Danahy (2007)	Hwa-Froelich & Matsuo (2005)
Procedures	**Test**: 20 vocab check **Teach:** 20 **Test:** 20	**Test**: 9 vocab check **Teach:** 9 + imitation **Feedback:** test 4 + accuracy for first 2, neutral for last 2 **Test**: 4 novel words, 2x	**Test**: 20 vocab check **Teach:** 20 (first with explicit instruction) **Test**: 10	**Test**: 10 vocab check **Teach:** 20 **Test:** 10	**Test**: 20 receptive & expressive vocab check **Teach:** 10 **Test:** 20 receptive, then expressive

Note. Ages presented as year;month with the range in the first line and Mean(Standard Deviation) in the lines below. The last two rows do not report the SD because it was not reported in the original articles. TD = Typically developing group, DLD = Developmental Language Disorder group.

[a]This study consisted of two conditions for each child. The pretest was administered in both languages across both conditions. In the bilingual condition, English and Spanish were used. In the English-only condition, only English was used for the teaching phase and posttest.

The earliest study to take after Roseberry and Connell (1991) was part of a larger study by Restrepo (1998) to compare the classification accuracy of multiple methods of evaluating Mexican-American school-age children (5;0 – 7;1) who primarily spoke Spanish and were beginning to be exposed to English in school, but the language of instruction was Spanish. The TD and DLD groups were matched for age, gender, and school. The reference standard was identification of DLD by a Spanish-English bilingual SLP, all of whom had identified the children with DLD through a language sample, teacher and parent interview, classroom observation, non-standardized measures of auditory comprehension and oral expression (not specified), and a standardized assessment that had been normed for Mexican-American children—Pruebas de Expresión Oral y de Percepción de La Lengua Español (Mares, 1981 as cited by Restrepo, 1998). A teacher questionnaire also had to reflect the determination of DLD. Notably, only the children with DLD were identified by an SLP, so only they were administered this reference standard, which is incongruent with best practice (Dollaghan, 2007b). The dynamic assessment was all administered in Spanish and presented the suffix of /-ono/ for partial items to be congruent with Spanish phonotactics. The pretest consisted of labeling nine vocabulary items. For the teaching phase, the child was shown the whole and partial version of the nine items from the pretest, and they repeated the examiner's label of the partial noun. The exact script was not provided. Then, the child chose four of the nine nouns, and the examiner asked them to label the partial form for each, providing positive feedback for the first two and neutral feedback for the next two. For the posttest, the child had to affix the bound morpheme to a set of four novel non-words in Spanish, three of which had been taught in the previous task. The child was asked to respond to each item twice, so the highest possible score was eight points. Approximations of the stem word or morpheme were considered correct. The score on the posttest was used (TD M = 4.35. SD = 3.32; DLD M = 2.80, SD = 3.41), and the measure had a sensitivity of 65.22% (15/23) and specificity of 60.87% (14/23). The final analysis in the study was a stepwise discriminant analysis, which was run with 23 randomly selected matched pairs from the total set of 31 pairs in the sample. Four of the measures tested resulted in good sensitivity and specificity, but this novel morpheme task was not one of those four.

Several differences may be the culprit for the reduced classification accuracy in Restrepo (1998) compared to Roseberry and Connell (1991). Children in Restrepo (1998) were older on average, which may have meant that children in both groups were more capable at learning the invented morpheme, reducing the potential difference in scores across groups. The teaching procedures and posttest stimuli also differed greatly. The first difference was that the number of stimuli were greatly reduced from twenty to nine in the teaching phase, reducing the number of learning opportunities for the child. Additionally, the posttest only had four unique items, again reducing the number of chances for the child to correctly respond. This limited the possible spread of scores, as well as the opportunities for accurate responses, illustrated by the fact that the TD group's mean (4.53) was similar to the DLD group's mean (4.7) in Roseberry and Connell (1991), yet their TD group mean was 17.2. The posttest also required the child to respond to each item twice, which could have introduced the bias of practice effects, potentially artificially increasing accuracy for both groups and therefore reducing the distance of scores between groups. Using novel, nonsense word nouns as the stems in the posttest also introduced variation from Roseberry and Connell (1991). At face value, an unfamiliar platform word appears to be more challenging to combine with the novel morpheme than a known word. However, some recent evidence questions that assumption, at least in statistical learning paradigms, revealing that unfamiliar words or structures may facilitate learning (Owen Van

Horne et al., 2017). Those studies, though, incorporated unfamiliar stimuli during teaching, which is the opposite to Restrepo (1998) who taught with familiar base words and tested with unfamiliar words. Thus, it is unclear how that methodological choice influenced the results. Finally, Restrepo (1998) used a reference standard of converging evidence, which is the gold standard in the absence of a single suitable measure (Castilla-Earls et al., 2020) and therefore more trustworthy than the reference standard used by Roseberry and Connell (1991). Thus, the two samples may not be directly comparable.

The other study inspired by Roseberry and Connell (1991) that may be presented as a classification accuracy was a study by Boyer and Martin (2012) of 29 Spanish-English emergent bilingual preschoolers in Head Start (3;1 – 5;10). The authors did not state that their goal was to establish the classification accuracy of this measure, but they sought to understand how the DLD and TD groups differed on this “method for identifying deficits in morphology, a critical element of identifying language disorder” (p. 616). The children in this study spoke Spanish or Spanish and English at home and received bilingual instruction in Head Start and had been doing so for at least a year. The reference standard was the identification of strong and weak learners with strong or weak communication skills (i.e., DLD was not specifically tested for) by teacher interview and classroom observation, respectively. Seven children were classified as weak learners and 14 as strong learners per the teacher interviews. The classroom observation classified a student with weak communication skills if they had one or no instances of initiating or responding to a verbal interaction with a peer during a 10-minute period of play; only four children in the sample were classified with weak communication skills. Only the teacher classification of strong/weak learners was used in their analyses.

Boyer and Martin (2012) used a within-participants design to teach the invented morpheme concept of a partial item in two different conditions to every child. The conditions were (1) teaching the invented morpheme /-ə/ affixed only to English platform words and (2) /-bei/ affixed to English words (e.g., “cat-bei”) and their translation equivalents in Spanish (e.g., “gato-bei”) in the English+Spanish condition. The conditions were counter-balanced for order of administration. The English condition was administered only by the trained graduate clinician while the English+Spanish condition was administered in English by the clinician and in Spanish by an interpreter. The pretest was the same expressive vocabulary check procedure as in Roseberry and Connell (1991) but with only 10 stimuli, and in the English+Spanish condition, the child was presented the instructions in both languages. These same words were used in the posttest. In that condition, a label in either language was accepted. During the teaching phase of 20 exemplars, the graduate clinician first told the child they would be learning a new rule, and then they had the child label the first picture in its whole form and then add a sound to that word when they saw the partial form. The clinician then modeled the word+correct invented morpheme. After that, each of the nineteen remaining items were presented as models without requests for or feedback for imitation. They used the script, “This is a boat. This is a boat. This is a boatbay. This is a boatbay.” Immediately following the teaching, the child was explicitly told to use the new rule on the posttest. The clinician labeled the whole form, “This is a sun.” Then they prompted, “Tell me what this is,” for the partial image. For the English+Spanish condition, the Spanish was presented by the interpreter for each item immediately after the clinician presented in English. Scoring was binary and was considered correct regardless of the language the child used for the platform word or morpheme.

Of the 29 participants, ten learned the rule to a predetermined criteria of 90% accuracy or higher in at least one condition (90% accuracy was to replicate the criteria set by Kohnert & Danahy, 2007), and these were the children who Boyer and Martin (2012) classified as having passed. However, a total of 16 children applied the rule some of the time (i.e., one or more correct) in at least one language. One child learned in both languages while the rest learned only one language's rule, four of whom applied it across both conditions (e.g., produced "sun-bei" in the English only and English+Spanish condition posttest). There was no difference by order or administration or by language condition, though there was a difference by age. Five-year-olds outperformed the three- and four-year-olds to a statistically significant degree. They reported a sensitivity of 71% and specificity of 57% for strong and weak learners who were four and five years old; the three-year-olds were excluded from this analysis because none of them learned the rule to 90% accuracy (though one did learn the rule to 80% accuracy in Spanish).

Although this study had the strength of presenting the stimuli in both languages, the poor reference standard was a significant weakness. Teacher report alone is not a gold standard for identifying DLD even at the time of Boyer and Martin (2012), and its classification accuracy is unknown. This rendered the results of this study incomparable to others. Additionally, the children had practice effects by hearing the stimuli with an invented morpheme applied to the platform word twice, and they saw the same words in the posttest condition twice (once in the English only condition, once in the English+Spanish condition). Another issue was the strong interference introduced by having two different phonological sequences for the same set of English words and for the same concept (a partial object). As the statistical learning literature has shown (e.g., Gómez & Lakusta, 2004), these can form counterexamples for each other, thereby reducing the likelihood of learning the invented morpheme. In fact, the number of children who did learn the morpheme was somewhat surprising given that 100% of the English stimuli had counterexamples. It may be that because the rule was presented explicitly at first and because the presentation of the stimuli did not result in optimal statistical learning conditions, the children were less engaged in statistical learning and more engaged in explicitly identifying the made-up suffixes.

The last study that based their design on Roseberry and Connell (1991) was by Kohnert and Danahy (2007) with 3;6 – 5;8 year-olds in Head Start (*n* = 20) who all heard Spanish at home and had English instruction with some Spanish in informal interactions at school. They compared one group who was only taught and tested in Spanish and the other in English. The pretest was an expressive vocabulary check of the 10 items that would be on the posttest. In the teaching phase, the children were taught 20 items with the partial-object morpheme (/-ə/ in English, /-be/ in Spanish) with the same labeling procedure as in Boyer and Martin (2012). The posttest used the same procedure and script as Roseberry and Connell (1991), presenting the child with the whole and partial images and having them label the partial form after being given the whole word label. Their criterion for passing was 90% accuracy or more (i.e., 9/10 or 10/10). Reasoning for this criterion was not provided. In the Spanish group, 7/10 children met the criterion, but only 1/10 children in the English group passed with 90% accuracy. However, four children in the English group had 10% accuracy or higher. A total of five children in the English-only group and two in the Spanish-only group did not get any correct. This study was an important replication because, even though there was no final classification accuracy to report, the results demonstrated a difference in performance by language of administration with a pattern of higher scores in the language of more experience. Additionally, not all TD children learned the invented morpheme rule in either language, further suggesting that the task was not optimal for identifying DLD given the false negatives it would produce. The participants in this study (*M* = 4;7) were about one year younger on average than in Roseberry and Connell (1991) (*M* = 5;7), which may have resulted in the poorer overall scores of these TD bilinguals, but Kohnert and Danahy (2007) reported no correlation between age and performance.

Finally, Hwa-Froelich and Matsuo (2005) assessed 21 Vietnamese-English speaking children ages 3;6 – 5;6 on a battery of assessments for TD bilinguals. The teaching phase consisted of teaching 10 pictures, presented once each with the script "This is a chair. This is a chair-po." The morpheme /-po/ was chosen to better resemble Vietnamese phonotactics, but the stimuli were administered in English. The posttest was first a receptive test, asking the children to identify the NOUN+po in an array of two pictures for 20 items and then an expressive test, asking them to label those same 20 partial items. This study was the only of the Roseberry and Connell (1991) replications to include modifiability, which consisted of three questions (child responsivity, examiner effort, child transfer based on Peña (1993) as cited by Hwa-Froelich & Matsuo, 2005) that were scored by the student clinician immediately after the posttests for all the assessments. On the receptive posttest, 18/21 children received a score of 18 or more and three received a score of 0. On the expressive posttest, 12/21 children scored at least one point with nine of those children receiving a score of 14 points or higher. They did find a statistically significant correlation with age ($r = .582$) and with their modifiability measure ($r = .467$), but they were also cautious about the modifiability measure because they noted a gender bias in favor of females. Again, not all the children learned the target morpheme, and they were all characterized as TD. This points to a potential issue with the teaching phase.

Theoretical Framework

The current study is a replication with modifications of Roseberry and Connell (1991) with the intent of improving the invented morpheme dynamic assessment. Despite the fact that not all TD bilingual children learned the invented morpheme in Roseberry and Connell (1991) or any of the replications, the target choice of a morpheme holds promise as an appropriate target for evaluating learning in children with DLD because children with DLD are expected to perform worse on morpheme-learning tasks than TD peers as difficulties with morphosyntax is the hallmark deficit of DLD (Bishop et al., 2016; Bishop et al., 2017). However, there was room for improvement on these invented morpheme dynamic assessments. The current study followed the format of a pre-test, a teaching phase, and then an immediate posttest, a one-hour posttest, and a one-day posttest. The pre-test had three parts: (1) checking that the morpheme was not guessable, (2) checking that the child had sufficient articulatory capability to produce the invented morpheme when affixed to a one- or two-syllable noun, and (3) checking that the child knew ten root words that would be used on the posttests. The child was taught the invented morpheme, /-əb/, meaning a partial object, and then tested on their generalization to a new set of stimuli in the posttest. In the rest of this section, I will highlight the major changes to the dynamic assessment before detailing the test creation process. See “Dynamic Assessment” in “Methods” for a full description of the dynamic assessment procedures.

While the original study reported a rationale for using modeling without imitation, the other parameters of the teaching phase were not based on a theory of learning. This lack of an empirically based teaching rationale may have been one cause for TD children not learning the invented morpheme. Only Boyer and Martin (2012) explicitly told the children that they would be learning a new rule, and they only explained the rule once. Most likely, the children were learning the pattern via statistical learning mechanisms because they implicitly learned the NOUN+morpheme pattern. Although the studies had methodological differences, each presented the invented morpheme in a context of high regularity but low variability, which is not ideal for implicit pattern learning via statistical learning mechanisms. The scripts revealed that the authors used the same platform utterance for each teaching dose (e.g., “This is a cat. This is a cat. This is a catə. This is a catə.”), so the most regular thing the child heard was “this is a ___ə.” The invented morpheme had less regularity than the full platform utterance, demoting its salience. Moreover, the most variable part of the utterance was the noun, making the child attend to it, rather than the target invented morpheme. Additionally, the lack of variability made it more likely that the children would learn the entire phrase as a chunk, “This is a ___ə,” rather than the invented morpheme itself. In contrast, the proposed study will have high regularity of the target morpheme and high variability around it. Finally, the dose of the teaching phase may need to be higher than just ten or twenty presentations.

Although children with DLD benefit from statistical learning, they perform more poorly on these and artificial grammar learning (e.g., invented morpheme) tasks than TD children (Ahufinger et al., 2021; West et al., 2021). Thus, statistical learning is hypothesized to provide a maximal distinction between children with DLD and typical language, making it a good theoretical basis for identifying DLD. In the current study, I made several key changes to the original design by Roseberry and Connell (1991). The most prominent change is that the teaching phase was changed to align with statistical learning principles. This was easiest to do through a scripted story to ensure fidelity across administrations. The child listened to the story and saw the partial objects presented as the assessor read about them in the script. To meet the regularity principle, the doses were presented frequently—48 doses—in a short time with no counterexamples. The first half of the teaching phase consisted of focused stimulation of 24 doses of the target morpheme. The second half of the teaching phase consisted of 24 doses of enhanced recasts, during which the child was asked to name the target morpheme after seeing and hearing the whole word in the story. The assessor then got the child's attention and said the correct response regardless of the child's answer. To meet the variability principle of statistical learning, every platform noun and surrounding utterance was unique. Moreover, the angle of the cut (top left or right, bottom left or right) and the style of cut (straight, large curve, multiple curves, zigzags) was randomized for every image. This ensured that the most consistently occurring features were the position of the target morpheme (i.e., NOUN+morpheme), the phonological sequence of the target morpheme, and the concept of partiality (not just a top right straight angle cut). Theoretically, this high regularity and high variability would enhance statistical learning to make it more likely that the TD children would learn the morpheme.

The other big change was the addition of a modifiability measure. Similar to Hwa-Froelich and Matsuo (2005), I borrowed from modifiability scales created by Peña (2000) and Peña et al. (2007). In most other iterations of this task, there was little or no interaction between the examiner and the child, but with the addition of enhanced recasts, the examiner got to engage the child. Thus, modifiability became a possibility to measure. This was important because Hunt et al. (2022) and Orellana et al. (2019) found that modifiability was one of the most informative outcomes of a dynamic assessment. Moreover, having this interaction and rating modifiability provided process information, which is critical to dynamic assessment. However, there could be bias in modifiability as it is a rating scale scored by the assessor rather than an objective measurement (e.g., Hwa-Froelich & Matsuo, 2005). I added in objective measures—time on the teaching phase and reaching a predetermined number of responses in a row correct—to supplement the subjective modifiability ratings.

Another big change was the posttest. No studies have systematically examined the impact of the transfer of morpheme knowledge from short-term memory to long-term memory (i.e., consolidation) on their diagnostic decisions. Dynamic assessments have ranged from 1 day (Kapantzoglou et al., 2012) to 12 weeks (Peña, 2000) between pre- and posttests without published justification, leaving SLPs to guess about optimal test timing. Consolidation after sleep plays a crucial role in word-learning (Davis & Gaskell, 2009), which is not dissimilar from derivational morpheme learning. Some studies have demonstrated that consolidation of learning may be impaired in children with DLD (Desmottes et al., 2017; Desmottes et al., 2016; Hedenius et al., 2011). In contrast, emergent literature suggested that encoding (forming the initial memory trace) rather than consolidation may be responsible for children with DLDs' difficulty with word learning (Alt, 2011; Gordon et al., 2021; Jackson et al., 2021). Thus, consolidation remains an open question for children with DLD. To assess whether consolidation changed diagnostic accuracy of the dynamic assessment, the posttest was conducted multiple times—immediately after teaching, at the end of the first day of assessments, and one day post-teaching. If consolidation is not a diagnostically determining factor, dynamic assessment may be effective over one day, saving SLPs valuable time. If it is diagnostically determinative, then SLPs can feel confident about having timing parameters for this dynamic assessment.

Finally, the dynamic assessment was created in English and in Spanish to be able to test each child in both their languages, thereby capturing their full linguistic abilities. Administering the dynamic assessment in both languages also reduced the possibility of assessment bias against a certain bilingual profile. For example, children with Spanish dominance would not be penalized for potentially weaker English because they could demonstrate their strengths in the Spanish version. The versions were designed to be administered independently, rather than an English sentence and a Spanish translation as was the procedure in Boyer and Martin (2012). Additionally, there were no counterexamples. There was only one version of the invented morpheme per language—/ -əb/ for English and /-ebo/ or /-eba/ (depending on grammatical gender of the platform word) for Spanish (see "Morpheme Choice" and The Unused Spanish Version of The Dynamic Assessment for more details). Because the evidence was mixed on whether monolingual English administration or bilingual Spanish-English administration was better, this study would have been able to add to the evidence for or against bilingual administration. Unfortunately, after testing the first six participants, only two produced enough Spanish (i.e., labeled ten vocabulary items on the pre-test) to be administered the Spanish version. Thus, the decision was made to keep the administration consistent across all children, using the same, English version for all participants to allow for comparison within the sample and to previous studies of the invented morpheme dynamic assessment.

Dynamic Assessment Creation

In the following sections, I will detail the decision-making process and rationale for the final choices made for the dynamic assessment itself. The first several subsections pertain to the entire dynamic assessment followed by subsections for decisions related to specific components of the dynamic assessment. This dynamic assessment went through several iterations since its initial creation in the Fall of 2023. Changes were made because of feedback, piloting with adults, and piloting with children. One of the biggest factors driving decisions was the need for a concise assessment, so almost every iteration resulted in a shorter version of the task.

Factors Guiding the Design. Throughout the entire design process, I was guided by the goal of implementing this dynamic assessment in a clinical setting. Before revamping the present assessment from its predecessor, I met informally with a group of SLPs from the Tucson Unified School District (TUSD) to understand their preferences for dynamic assessment (N. Fry, C. Jones, G. Nicolette, G. Santos, K. Worden, personal communication, April 20, 2022). The SLPs mentioned time as a crucial factor both for practicality of scheduling and to account for children's attention spans—thirty minutes or less was mentioned by one SLP. All stated that it would be best to complete a dynamic assessment in one day. Many SLPs also shared the need for something quickly useable without many manipulatives. They discussed how other dynamic assessments lacked clear instructions of what to do and how to implement each phase. Based on the need for brevity, structure, and simplicity, I chose a scripted storybook format (rather than a play format) and attempted to keep the dynamic assessment time to about 15 minutes.

For multiple reasons, the clinicians wanted concrete scores. The first was the need to report and interpret the dynamic assessment results to parents. The second was the need to mitigate bias. The SLPs brought up how their own cultural backgrounds may bias how they understand a child's development, and they did not want to make a clinical judgment solely based on how they perceived the child's performance. It is likely they were referring to the modifiability ratings being subjective as well as other currently available dynamic assessments not having concrete metrics for understanding the child's performance. Thus, the current dynamic assessment was designed to yield scores based on operationalized definitions.

I also kept the ease of use in mind; clinicians should be able to learn how to administer and score this dynamic assessment quickly. While we measured many variables that could be indicative of the child's performance on the dynamic assessment in the current study, I envisioned that only those most necessary for diagnostic accuracy would be recommended to clinicians for use.

Morpheme Choice. In the original iteration of the current dynamic assessment, Roseberry and Connell (1991) used the schwa /-ə/ to represent the concept of "part of." All subsequent replications used the same morpheme for English words. In the current iteration, the schwa was replaced with /-əb/ for the English morpheme because schwa could potentially be mistaken for an emphatic pronunciation of a word, which could have added error into the scoring. Additionally, /-əb/ has the VC syllable structure, which is common to word-final morphemes in English (Ebbers, 2010). The suffix /-əb/ was chosen because it did not approximate any currently existing free or bound morphemes in English and, given its status as an early 8 phoneme that is typically acquired by 2;0-2;11 (Crowe & McLeod, 2020; Shriberg, 1993), was likely easily pronounced by preschoolers who are still developing their articulatory abilities. See Table 2 for a list of the final consonants considered and reasons for their rejection.

Table 2

Considerations for the Target English Morpheme

Potential Morpheme	Reason for Rejecting
əp	Like the free morpheme "up"
ət	Like the bound morpheme for regular past tense "-ed"
əd	Like the bound morpheme for regular past tense "-ed"
ək	Children with phonological difficulties may be less likely to produce a velar consonant
əg	Like the bound morpheme for present progressive "-ing"
əm	May sound like filler "um"
ən	Like the free morpheme "an"
əŋ	Like the bound morpheme for present progressive "-ing"
əf	May be hard to hear; children with phonological difficulties may be less likely to produce a fricative

Potential Morpheme	Reason for Rejecting
əv	Children with phonological difficulties may be less likely to produce a fricative
əθ	Children with phonological difficulties may be less likely to produce a fricative
əð	Children with phonological difficulties may be less likely to produce a fricative
əs	Like the bound morpheme for plural "-s/-es"; children with phonological difficulties may be less likely to produce a fricative
əz	Like the bound morpheme for plural "-s/-es"; children with phonological difficulties may be less likely to produce a fricative
əʃ	Children with phonological difficulties may be less likely to produce a fricative
əʒ	Children with phonological difficulties may be less likely to produce a fricative
ət͡ʃ	Children with phonological difficulties may be less likely to produce an affricate
əd͡ʒ	Children with phonological difficulties may be less likely to produce an affricate
əh	Violates English phonological patterns (i.e., no final "-h")
ər	Like the bound derivational morpheme for "-er" (as in "one who Xs is a Xer"); preschool children are not expected to produce /r/
əj	Violates English phonological patterns (i.e., no final "-j")
əl	Children with phonological difficulties may be less likely to produce the liquid /l/

Word Choices. Words for the teaching and testing phases needed to be imageable, short, familiar words for preschoolers. Although some studies have shown that children learn morphemes best in statistical learning paradigms when the platform words are more challenging (Owen Van Horne et al., 2017), I chose to use familiar platform words on both the teaching phase and posttests because children only had one session to learn the target morpheme. Making an ideal teaching condition required me to balance creating an optimal statistical learning environment with ensuring the child could learn their job in the task, which was to affix the target morpheme onto a word, not to learn a new word. Thus, I opted for familiar words.

The English words were selected by generating a search on SCOPE: The South Carolina Psycholinguistic METAbase (Gao et al., 2022) using the search criteria listed in Table 3. After an initial pass, I deleted words that were not highly imageable (e.g., "day"). I also removed family member labels (e.g., "brother") and pronouns.

Table 3

Psycholinguistic Properties of Dynamic Assessment Teach and Posttest Stimuli

Psycholinguistic Characteristic	Value Range	Original Database
Word Frequency	4.00 – 7.62	Freq_SUBTLEXUS_Zipf: (Brysbaert & New, 2009)
Age of Acquisition	1.58 – 3.80	AoA_Kuper: (Kuperman et al., 2012)
Number of Syllables	1 - 2	NSyll: (Balota et al., 2007)
Number of Phonological Neighbors	3 – 75	Phon_N: (Balota et al., 2007)

Note. The SCOPE Metabase generates lists of words based on the determined search criteria by incorporating selected databases for linguistic features within the pre-determined ranges. The original databases are cited in this table.

Image Development. The images for all sections of the dynamic assessment were developed by an illustrator, Victoria Choi, with feedback from me. For the teaching phase, we prioritized clear, representative stimuli with few, if any, images on the page other than the target word (maximum of five other items, e.g., for "breadəb", there was the bread (1), on a plate (2), and a dog (3)). The reason non-target images were included with the targets was to help the child follow the story and maintain interest in the task. Ms. Choi also developed some visual scenes for the beginning, middle, and end of the stories to facilitate the child's understanding about the narrative. Pages that were solely for storytelling did not have an item limit. Only the necessary stimuli were on the pre-teaching checks and posttest.

To maximize the salience of the partiality for the target morpheme images, we cut a large chunk of the image and ensured the remaining image had a strong contrasting color with the background. The cut angle (top right, bottom right, top left, bottom left) and type (straight cut, multiple curves, single curve, jagged) were randomized within the teaching and posttest phases to have high variability to facilitate statistical learning of the target structure (Plante & Gómez, 2018). See Figure 2 for example visual stimuli.

Randomization. I have mentioned the randomization of test elements in their respective sections. For all randomizations, I used Random.org (Haahr, n.d.). Table 4 has an exhaustive list of elements that were randomized.

Table 4

Elements Randomized in the Procedures and in the Dynamic Assessment

Test	**Randomized Elements**	**Scope**
Administration		
Order of Language Assessment Administration	BESOS subtests, language samples (one per language)	Across entire study population
Dynamic Assessment		
Vocabulary Check/Posttest	Word list	Entire stimuli list
Speech Screen/Guessability Check	Word list	Entire stimuli list
Guessability Check	Visual stimuli position on page	Each page for entire stimuli list
Guessability Check	Visual stimuli cut location & type (target morpheme), number of objects (plural), curve axis (curved objects)	Each page for entire stimuli list
Teaching: Story	Word list	Entire stimuli list
Teaching: Story	Reason for object partiality	Entire stimuli list[a]

Test	Randomized Elements	Scope
Teaching: Story	Visual stimuli cut location and type	Entire stimuli list

[a]Some objects were originally considered “hidden,” but I excluded that as an allowable cause for the partiality. “Hidden” causes were replaced with other causalities in a non-randomized manner.

Language of Administration. Although the dynamic assessment was originally intended to be administered once in English and once in Spanish across two different scripted narratives, early participants did not demonstrate sufficient Spanish usage to participate in both languages. Therefore, the dynamic assessment was administered in English only to every child.[2] Children were allowed to codeswitch on anything with the exception of the target word and morpheme. Allowing codeswitching but not requiring it reflected the administration parameters of the reference standard (Peña et al., 2008, in development), and it allowed children to take advantage of their full linguistic repertoire (Kapantzoglou et al., 2021; Yow et al., 2017).

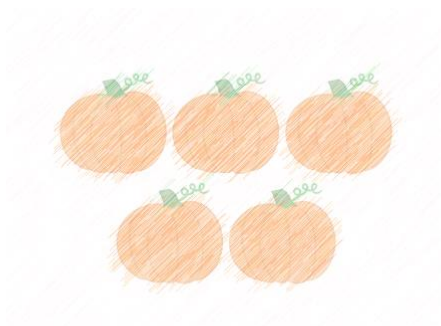
Pre-Teaching Checks. The three distractor items on the guessability check were each chosen to share at least one aspect with the target morpheme; see Table 5 for the list of specific distractor items. The criteria for each distractor were that it had to be imageable and visually salient so that a child might be similarly attracted to choose it as they would be to choose a partial object. It had to be a noun since the target morpheme was a noun, and it had to be a naturally occurring concept in both Spanish and English. For the inflectional morpheme distractor, the plural form was chosen because it was the only inflectional morpheme for a noun that occurred across Spanish and English. For the derivational morpheme, ‘micro-/mini-’ was chosen because it was one of the few easily imageable derivational noun morphemes that had an

[2] One child had participated in the Spanish dynamic assessment originally and then was administered the English version a week later.

equivalent derivational morpheme in English and in Spanish. For the visual distractor morpheme, adding a curve to the image was chosen because it was highly visually salient. In addition, in the real world, a curve in an object could be the result of damage, and being damaged (specifically, broken) was also a potential cause to make a partial object, which was the target morpheme. In addition, a curved object could be curved in any of the four sides of the object, which paralleled the variability of how the partial morpheme was presented.

Table 5

Guessability Check Target and Distractor Items

Item Type	Meaning	Natural Equivalent	Justification	Example Image
Whole image	N/A	N/A	Comparison target and object to be labeled for the speech screen	
Target morpheme	Part of	Broken or cut objects (like hemi-/demi-)	Target morpheme	
Distractor: inflectional morpheme	More than one	-s (plural)	Visually salient, naturally occurring in Spanish and English, imageable inflectional	

Item Type	Meaning	Natural Equivalent	Justification	Example Image
			noun morpheme	
Distractor: derivational morpheme	Very small	Micro- or mini-, -ito/ico/illo (Spanish)	Visually salient, naturally occurring, imageable, noun	
Distractor: visual distractor	N/A	Curved or warped objects (no naturally occurring similar morpheme)	Visually salient, a visual alteration that could occur naturally (like breaking or cutting an object), can occur in any of 4 directions like the target morpheme, imageable, noun	

Randomization of the guessability check stimuli was generated using the random list and random integer generators from Random.org. The order of presentation of the items and distractor images' page positioning (i.e., position 1, 2, 3, or 4) were randomized. In addition, the angle and type of partial cut were randomized as was the side of the curve for the curved distractor items. The number of plural objects was randomly generated to be between two and ten objects. See Table 4 for a list of all the elements randomized in the dynamic assessment.

In addition, a cutoff score needed to be decided for what would be considered an above-chance probability of guessing the target morpheme. There were ten items on the guessability check, so there was a 14.6% probability that a child would guess exactly four items by chance and a 5.8% probability that a child would guess exactly five items by chance (calculated by Berman (n.d.); see Table 6 for a list of the probability by number of successes). The cutoff for correct guesses was no more than four items to be able to continue with the dynamic assessment. This threshold was chosen for multiple reasons. The first was that if the child guessed the same area of the quadrant for every response, they would receive up to three correct responses. The probability of getting more than three correct responses, though, was 22.4%, which was relatively high. The probability of getting more than four correct responses dropped to 7.8%. Thus, if a child correctly guessed five or more items correctly, the likelihood of that response being by chance alone was low enough to suggest that some factor other than chance was driving their response behavior. If a child guessed five or more correctly, they were withdrawn from the study.

Table 6

Probability of Guessing the Target Morpheme by Chance for 10 Trials

Number of successes (r)	Binomial probability of exactly r number of successes (P(x) = r)	Cumulative probability of getting FEWER THAN r successes (P(x) < r)	Cumulative probability of getting AT MOST r successes (P(x) ≤ r)	Cumulative probability of getting MORE THAN r successes (P(x) > r)	Cumulative probability of getting AT LEAST r successes (P(x) ≥ r)
0	.056	0	.056	.944	1.00
1	.188	.056	.244	.756	.944
2	.282	.244	.526	.474	.756
3	.250	.526	.776	.224	.474
4*	.146	.776	.922	.078	.224
5	.058	.922	.980	.020	.078
6	.016	.980	.996	.004	.020
7	.003	.997	1	0	.003
8	0	1	1	0	0
9	0	1	1	0	0
10	0	1	1	0	0

Note. The calculator (Berman, n.d.) only calculated up to three decimals. The cumulative probabilities are additive or subtractive. For instance, the cumulative probability of getting AT MOST 2 correct guesses (.526) is the probability of 0 correct (.056) + probability of 1 correct (.188) + probability of 2 correct guesses (.282).

*The cutoff score was decided to be 4 or fewer correct guesses.

The vocabulary check went through multiple iterations. It initially included all the words that the child would be asked to say across the teaching and posttest phases, but pilot testing revealed that it was burdensomely long, so only the posttest vocabulary was checked.

Focused Stimulation and Enhanced Conversational Recasting (Teaching Phase). In my prospectus document, I proposed to include 69 doses of the morpheme through focused stimulation to align with the requirement of 67-68 doses for Vocabulary Acquisition and Usage for Late Talkers (VAULT), an efficacious word-learning treatment for late talkers (Alt et al., 2021; Alt et al., 2020). Time was of the essence, so I reduced the focused stimulation dose number to 24. Additionally, I had proposed to provide two doses of the partial morpheme in a focused stimulation manner and then have the third be an enhanced conversational recast, totaling in 23 recasts, which aligned with the enhanced conversational recast dose target of 24 in studies by Plante et al. (Meyers-Denman & Plante, 2016; Plante et al., 2019; Plante et al., 2014). To reduce children's anxiety at having to produce an unfamiliar morpheme after only hearing it twice, though, I changed the dose structure so that the child would be exposed to the invented morpheme more before being required to respond.

In the final task, the child heard 24 doses of focused stimulation and then was required to respond to 24 doses of enhanced conversational recasts. Given that Plante et al. (2018) found a small effect for focused stimulation (called auditory bombardment) after enhanced conversational recasting, I had considered reversing the order to make focused stimulation last. However, I kept focused stimulation first because the child had never been exposed to this invented morpheme prior to this assessment. It is unknown whether moving focused stimulation second would have a similar benefit as in Plante et al. (2018) because that study used real morphemes to which the child had ostensibly been exposed for years in daily life. Thus, focused stimulation remained first to set children up for success for the enhanced conversational recasts.

All doses were provided with high regularity of the target and high variability around the target to maximize the likelihood of TD children learning the target morpheme (Plante & Gómez, 2018). The target morpheme was presented in 48 doses total (24 focused stimulation, 24 enhanced recasts) and did not have counterexamples or irregular presentations to ensure the regularity of the morpheme presentation was high (Plante & Gómez, 2018). For each dose, the

NOUN+əb exemplar was either at the end of a sentence or was followed by a new phrase (e.g., prepositional phrase), creating the peaks and valleys of transitional phrase probabilities that Thompson and Newport (2007) found to be crucial to the learning of syntax. Furthermore, having one invented grammatical structure (i.e., the artificial morpheme "/-əb/") in a real language may be expected to facilitate learning beyond artificial grammar paradigms because of the many complexities in natural language, and Gómez (2002) demonstrated that variability is critical to effective statistical learning. Moreover, each platform word and utterance was unique to further ensure the variability principle was met (Plante & Gómez, 2018).

Children in the current study were also adding semantic information to the target structure. They regularly saw the broken object with few, if any, competing visual exemplars (see "Image Development"), while hearing the NOUN+əb construction within an utterance. This regular presentation of the target morpheme /-əb/ with high variability of the visual and linguistic input around /-əb/ directed the learner to pick up on the most regularly occurring features of the input (Aslin & Newport, 2012; Plante & Gómez, 2018)—the phonological sequence /-əb/ at the end of nouns that co-occurred with a partial object. Thus, the children in this study were not only statistically learning the sequential dependency of NOUN+əb, but they also were engaging in mapping semantic meaning onto /-əb/. This is important because not every noun had an invented morpheme affix, so children had to rely upon the image (and linguistic information from the story) to understand when the morpheme /-əb/ should be applied.

Posttests. While the original studies of this dynamic assessment only administered one posttest immediately after the teaching phase (Boyer & Martin, 2012; Hwa-Froelich & Matsuo, 2005; Kohnert & Danahy, 2007; Restrepo, 1998; Roseberry & Connell, 1991), the current study had two extra posttests to see if their inclusion increased diagnostic accuracy of the dynamic assessment as some emerging evidence has suggested it might (Hunt et al., 2022). The one-hour interval/end-of-session was chosen because it allowed for time for consolidation, and the previous literature has not clearly reported the timing of a delayed posttest (Camilleri & Law, 2007; Hasson et al., 2012; Maragkaki & Hessels, 2016). The one-day interval was chosen because of the role of sleep in consolidation (Davis & Gaskell, 2009).

The Current Study

The purpose of the current study was to examine the feasibility of the proposed dynamic assessment in terms of its replicability and classification accuracy. The clinical role of this dynamic assessment would be to identify DLD with the objective of that identification leading to the child receiving SLP services (STARD #3). Akin to the levels of intervention research proposed by Fey and Finestack (2009), this study was intended to be at the early stages of evaluating the proposed dynamic assessment with the knowledge that its conclusions would not be broadly generalizable. Rather, this was a first step that would potentially inform a line of studies that could increase the confidence in the findings and advance the generalizability of the results.

Research Aims

The current study will address the following aims (STARD #4):

1. To create a replicable dynamic assessment procedure for evaluating the expressive morphology of Spanish-English speaking children ages 4;0 – 5;11 suspected of having DLD.
    a. **Hypothesis**: I predict that the procedures will be replicable given their scripted nature.

2. To determine which elements of administration (modifiability scoring, language of administration, posttests at varying time intervals) yield the best differentiation between children with DLD and typical language.
    a. **Hypothesis**: Given that this dynamic assessment is introducing a more cognitively appropriate teaching method than past iterations (e.g., Roseberry & Connell, 1991), I predict that the dynamic assessment will reach fair to good classification accuracy per the standards by Plante and Vance (1994).
        i. **Modifiability**: Given that modifiability is one of the most diagnostically informative factors in other dynamic assessments (Hunt et al., 2022; Orellana et al., 2019), I predict that it will also be diagnostically informative in this dynamic assessment.
        ii. **Consolidation**: Given opposing findings regarding consolidation for children with DLD, it is unclear whether consolidation will be a diagnostically determining variable.

Originally, the second part of the hypothesis for aim 2 regarded language of administration. Given that Kohnert and Danahy (2007), which was more methodologically rigorous than the other Spanish-English presentation of the invented morpheme by Boyer and Martin (2012), demonstrated a language effect on a dynamic assessment protocol similar to the current study, I predict that there will be an effect of language on diagnostic accuracy of this assessment. It is unclear whether using the best score of one language or the combined scores of both languages will be more diagnostically accurate. Unfortunately, due to the sample not speaking enough Spanish to run the dynamic assessment in both languages, this hypothesis was not able to be evaluated in the current study. Only the English version of the dynamic assessment was used for all participants and will the focus of the rest of this dissertation manuscript.

Method

Study Design

This study was approved by the University of Arizona's IRB. This was a prospective study in which the data collection procedures were planned before the index test and reference standard data were collected (STARD #5). The goal of the study was to determine if this dynamic assessment would be diagnostically accurate, and if so, to ascertain which pieces of data from the dynamic assessment were most informative about the child's language status as being DLD or TD. As a feasibility study, a case-control design was deemed appropriate and was operationalized as testing one case (i.e., participant) per profile (see "Participant Profiles") and at least one replication (i.e., another participant of that profile).

Data were collected from children with DLD and TD across a range of bilingual profiles to compare the dynamic assessment's accuracy for the two groups. Their language was measured by the reference standard, which is the pre-existing assessment tool with acceptable sensitivity and specificity that determines the child's status of DLD or TD. The reference standard was the Bilingual English-Spanish Oral Screener (BESOS; Peña et al., 2008, in development). The index test, which is the assessment whose classification accuracy is being studied, was the dynamic assessment. It included a set of pre-teaching eligibility checks, a teaching phase, and three posttests (immediate, 1 hour post, 1 day post). The target skill being taught in the dynamic assessment was learning the invented English derivational morpheme, /-əb/, which meant a partial object. Modifiability was measured in response to teaching. Children also provided narrative language samples to further characterize the sample. Receiver Operator Characteristic (ROC) curves compared children's identification as DLD or TD by the index test (dynamic assessment) and the reference standard (BESOS) to yield a classification accuracy of the index test.

The following sections detail the participants, measures, procedures, and analysis plan, addressing the STARD criteria 6 - 18.

Participants

Eligibility Criteria

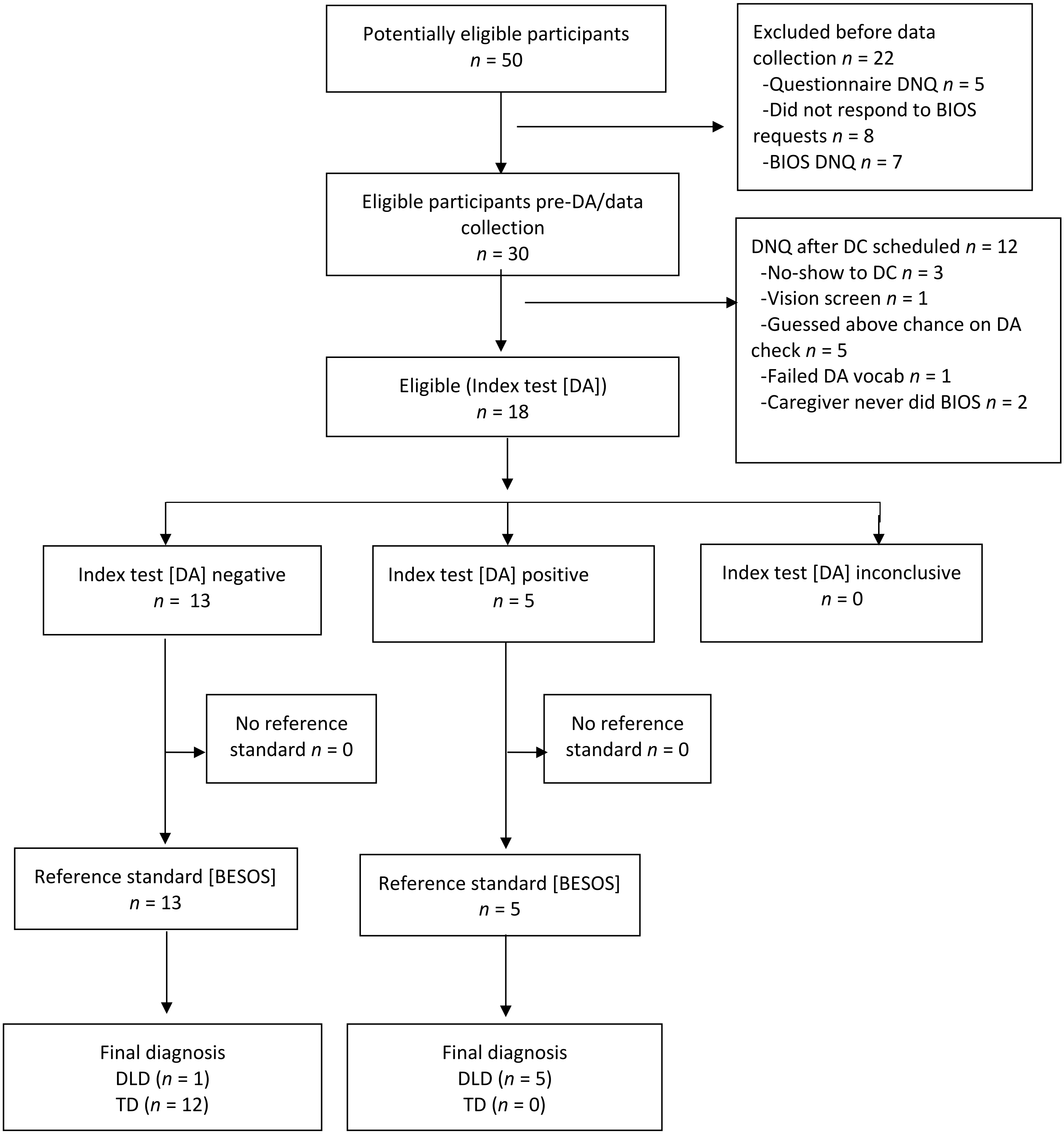
Eligible participants were children between 4 years, 0 months and 5 years, 11 months of age who used Spanish and English regularly in daily life at the time of participation (i.e., at least 21% Spanish use and 79% English use or vice versa). Children were excluded if they did not meet the inclusionary criteria or if they had a diagnosis of neurological, cognitive, motor, or behavioral problems that would preclude a DLD diagnosis (Bishop et al., 2017) or participation in the assessment without accommodation (e.g., Down Syndrome, Developmental Delay, AAC user). Children of all races/ethnicities and gender identities were welcome to participate (STARD #6). Initial eligibility was determined by caregiver report on a caregiver questionnaire, and continued eligibility was dependent on passing the eligibility screenings during the first day of data collection (STARD #7).

Children who were initially eligible could be withdrawn for any of the following reasons: (1) the caregiver did not complete the caregiver questionnaire or language use survey (Bilingual Input-Output Survey; BIOS), (2) the child failed the hearing, vision, or cognitive screening, (3) on the dynamic assessment pre-teaching checks, the child was unable to name 10 pictures, failed the speech screen, or guessed the invented morpheme above chance level, (4) the child missed a data collection session and there was difficulty rescheduling, or (5) the child engaged in behavior that stopped them from being able to do the tasks in time.

A total of 50 families signed and submitted consent (an additional 5 were signed but not officially submitted via REDCap). Five participants were not eligible based on their caregiver questionnaire responses. Eight families did not respond to contact attempts for the language use survey. A total of seven families were rendered ineligible due to their child being functionally monolingual on the language use survey (the BIOS; see “Participant Profiles”). Three families were scheduled but did not attend the data collection. Nine children started data collection but were found to be ineligible at the beginning of data collection due to not passing the vision screen (*n* = 1), guessing too many items on the dynamic assessment pre-teaching guessability check (*n* = 5), and not being able to label ten words in English on the dynamic assessment vocabulary check (*n* = 1). Of the 27 children who began data collection, 18 were eligible and completed data collection. See Figure 1.

Figure 1

Standard for Reporting Diagnostic Accuracy (STARD) Flow Chart of Participants (STARD #19)



Note. DA = dynamic assessment, DNQ = did not qualify, DC = data collection, BIOS = Bilingual Input-Output Survey, BESOS = Bilingual English-Spanish Oral Screener, DLD = Developmental Language Disorder, TD = Typical Language Development

Recruiting

The study population came from a convenience sample of participants (STARD #9) who were recruited from October 2023 through April 2024 (STARD #8). Recruitment methods varied. Online, I posted social media ads (Facebook and Instagram) and a description of the study on the L4 Lab website. In person, I spoke with parents during pick-up times to recruit at private and public daycares and preschools, particularly in zip codes with high percentages of Spanish-speakers. Through the largest local school district in Tucson, AZ, I received permission from principals to distribute packets through students' backpacks. Packets included a cover letter, a consent form, and a caregiver questionnaire, all of which were in Spanish and English. Caregivers could return signed packets through the children's backpacks for the school to collect or could go to the REDCap website to sign up. We also received permission from local entities to post or pass out flyers (e.g., local libraries, clinics, Casa de los Niños) and make announcements (e.g., bilingual story time at local libraries). Recruiting efforts were targeted at Latine-serving institutions and highly Spanish-speaking areas around Tucson, Arizona.

Participant Profiles

Participants were recruited until the desired number of participants were found that represented the target profiles (see Table 7 for a list of participant profiles). To determine a target sample size, I calculated a power analysis via G*Power version 3.1 (Faul et al., 2007) for a matched pairs two-tailed t test with equal group sizes using the Roseberry and Connell (1991) data. It revealed that I would need 12 participants total to achieve 95.5% actual power to detect a between-group difference. I aimed to recruit at least two participants in each of eight profiles, yielding a total N of 16, yielding adequate power (STARD #18). The profiles included the variables of language status (DLD or TD), current language use, and, secondarily, language ability.

Language status (DLD or TD) was determined by the reference standard, the BESOS. Language status was included because it was the primary variable used to evaluate the classification accuracy of the dynamic assessment. Current language use was characterized as balanced or unbalanced across Spanish and English and was included because bilingual language use may influence dominance and therefore performance on the dynamic assessment (Peña et al., 2023). It would be important to know if one group performed differentially due to their language use profile, so we could make appropriate recommendations for future use of this dynamic assessment. A language use survey (the Bilingual Input-Output Survey; BIOS) was given to each child's caregiver and teacher and calculated their percentage use from hour-by-hour estimates of Spanish and English input and output. Based on cut-offs used by Bedore et al. (2012), balanced use was classified as current Spanish and English use each between 40% and 60%. Unbalanced use was considered 21-39% use in one language and 61-79% use in the other language. Children were considered functionally monolingual at use rates of 0-20% in one language and 80-100% in the other; these children were not eligible to participate. I recruited for the unbalanced category, not for unbalanced—higher Spanish use separately from unbalanced—higher English use.

The third variable, language ability, was characterized as low, medium, or high based on a continuous measure, the narrative language samples (see "Language Samples" in "Language Measures Administration"). Language ability as a continuous measure was included to help avoid spectrum bias, a case of only sampling the extreme ends of the language ability spectrum, which can lead to over-confidence in a measure's classification accuracy (Ransohoff & Feinstein, 1978). Each child's language sample was coded and compared to a previously collected sample's distribution from Hernandez et al. (2024) to determine low (< -1 SD), medium (-1SD – 1 SD), and high language (> 1 SD) ability. Children at the low end or high end of the spectrum were also assumed to be characterized as DLD or TD, respectively. Children in the middle of the spectrum were those who are the hardest to identify because their status of DLD or TD was less clear cut as their language scores were below average but not extremely so. Thus, in this case, "medium" language ability did not mean average, but rather it was the middle of the spectrum that I was trying to capture in this diagnostic accuracy study.

Profiles were:

1. DLD with unbalanced use, low language ability
2. DLD with unbalanced use, medium language ability
3. DLD with balanced use, low language ability
4. DLD with balanced use, medium language ability
5. TD with unbalanced use, medium language
6. TD with unbalanced use, high language
7. TD with balanced use, medium language
8. TD with balanced use, high language

Table 7

Participant Profile Categories

Language Use	**Language Ability**			
	Low	Medium		High
Unbalanced	1	2	5	6
Balanced	3	4	7	8

Note. The shaded boxes represent children with DLD, and the unshaded boxes represent children with TD.

All eligible children who completed all tasks were retained for analyses regardless of profile. Language ability was recruited for after considering language status and current use. Participants were recruited to cover the language spectrum to the extent feasible. Thus, the main four profiles of focus for recruitment were for language status (DLD or TD) and language use (balanced or unbalanced). Recruitment continued until each of the four main profiles (i.e., DLD-balanced, DLD-unbalanced, TD-balanced, TD-unbalanced) had the minimum of two participants.

Participant Characteristics (STARD #20)

A total of 12 children were identified as TD and 6 identified as DLD by the reference standard. There were 11 girls and 7 boys with a mean age of 4;7 (years;months, *SD* = 6.37 months, range = 4;0 – 5;7). All participants (*N* = 18) were Hispanic/Latine as reported by their caregivers, and three children were also reported to be described by other racial/ethnic identifiers (*n* = 1 American Indian/Alaskan Native, *n* = 2 White). Without a standard way to ask demographic questions, I used a combination of precedence from other researchers and the literature. For gender, I followed precedence from the POWER project and honored the gender spectrum by having three options: boy, girl, and prefer to self-describe (C. Figueroa, personal communication, June 13, 2023). For race/ethnicity, Dowling and Hugo Lopez (2022) reported that, particularly for Hispanic/Latine people, the separation of ethnicity (Hispanic/Latino or not) and race (e.g., White, Black, Asian, American Indiana/Alaska Native) did not match self-perceptions, so I combined race and ethnicity in the format their research suggested was most apt. See Table 8 for more details on participants' demographic information.

Table 8

Participant Demographics

Demographic Categories	Balanced Spanish-English use				Unbalanced Spanish-English use			
	TD		DLD		TD		DLD	
	n	**%**	***n***	**%**	***n***	**%**	***n***	**%**
Count	4[a]		4		8		2	
Age[b]	4;7 (6.18)		4;6 (5.92)		4;9 (6.72)		4;8 (11.31)	
Gender								
Girl	3	75	2	50	4	50	2	100
Boy	1	25	2	50	4	50	0	0
Ethnicity[c]								
Latino/a/e	3	75	4	100	6	75	2	100
Multiracial	1 Latine-White	25	0	0	1 Latine-American Indian/Alaskan Native, 1 Latine-White	25	0	0
Maternal Education[d]								
8th grade or less	0	0	0	0	1	12.5	0	0
8th-12th grade	0	0	1	25	0	0	0	0
HS diploma	1	25	0	0	1	12.5	0	0
Technical school or other further education	0	0	2	50	2	25	0	0
Undergraduate college (1 year or more)	1	25	1	25	0	0	2	100
Undergraduate degree	0	0	0	0	3	37.5	0	0

Demographic Categories	Balanced Spanish-English use				Unbalanced Spanish-English use			
	TD		DLD		TD		DLD	
	n	**%**	***n***	**%**	***n***	**%**	***n***	**%**
Graduate school (1 year or more)	2	50	0	0	1	12.5	0	0
Prefer not to respond	0	0	0	0	0	0	0	0
Paternal Education[d]								
8th grade or less	0	0	0	0	0	0	0	0
8th-12th grade	0	0	2	50	0	0	0	0
HS diploma	0	0	0	0	1	12.5	0	0
Technical school or other further education	1	25	0	0	1	12.5	1	50
Undergraduate college (1 year or more)	0	0	1	25	2	25	0	0
Undergraduate degree	1	25	0	0	2	25	0	0
Graduate school (1 year or more)	1	25	0	0	2	25	1	50
Prefer not to respond	1	25	1	25	0	0	0	0
Co-occurring ADHD[e]	0	0	1	25	0	0	0	0
SLP services[f]	0	0	1	25	1	12.5	0	0
Language ability[g]								
Low	0	0	1	25	1	12.5	0	0
Medium	4	100	3	75	7	87.5	2	100

[a]One child in the Balanced TD group had 39.5% Spanish use and 60.5% English use. They were categorized as balanced rather than unbalanced given their proximity to the cutoff.

[b]Age at the time of the first day of data collection. The mean and standard deviations are presented as M (SD) in year;month (months) format.

[c]Participants could check all races/ethnicities that applied. All participants indicated that their child was Latine, so other categories listed indicate that the child was multiracial. Other race/ethnicity categories were listed and were not checked by any respondents: Black, Asian, Middle Eastern/North African, Hawaiian/Pacific Islander, self-describe.

[d]Education reported as each caregiver's highest level of education.

[e]ADHD was identified by caregiver self-report. Some children may have had undiagnosed ADHD given that the minimum recommended age for diagnosis is four years old (Subcommittee on Attention-Deficit/Hyperactivity et al., 2011).

[f]SLP services refers to children who were ever in speech-language therapy services in any setting (e.g., schools, private practice).

[g]Language ability was determined by the average of the *z*-scores of the child's best language MLU-W and their combined-language PGU scores. The mean and standard deviation of the TD group from Hernandez et al. (2024) were used to create the *z*-scores. Categories were low (< -1 SD), medium (-1 – 1 SD), and high (> 1 SD), but no children scored in the "high" range. This fulfills STARD #21a; there were no alternative diagnoses (STARD #21b).

Participants' Bilingual Profiles. There is no current agreed-upon standard for reporting bilingual profiles. The choice lies with the researcher. My choice was based upon (1) the availability of evidence for variables that were relevant to the current study, (2) the availability of reliable tools to measure the variables, and (3) the need to balance informativeness with respect for the family's privacy. I will report on age of first exposure, current use, proficiency, language dominance, dialects, and community context.

All children heard Spanish from birth, and the average age of first exposure to English was 0;10 (years;months, *SD* = 1;4). For current use, seven children had balanced Spanish-English use while eleven children had unbalanced use, four using more Spanish and seven using more English. Language proficiency was the child's standard score on the BESOS Semantics and Morphosyntax subtests; see Table 9 for mean scores. Language dominance was determined by the child's performance on the reference standard (the BESOS). Children with mixed dominance (*n* = 5, 27.78%) performed better on the Semantics subtest in one language and in Morphosyntax in the other language. All but one child with mixed dominance were also classified as balanced bilinguals. Children with congruent dominance (*n* = 13, 72.22%) performed best on Semantics and Morphosyntax in the same language. Eleven families responded about their dialect use, which was asked after data collection. All children of the families who responded heard Mexican Spanish, and one child also heard Puerto Rican Spanish. For English, 90.91% (*n* = 10) heard General American English while 9.09% (*n* = 1) heard Southern American English.

The following is a brief description of the community context of the current study, which took place in Tucson, AZ and surrounding areas. Tucson is made up of 44.8% Hispanic/Latino residents (United States Census Bureau, n.d.), and an estimated one-third of Tucson residents speak Spanish (Mora & Lopez, 2023). Anecdotally, many families regularly visited Mexico or were visited by family members who traveled back and forth from the US and Mexico. Some of the children in this sample attended public schools in Arizona while others were in private daycares or preschools, Head Start programs, or not yet attending school. Thus, some were impacted by Arizona state law's English-only schooling policies but not all. Even in the public schools, some teachers told me that they and/or peers of the child spoke Spanish with them in the classroom. Additionally, Tucson hosts several events and has numerous stores and restaurants that highlight local Sonoran, Southwestern, and Mexican culture (e.g., Tucson Meet Yourself, Tucson International Mariachi Conference, UNESCO City of Gastronomy). See "Characterizing Bilingualism" in the introduction for more information on the state and federal community context.

Table 9

Bilingual Profiles

Variables	**Balanced Spanish-English Use**				**Unbalanced Spanish-English Use**			
	TD		**DLD**		**TD**		**DLD**	
	M	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***
Age of first exposure[a]								
English	0;9	1;6	0;9	0;11	0;11	1;8	1;0	1;5
Current use[b]								
Spanish	47.58	8.34	48.63	7.25	70.00 (Sp)	4.24	62.50 (Sp)	0.71
					33.91 (Eng)	5.03		
English	52.42	8.34	51.38	7.25	30.00 (Sp)	4.24	37.50 (Sp)	0.71
					66.09 (Eng)	5.03		
Proficiency[c]								
Spanish Morphosyntax	91.25	10.24	70.75	7.63	72.75	11.41	74.00	7.07
English Morphosyntax	110.25	8.88	76.75	3.50	90.50	11.45	76.50	2.12
Spanish Semantics	99.50	9.04	72.00	12.88	84.25	21.84	72.00	14.14
English Semantics	96.75	9.18	71.25	11.24	90.75	10.85	79.50	7.78
MLU-W$_{Sp}$	6.04	1.98	4.73	N/A	5.24	1.34	4.42	0.44
MLU-W$_{Sp}$ No CS	6.54	2.24	4.70	N/A	5.42	1.87	4.16	N/A
MLU-W$_{Eng}$	6.46	0.43	4.73	0.91	5.79	0.97	4.53	1.36

Variables	Balanced Spanish-English Use				Unbalanced Spanish-English Use			
	TD		DLD		TD		DLD	
MLU-W$_{Eng}$ No CS	6.06	0.58	4.67	1.00	5.25	1.19	2.96	0.53
	n	**%**	***n***	**%**	***n***	**%**	***n***	**%**
Dominance								
Mixed	2	50	2	50	0	0	1	50
Congruent	2	50	2	50	8	100	1	50
Dialects [d]								
Spanish								
Mexican Spanish	3	75	1	25	5	62.5	1	50
Multiple Spanish dialects	0	0	0	0	1	12.5	0	0
No Response	1	25	3	75	2	25	1	50
English								
General American English	2	50	1	25	6	75	1	50
Southern American English	1	25	0	0	0	0	0	0
No Response	1	25	3	75	2	25	1	50

[a]Age of first exposure listed in years;months. All children in the sample were exposed to Spanish from birth.

[b]Current use was based on the combined input and output across the home and school BIOS data if available and applicable. If the child did not go to school, their home BIOS was used. For the unbalanced groups, (Sp) indicates the children who had more Spanish use and (Eng) indicates the children who had more English use. The unbalanced DLD group only had children who used more Spanish than English.

[c]Proficiency scores were from the BESOS. Those that were lower than the lowest standard score (e.g., raw score of 0 yields a standard score of < 63 on English Semantics) were substituted with

one standard score point lower than the minimum score (i.e., < 63 became 62) to include them in the averages. The MLU-W_{Sp} and MLU-W_{Eng} refer to Mean Length of Utterance-$Word_{(language)}$ and came from the narrative retell language samples. These values include codeswitched utterances; MLU-$W_{Sp/Eng}$ No CS are for the Spanish and English MLU-W values excluding codeswitched utterances, respectively. Some children did not have language samples in both languages, the averages were for the available data.

[d]Dialect was asked after data collection, so there was some missing data, coded as "no response." The one child who heard multiple Spanish dialects heard Mexican Spanish and Puerto Rican Spanish.

Measures

In the following section, I will provide the description and rationale for each measure. See Table 10 for a list of each measure and its purpose. The dynamic assessment is elaborated in its own section after "Measures."

Table 10

List of Measures

Measure name	Purpose	Reference[a]
Caregiver		
Caregiver questionnaire	Initial eligibility screen Demographic information	N/A
Bilingual Input Output Survey (BIOS)	Initial eligibility screen Bilingual profile description	Peña et al. (2018)
Child		
Hearing screening	Ensuring sufficient hearing acuity to participate	American Academy of Audiology (2011)
Vision screening	Ensuring sufficient visual acuity to participate	Hyvarinen et al. (1980)

Measure name	Purpose	Reference[a]
Primary Test of Nonverbal Intelligence (PTONI)	Ruling out cognitive impairment due to incompatibility with DLD	Ehrler and McGhee (2008)
Dynamic assessment	Index test	N/A
Bilingual English Spanish Oral Screener (BESOS)	Reference standard—characterizing child as DLD or TD	Peña et al. (2008, in development)
Language sample	Characterizing child's language ability on a continuous measure Part of bilingual profile	Hernandez et al. (2024); SALT Software LLC (n.d.-a)

[a]Reference for the assessment or protocol.

Eligibility Assessments

There were two timepoints at which eligibility was screened—before data collection and the first day of data collection. Before data collection, the caregiver questionnaire and BIOS results determined if a child was initially eligible to start data collection. If so, on the first day of data collection, the child's eligibility was screened through the hearing screening, vision screening, the nonverbal cognitive screener (Primary Test of Nonverbal Intelligence; PTONI), and the pre-teaching checks on the dynamic assessment.

Caregiver Questionnaire. The caregiver questionnaire consisted of eligibility screening questions and demographic questions. The eligibility questions included the child's age, whether the child had been exposed to Spanish and English regularly, and whether the child had been diagnosed with a head injury, autism, or other medical condition. To provide background on the child's speech-language history, I asked if they had ever received special education, speech or language therapy, and if the caregiver or if the child's teacher or doctor had any concerns about the child's development. I also asked if the child had attention deficit/hyperactivity disorder (ADHD) as part of the child's profile; the presence of ADHD was not disqualifying given that

ADHD is commonly co-occurring but not exclusive of a diagnosis of DLD (Bishop et al., 2017). Demographic questions included the child's gender, race/ethnicity, and the caregivers' highest levels of education.

For SES, the combination of education and occupational status measures has been shown to be robust in relation to predicting language abilities (De Cat, 2021; Gatt et al., 2020). I considered following the recommendation of best practice to report a composite measure that would include caregivers' highest levels of education, income, and occupation status ranking (Broer et al., 2019). However, given the lack of a national comparative database for status or levels of occupation in the United States, I decided to remove occupational status. Income was also excluded because it felt invasive, especially because the caregivers were responding to this questionnaire without ever having developed trust with the research team; I did not want to risk losing potential participants. Thus, I opted for caregivers' highest levels of education.

Early childhood language research tends to ask maternal level of education (Ensminger & Fothergill, 2002), and maternal education has been shown to be the most predictive SES metric for child language (Hoff, 2006), which allowed me to compare my results to the literature. I asked for both caregivers' highest levels of education because some research has shown that it is more robust than the maternal education alone (Gatt et al., 2020). Initially, I delineated the level of education into the following categories: No education, 8^{th} grade or less, Between 8^{th} – 12^{th} grade or high school diploma, Technical school or other further education, Undergraduate college (1 year or more), Graduate school (1 year or more), or Prefer not to respond. After feedback (H. Mettler, personal communication, June 13, 2023), I expanded it to include "Undergraduate degree," and I separated "Between 8^{th} – 12^{th} grade or high school diploma" into two categories given the suggestion that there was a qualitative difference between those respective categories.

Bilingual Input Output Survey (BIOS; Peña et al., 2018). As part of the Bilingual English Spanish Assessment (BESA; Peña et al., 2018) battery, the BIOS was the primary measure to describe the child's bilingual profile because it provides data on the variables of age of onset and language use. It asks about the child's years of exposure to both Spanish and English and then details which languages the child hears and speaks for each waking hour on a typical day. The BIOS-Home is to be administered as an interview or fillable form to the child's caregiver while the BIOS-School is to be administered as an interview or fillable form to the child's teacher if the child attends school. If a child is in school, the BIOS-School information accounts for the child's language use during school hours while the BIOS-Home provides language use data for before and after school hours on weekdays. If the child does not attend school, the caregiver's responses will provide all the language use information for the weekday. The BIOS-Home also provides weekend language use data.

The primary outcome variable is the child's percentage Input, Output, and total Use (average of the input and output) for both English and Spanish. To get one variable for each language, I substituted the BIOS-School information for any hours the child was at school into the BIOS-Home. At times, the hours conflicted (i.e., the caregiver said the child was at home, and the school reported it was an hour at school). In those circumstances, the BIOS-Home response was used instead of the BIOS-school. Additionally, some families discussed the child's language use at school. If the BIOS-school was obtained, then that information was used. If it was not obtained, then caregiver report about school language use was used.

Hearing Screening. A pure tone hearing screening was administered to each child to rule out the possibility of hearing loss, which would interfere with their ability to interact with the auditory stimuli. Per the American Academy of Audiology (2011) guidelines, children were required to respond to 1,000, 2,000, and 4,000 Hz at 20dB or 25dB once to pass (see "Hearing Screening" in "Eligibility Assessment Administration"). These frequencies and this threshold were selected to maximize the specificity (92.2%) and sensitivity (100%) for detecting minimal hearing loss Dodd-Murphy et al. (2003) [poster presentation] as cited in American Academy of Audiology (2011).

Vision Screening. To ensure children could reasonably engage with the visual stimuli in the other study tasks, each child was tested for binocular near visual acuity using the LEA symbol chart (Hyvarinen et al., 1980). The LEA symbol chart is the preferred method of measuring near vision in this age range "because children know the symbols, left-right confusion cannot affect measurements, symbols have been validated properly against Snellen E's and Landolt C's for size (LEA symbols have to be 1.5x larger than the E in order to result in the same VA), symbols have equal discriminability, and good test-retest reliability" (Hyvarinen et al. (1980) as cited in Huurneman & Boonstra, 2016, p. 7). It consists of three columns with five shapes per row in each column, and the symbols decrease in size on each row, corresponding to visual acuity measurements (e.g., 20/20). Children were required to have 20/40 vision or better to pass.

Primary Test of Nonverbal Intelligence (PTONI; Ehrler & McGhee, 2008). The PTONI is a standardized assessment that measures one aspect of nonverbal intelligence, which is how well children identify an image that does not belong with others. Each item consists of a horizontal array of images, and the child is asked to identify the image that does not belong to the group. It was used to rule out cognitive impairment as cognitive impairment precludes a diagnosis of DLD (Bishop et al., 2017), and the current study is evaluating the dynamic assessment's classification accuracy for DLD rather than language impairments writ large. I chose the PTONI due to its strong psychometric properties and for the fact that part of its normative sample included Spanish speakers who were provided instructions in Spanish (n = 36 of N = 1,010). With a test-retest reliability of .97, an interscorer reliability of .99, and a Cronbach's alpha of .93 for item consistency, the PTONI demonstrated sufficient reliability.

To be considered eligible to participate in the study, a child had to have nonverbal IQ scores high enough to allow for a diagnosis of DLD (Bishop et al., 2017), which was decided to be above 73 (4 years) or 74 (5 years). These cut-off points were determined by subtracting two standard deviations (SD = 15) from the mean score of 100, resulting in a score of 70. This value was then adjusted by adding one standard error of measurement (SEM), which corresponds to the 68% confidence level.

Language Measures

Other than the index measure (i.e., the dynamic assessment) two measures were used to evaluate the child's language ability. After the dynamic assessment, the BESOS (reference standard) and language samples were administered in random order.

Reference Standard (BESOS; STARD #10b). The reference standard is the test that determines a child's language status of DLD or TD, and the index test's classification accuracy is compared to it. In the present study, the reference standard was the composite standard score from the Bilingual English Spanish Oral Screener (BESOS; Peña et al., 2008, in development). Each child took four subtests—Morphosyntax and Semantics for both English and Spanish. All subtests had age-specific items. The Morphosyntax subtests consisted of sentence repetition and cloze task responses for key grammatical skills, such as negatives in English and clitic pronouns in Spanish. The Semantics subtests examined word knowledge through receptive and expressive tasks and allowed the child's responses to be in either language regardless of the language of administration. Raw scores were converted to standard scores for each subtest, and a composite score was determined by combining the child's best language scores from the Morphosyntax and Semantics subtests. The composite score was based on a weighted average, which was found in the Bilingual English-Spanish Assessment test manual, Appendix F (Peña et al., 2018). A cutoff of -1*SD* below the mean (i.e., standard score of 85) was used, such that children scoring 85 and above were considered to have TD, and children with a score of 84 and below were considered to have DLD.

The BESOS has a concurrent sensitivity of 90% and specificity of 91% (Lugo-Neris et al., 2015; Peña et al., 2023), which is the true positive and true negative identification rate, respectively, at the time of evaluation. Thus, the classification accuracy of the BESOS is above the 90% criteria to be considered "optimal," which is why it was chosen as the current reference standard (STARD #11; Plante & Vance, 1994).

Current prevailing clinical recommendation for determining a child's language status is to use the combination of ethnographic interviews, language sampling, dynamic assessment, and, if available, standardized measures (Castilla-Earls et al., 2020). However, these recommendations assume that most single measures are not adequate on their own to render an accurate diagnosis. This study relied on this single measure, the BESOS, because it reached the "optimal" level of classification accuracy when used alone (Plante & Vance, 1994), and introducing other measures would add to measurement error.

Language Sample. The language sample was used as a continuous measure of language proficiency. Each child was asked to produce one narrative retell language sample in English and one in Spanish. The English story was *Frog Where Are You?* (Mayer, 1969), and the Spanish story was *Frog on His Own* (Mayer, 1975). These stories were chosen for parity with the comparative sample in the SALT database (Miller & Iglesias, 2020) and the comparative sample data from Hernandez et al. (2024) used to create a single, continuous score of language proficiency. The child's story retell was recorded by Zoom H4N Pro recorders and was transcribed and coded. See "Language Sample Transcription and Coding" and "Language Sample Analysis" in "Procedures" for more details.

The Dynamic Assessment

The index test is the experimental measure, in this case, the dynamic assessment. It is compared to the reference standard to determine its diagnostic accuracy. Using the framework proposed by Bamford et al. (2022), I will describe the dynamic assessment format before detailing the decision-making and creation process. Administration procedures are described in the "Dynamic Assessment Administration" subsection in "Procedures." Bamford et al. (2022) organized their descriptive categories under two domains: scope and method. Scope includes the purpose of the assessment, area of assessment, and population. Method includes the dynamic assessment format, the type of instruction, amount of instruction, measurement methods, and level of prescriptiveness. See Table 11 for the description of the current dynamic assessment within the Bamford et al. (2022) framework and Figure 2 for a visual example of each phase.

Table 11

Proposed Dynamic Assessment Within the Dynamic Assessment Framework

Category	Element
Domain: Scope	
Purpose	Diagnosis
Area of Assessment	Language (morphological learning)
Population	Spanish-English preschoolers with and without DLD
Domain: Method	
Format	Modified “Sandwich”: pre-teaching checks—teach—posttest + posttest + posttest
Type of Instruction	Statistical learning via focused stimulation and enhanced conversational recasting
Amount of Instruction	24 doses of focused stimulation, 24 enhanced conversational recasts
Measurement	-Teaching phase accuracy score -Posttest accuracy score -Modifiability score -Number of items to reach criterion -Time to completion of test
Level of Prescriptiveness	Prescripted[a]

Note. The Dynamic Assessment Framework was established by Bamford et al. (2022).

[a]Levels of prescriptiveness were characterized by Bamford et al. (2022) as “Prescripted,” “Planned but not scripted,” “Flexible,” and “Not specified.” These terms were not operationalized. Because the current dynamic assessment used a script, “prescripted” was chosen.

Dynamic Assessment Components (STARD #10a)

The dynamic assessment involved a pre-teaching phase consisting of three eligibility checks. If the child passed those, the teaching phase began, during which the child was taught the target morpheme, /-əb/, which meant “part of” an object. Finally, the child’s skill with the morpheme were tested on a posttest at three time points—immediately after the teaching phase, one hour post teaching, and one day post teaching.

Pre-Teaching Checks. The pre-teaching checks involved three checks to ensure the child could proceed to the teaching phase: (1) a speech screen, (2) a receptive guessability check, and (3) a vocabulary check. This speech screen was included to ensure the child could produce the target morpheme, /-əb/, when added to a one- or two-syllable word. Thus, articulation could be ruled out as a cause for lack of production of the morpheme. The speech screen was a ten-item screen of five words ending in /-b/ and five three-syllable words (see Figure 2 for an example). The child saw color pictures of each item and was asked to name it, and their response was recorded as “yes” or “no” for including a final consonant (for the /-b/ words) or three syllables, respectively.

The goal of administering the receptive guessability check was to ensure that there was not something inherently obvious about the target morpheme such that the child would know it automatically. If the child already “knew” the target morpheme above chance level, that would have been a confound. In other words, we wanted to be sure the child did not know the target morpheme so that there was something to teach during the teaching phase. The guessability check consisted of the same ten base words as in the speech check, and each item was presented in a quadrant of four images. The four image choices included the target morpheme (i.e., a partial image) and three distractors: an inflectional morpheme, a derivational morpheme, and an object alteration. Because the inflectional and derivational morpheme distractors already had known linguistic correlates, other visual alterations were added to increase their viability as true possibilities to be guessed as the target morpheme (see Figure 2 for an example).

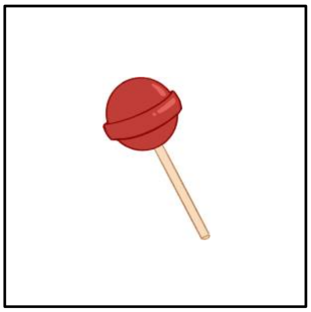
The vocabulary check ensured that participants had sufficient English vocabulary such that they could focus on producing the target morpheme during the posttest without having to be linguistically taxed by unfamiliar vocabulary. The child was shown a picture (one per page) and asked to label it until they correctly labeled ten images. There were twenty images available in case a child did not know the first ten items. This allowed for variability across children's English semantic knowledge. Images were in their whole form (i.e., without visual alterations and without the target morpheme affixed). Scoring was binary—correct or incorrect.

Figure 2

Dynamic Assessment Example by Phase

Checks (pre-test)

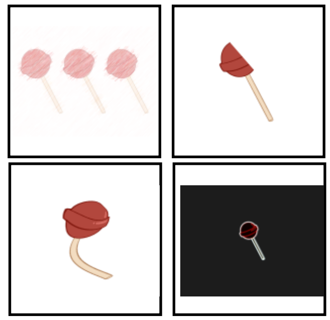
Speech



"What is this?"

Child: "lollipop"

Guessability



"Show me the lollipopəb."

Child: points

Vocabulary

"What is this?"

Child: "mouth"

Teaching: FS

"She sought refuge on the old couchə**b** and buried herself under a shirtə**b**."

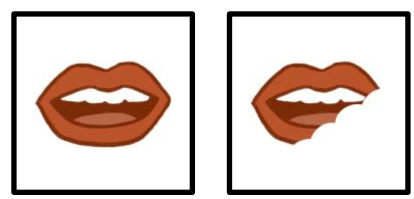
Teaching: ECR

"Isa scurried away but bumped into a cup. The cup fell. Then it was a…"

Child: "cupə**b**"

ECR: 1) *cue for child to look at assessor*
2) "Then it was a cupə**b**."

Posttests



"This is a mouth. What is this?"

Child: Mouthə**b**

Teaching Phase. The teaching phase consisted of a scripted story with corresponding images. The story presented 24 doses of the target morpheme, /-əb/ (partial object), affixed to a platform word via focused stimulation in the first section and 24 enhanced conversational recasts of the target morpheme in the second section. This order of presentation allowed the child to first be exposed to the target morpheme (focused stimulation) before being required to respond (enhanced conversational recast). This was in opposition with the original assessment by Roseberry and Connell (1991), which had the child only listen during the teaching phase. Having the child respond allowed for an interaction, which aligned the procedures more closely with the traditional definition criteria of a dynamic assessment. Additionally, this interaction provided the opportunity to observe the child's modifiability in response to teaching.

For focused stimulation, one dose was considered having the child view an image of a partial object while hearing one instance of a platform word with the target morpheme affixed, and each dose occurred within a grammatically correct sentence. A platform word is the root word to which the target morpheme was affixed (e.g., "cat" is the platform word in "catəb"). Most doses were presented as the only dose in the sentence, but twice, they were presented in a list order to reduce administration time (e.g., "I see a trainəb, a starəb, and a truckəb.").

The enhanced conversational recast involved presenting the platform word without the target morpheme and its visual whole-object correlate twice (e.g., "bed"). The child was cued to the child to finish a sentence by labeling the partial object that they saw. A correct response would be the platform word plus /-əb/ (e.g., "bedəb"). One dose was considered the enhanced recast, which entailed obtaining the child's attention and saying the correct answer (e.g., "bedəb") (Plante et al., 2014). Across both focused stimulation and enhanced recasting, every platform word and sentence was unique to maximize the variability principle of statistical learning (Plante & Gómez, 2018).

All doses were presented in the context of a scripted narrative of a cat and dog helping their family. The plot structure followed a typical North American, Western plot with the exposition consisting of a presentation of the characters and setting and an initial problem. The characters worked through a series of obstacles as they attempted to solve the initial problem as the storyline followed their journey. The main characters were in the real world for the focused stimulation part of the story and then transitioned to a magical world for the enhanced conversational recasting part of the story. They came back to the real world toward the end of the ECR section. The story ended with the characters solving the problem and the characters coming together as a happy family. The narrative was developed by me and was shortened and enhanced with feedback from lab members.

The story served to keep the child interested in a familiar task and motivated to respond. In addition, it allowed for all doses and recasts to be presented in the context of a semantically appropriate, grammatical utterance, which allowed for the regularity and variability principles of statistical learning to be fulfilled. Additionally, having the morpheme presented in a story context may have enhanced learning the semantic component of the derivational morpheme (Dautriche & Chemla, 2014). Finally, the story followed a script, which may facilitate the implementation of this dynamic assessment in the clinical setting as well as support replicability across future studies.

Modifiability. The first modifiability scale, developed by Peña (2000), consisted of rating the examiner's effort on a scale of 0 ("extreme") to 3 ("slight") and rating the child's responsivity on a scale of 0 ("none at all") to 3 ("high"). Child responsivity referred to the how well the child seemed to be learning the target skill (Lidz, 1987, 1991). The third part of the originally developed scale was a rating of transfer on a scale of 0 (“low”) to 2 (“high”). Because transfer is dependent on how well the child transfers the target skill from the teaching phase to the posttest, the child’s transfer was the only piece of modifiability that was rated after each posttest.

The second scale, called the Mediated Learning Observation (Peña & Villarreal, 2000, as cited in Peña et al., 2007), consisted of four constructs each measured by individual behaviors rated on a 1 to 5 scale. The constructs were internal social-emotional (affect), cognitive arousal, cognitive elaboration, and external social-emotional (behavior). Because the teaching phase in the current dynamic assessment did not require the child to plan, reflect, and use strategies, the cognitive elaboration construct was removed from the present rating scale. Additionally, within the external social-emotional (behavior) construct, the rating for “responsiveness to feedback” was removed given that the child was not given direct feedback due to the nature of the statistical learning task.

In addition to the rating scales, objective measures of modifiability were measured. The time to do the teaching phase was recorded because time may be a proxy for how well the child engaged in the task. A child who needed to be reminded to focus presumably would take longer to complete the teaching phase than a child with sustained on-task behavior. Similarly, a child who knew the answer may have been quicker to respond than a child who was unsure of the task and who may have taken more time to generate a response or who needed to be prompted. Time was recorded at the beginning and end of each dynamic assessment phase.

A second objective measure of modifiability was the criterion during the teaching phase, which was the first three correct consecutive responses the child produced. This measure was intended to indicate how quickly each child grasped the learning task. The criterion provided information about the child's speed of learning, which helped fulfill one of the key goals of dynamic assessment—understanding the child's learning ability.

Posttest. The goal of the posttest was to measure if the child had generalized their learning of the invented morpheme to a new set of platform words without support or feedback. The posttest consisted of a whole image next to a partial image. The assessor labeled the whole image and asked the child to label the partial image to elicit the target morpheme. The ten correctly labeled words on the pre-teaching vocabulary check were the ten platform words used on the posttest.

The same posttest was administered three separate times. The first posttest was administered immediately after the assessor scored the child's modifiability after the teaching phase. Subsequently, a posttest was administered one hour after the end of the teaching phase or as late as possible within the first session if an hour had not yet been reached. The third was administered one day after the first posttest.

Procedures

Study Flow

This section outlines the chronological flow of the study. See Table 10 for a list of all measures used. Recruitment flyers and information packets directed caregivers to fill out the consent form and, subsequently, the caregiver questionnaire. Most caregivers filled out the forms on REDCap, a HIPAA-compliant online platform and a few filled out paper forms, which I transferred to REDCap. Once both documents were received and I determined the child met the initial eligibility criteria of (1) age, (2) lack of diagnoses that would preclude DLD, and (3) regular exposure to Spanish and English, I called the caregiver within 1 – 2 business days to collect the Home BIOS information. In that call, I scheduled and confirmed the location for data collection (e.g., the child's school or a location convenient to the family). If the child was in a school or daycare setting, then I would contact the child's teacher to collect the School BIOS survey. Assessors were not given any information about the child's developmental or medical history prior to data collection (STARD #13a-13b).

Data collection with the child proceeded once the caregiver's consent, questionnaire, and home BIOS information were received. We did not wait for the teacher's response for school BIOSes but attempted to contact teachers periodically for up to 8 weeks after data collection. Of the children who attended school, we obtained school BIOS information for 93.33% of the children in a school setting (*n* = 14 school BIOS questionnaires/15 children in school). If data collection occurred in a school setting, the time was agreed upon with the administrators and teachers prior to starting. When data collection was during school hours, we attempted to minimize interruption to the important parts of the child's day. Data collection occurred outside of the classroom in a private, quiet space that was well-lit and had minimal distractions.

Data collection was scheduled to last up to 1 hour, 30 minutes per day for two days with the option to add up to two more days if the child did not finish within the scheduled time span. By way of a visual schedule, children learned what to expect during data collection, and they added stickers to each slot on their visual schedule when they finished a data collection activity. Children were offered frequent breaks and were allowed to take a break whenever needed, though we tried to avoid taking a break in the middle of an assessment.

After briefly establishing rapport with the child, the assessors administered the eligibility assessments: the hearing screening and vision screening (see "Eligibility Assessments"). Because these were to establish eligibility, order effects were not a concern, so the hearing screening was always followed by the vision screening. The child was required to pass the hearing and vision screening contemporaneously for other assessments to be administered. The third eligibility assessment, the PTONI, was administered either immediately after the initial screenings ($n = 1$) or after the dynamic assessment ($n = 15$)[3] to ensure we had enough time for the dynamic assessment to be fully administered the first day. To save time in the moment of data collection, the PTONI was scored after the first day of data collection, and the child had to have a passing score to proceed on day 2.

The dynamic assessment was administered after the hearing and vision screenings and before all other assessments of language (see "Dynamic Assessment Administration" for details). The dynamic assessment always came first because we needed to conduct a posttest after one hour, and we only had 1.5 hours for each data collection session. Additionally, having the index test come before the reference standard reduced bias in our modifiability scoring because we only knew how the child performed on the dynamic assessment and had not yet formed an opinion of their language from the reference standard and language samples (STARD #13a).

After the dynamic assessment was administered, the reference standard (BESOS) subtests and the language sample were administered in a randomized order. All language measures (i.e., the BESOS, language samples, and dynamic assessment) were audio recorded on Zoom H4N pro recorders. The delayed posttests for the dynamic assessment came as close as possible to one hour after and then one day after the end of the immediate posttest, respectively, without interrupting the administration of another activity. Most often, it was administered between the BESOS subtests or language samples. After 1.5 hours or after the child demonstrated the need to finish for the day, whichever came first, data collection for the first day ceased. The second day

[3] Two participants did not have the order of the PTONi recorded, so it is unclear if they received it before or after the dynamic assessment.

of data collection began with the next item on the randomized order and proceeded until either all activities were complete, the time limit was met, or the child showed signs of needing to finish (STARD #22).

After each day of data collection, the child selected a small, toy prize to take home. If the child was at home or with their caregiver for data collection, they took the prize immediately. If the child was at school for data collection, the prize was put in a bag and sent home when all data collection was completed for the children at their school. Once the child finished all data collection activities, the family was sent a $30 Amazon gift card via email. The family was offered a prorated compensation of one small, toy prize and $5 if the child began data collection but did not complete it for whatever reason (e.g., failing the eligibility screenings, deciding to withdraw, not being able to schedule day 2).

A few weeks after data collection, families received a report by hard copy or email of their child's performance in the study. Families were reminded that this was not a formal speech-language therapy evaluation and were given the phone numbers to the Child Find centers for the two largest local school districts where they could request an evaluation if desired. If a child did not pass a screening or demonstrated below average language abilities, I called the caregiver to discuss the results of the study and answer questions before sending home the report.

Eligibility Assessment Administration

For the hearing screening, vision screening, and PTONI, the assessor provided instructions and interacted with the child in their language of greater use, based on the BIOS data. If the child's use was balanced across languages, if we did not have the complete BIOS data yet, or if the child seemed not to understand, the assessor interacted with the child in both languages. Additionally, if the child began to use one language primarily or respond only to one language, the assessor was trained to follow the child's lead and use the language of the child's preference.

Hearing Screening. At the beginning of each day of data collection, the audiometer was checked to ensure it was working, and because testing did not occur in a soundproof environment, a biologic noise level check was conducted (American Academy of Audiology, 2011). The biologic noise level check involved having an assessor with known normal hearing respond to the frequencies to be tested at 10dB below the target threshold of 20dB in the environment in which testing would occur. This ensured the frequencies of interest could be heard despite ambient noise. If the adult listener did not pass at the target frequencies, we attempted to find an area with less ambient noise. If that was not possible, we raised the required pass threshold by 5dB to be 25dB and ensured the assessor could detect each frequency at 15dB.

During the hearing screening, children were told, "You are going to hear some beeps. Every time you hear a beep, raise your hand. Sometimes the beeps will be very soft. Raise your hand as soon as you hear the beep. Do you understand?" (See Appendix B. Spanish Translations of Instructions for Tasks). After giving the initial instructions, the assessor put the over-the-ear headphones on the child's ears, ensuring that they fit snugly over the child's ears without obstructions between the headphones and the ear (e.g., hair). The assessor played at least two test tones of 1,000 Hz at 40 dB to teach the child to respond (American Speech-Language-Hearing Association, n.d.-a). Once the child demonstrated they understood how to respond, either by raising their hand or placing a coin in a piggy bank, the assessor began the test. Children were positioned to sit facing away from the audiometer to reduce the likelihood of responding due to seeing the button pressed. The assessor tested all frequencies in one ear before switching to the other ear.

A child passed the hearing screening if they responded to one beep per frequency in each ear. However, if a child ever did not respond to a beep at a frequency, at least a second beep was presented, up to four beeps total per the American Academy of Audiology (2011) guidelines. If a child failed at any single frequency, the child did not pass the hearing screening. The child was rescreened soon after the initial hearing screening using the same procedures as in the initial hearing screening. If the child did not pass the second hearing screening, they were withdrawn from the study. If the child did pass the initial or second hearing screening, they continued with the other eligibility checking procedures.

Vision Screening. On the LEA symbols card, the child was first asked to label the symbols that they would see in the exam (circle, house, apple/heart, square); the child's labels were used (e.g., if a child said "triangle" for "square," then "triangle" was counted as correct during the exam). One child had inconsistent shape labels and so was asked to point to corresponding picture cards to identify the symbols they saw. The examiner ensured the child's face was 40cm from the LEA symbols card by using a 40cm string that was attached to the LEA symbols card and anchoring it on the child's temple. The child was asked to label the first symbol on each row in the middle of the LEA symbol card. Once the child hesitated, squinted, or was incorrect, the examiner moved up one line and prompted the child to name all five symbols in that line. They moved down one line at a time, asking the child to name the symbols for the entire line until the child got two or fewer symbols correct on a line (i.e., three or more errors). The line above that was the child's near visual acuity. Children were required to have 20/40 vision or better to pass. If a child did not pass on their first try, the caregiver was notified and offered to readminister the vision screening on a different day.

If a child wore glasses, they were asked to first conduct the vision screening without their glasses and then again with their glasses. This was to determine whether testing could continue if the child were to take off their glasses or forget them on a subsequent day. The same passing score of 20/40 or better was required. If the child passed without their glasses, they could continue data collection with or without their glasses. If they passed only with their glasses on, they were required to wear their glasses for each subsequent data collection activity.

PTONI. To take the PTONI, each child practiced identifying the dissimilar picture in an array of images for five sample items. If the child got an answer incorrect, the child was told their answer was incorrect and prompted to try again after the instructions were presented again. If still incorrect, the child was told the correct answer and why it was correct. After the sample items, the child began the test at the appropriate start point for their age (4-years at item 1, 5 years at item 15). The child's answer was recorded on the data sheet and scored as 1 (correct) or 0 (incorrect). The assessor tested the child until they reached a ceiling of 5 items incorrect within a series of 7 consecutive items. Once the child hit the ceiling, if a basal of 7 consecutive correct items had not been reached for a 5-year-old, the assessor went back to the start point and administered the items in descending order until the basal criteria had been met.

Language Measures Administration

Reference Standard (BESOS). Each BESOS subtest was administered in the standardized procedure required by the test manual and instructions. There were no basal or ceiling rules, so all items were administered to all children on their age-specific subtests (i.e., 4-year-olds received all the items for 4-year-olds). For the Morphosyntax subtest, codeswitched responses for the skill of interest were marked as incorrect; codeswitched words that were not the target skill were acceptable and did not count against scoring. For the Semantics subtest, all responses could be in either language. The pre-determined cut score was a standard score of 85 given its optimal classification accuracy (STARD #12b; Peña et al., 2023).

Language Samples. On an iPad, the assessor played an audio recording of the script provided by the Systematic Analysis of Language Transcripts (SALT) software (SALT Software LLC, n.d.-a). The child heard *Frog Where Are You?* (Mayer, 1969) in English and *Frog on His Own* (Mayer, 1975) in Spanish while looking at the corresponding pages in the wordless picture book. Before listening to the story in the respective language, the child was told, "Here is a book. We are going to listen to this story while we look at the book together. When we finish, I want you to tell the story back to me in English. Ok? Let's look at the book." See Appendix B. Spanish Translations of Instructions for Tasks. After hearing the story, the child was told, "Okay, now I would you like to tell me the story." During the story, the assessor only cued the child if they paused, indicated uncertainty, or needed reminders to stay on task. Cues were as general as possible, such as "You're doing great. Keep going." If the child asked a question, the assessor responded that they did not know (e.g., in response to "Why is he doing that?") or redirected the child to do their best (e.g., in response to "What is that animal?"). The child was cued to stay in the target language if they began to speak in phrases or whole utterances in the non-target language. Most of the language sample had to be in the target language to be transcribed and analyzed, so some children only had a language sample in one language (n = 8 both languages, n = 8 English only, n = 2 Spanish only).

Language Sample Transcription and Coding. Each language sample was transcribed by a native speaker of its respective language and checked by a second listener (double scorer) for inter-rater reliability. Disagreements were resolved by me, which we deemed a revision. If reliability was below 90% at the revision phase, a trained research assistant was asked to review the transcript for what we called a “triple score.” In the triple score, the research assistant commented on each revision stating their agreement or disagreement, providing a reason for disagreement. Disagreements were resolved by me as an expert rater with consensus discussion if needed.

Subsequently, the transcript was divided into C-units and modified C-units for Spanish and English, respectively. A C-unit is an independent clause with its dependent clauses. A modified C-unit is a typical C-unit that is also parsed when there are multiple verbs in a row without restating overt subjects. For example, the utterance, “he called and listened for the frog” would be one C-Unit but two modified C-units, parsed after “called.” The modified C-unit is particular to English samples from Spanish-English speakers because Spanish, as a pro-drop language, allows for a verb to be an independent clause. Thus, there is parity in how the two languages are segmented, which is important for bilingual samples.

Next, samples were coded for SALT grammatical conventions and errors (Miller & Iglesias, 2020). Both C-unit parsing and error coding followed the same process as the transcriptions: an initial attempt, a double score, and a revision with a triple score as needed. See Table 12 for the reliability for each part of the transcription process. See Appendix C. Language Sample Codes for a list of the SALT conventions and errors marked. Finally, the coded transcriptions were analyzed in SALT using the Bilingual English (or Spanish) Story Retell databases.

Table 12.

Interrater Reliability for Language Sample Analyses

Language	Transcription		C-Units		Error Codes		Grammar Codes		CS and F codes		Overall	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
English	95.64	4.69	99.03	1.60	93.52	7.73	95.79	5.83	83.13	34.78	93.42	13.52
Spanish	93.36	6.52	95.55	0.95	93.77	11.44	97.09	4.97	95.25	14.50	95.80	5.35
Total	94.50	5.60	99.29	1.28	93.65	9.58	96.44	5.40	89.19	24.64	94.61	9.44

Note. Values reported in percentages. *M* = mean, *SD* = standard deviation, CS = codeswitch, F = Fragment codes used to denote a modified C-Unit.

Language Sample Analysis. To have a continuous measure of language proficiency, we calculated each child's Percent Grammatical Utterances (PGU) and Mean Length of Utterance Word (MLU-W). These indices were chosen for their predictive value of bilingual children's language abilities in narrative retell language samples (Bedore et al., 2010; Hernandez & Castilla-Earls, 2022; Kapantzoglou et al., 2017; Kapantzoglou et al., 2015; Restrepo, 1998). Similarly, Kapantzoglou et al. (2017) found that grammatical errors per communication unit (GE/CU), similar to PGU, was predictive of Primary Language Impairment (i.e., DLD), but MLU-W was no longer a significant predictor once GE/CU and lexical diversity were added to their model. Given that this was a finding from only one study, MLU-W was still included in the current analyses. Additionally, having PGU and MLU-W allowed me to compare each child's language ability to a sample of 4-year-old and 5-year-old children from Hernandez et al. (2024), which allowed for the characterization of a spectrum of "low" to "high" language ability in the current sample.

To calculate PGU, the number of agrammatical C-units and modified C-units across Spanish and English, respectively, was subtracted from the number of grammatical (modified) C-units for both languages divided by the total number of (modified) C-units combined from both language samples (Eisenberg & Guo, 2013; Eisenberg et al., 2012); using SALT, this was 1 – Percent Utterances with Errors (Percent Utterances with Errors was obtained from the Errors Summary). The total of each child's Spanish and English utterances were included in their PGU scores because combined scores yielded higher sensitivity (91%) and specificity (89%) per Hernandez et al. (2024) than using the best language PGU score, which had a sensitivity of 88.9% or 86% and specificity of 81.6% or 57% per Hernandez et al. (2024) and Lazewnik et al. (2019), respectively. Although PGU is not used for diagnostic purposes in the current study, the better diagnostic accuracy of the combined language PGU indicates that it is a better measure of a child's language abilities.

For MLU-W, each child received one value for Spanish and one for English, calculated by averaging the length of each utterance and dividing by the number of words in the transcript. This value was provided by SALT.

Utterances in which the child codeswitched were given the code [CS], and analyses were run with and without codeswitched utterances. Including codeswitched utterances represented the child's true, naturalistic use of their bilingualism. Excluding codeswitching allowed me to match the procedure done by our comparative database by Hernandez and Castilla-Earls (2022). Three children had a significant reduction in their number of utterances in Spanish, to the extent that analyzing the language sample did not make sense, when their codeswitched utterances were excluded (e.g., from 66 C-Units to 9 C-Units). In those cases, the English PGU and MLU-W were used, and the Spanish data was excluded for the analyses without codeswitched utterances. To minimize the number of excluded utterances, children were encouraged to use the target language during their language samples if they codeswitched more than a few words.

When coding for errors of codeswitched utterances, we left instances of ambiguous errors uncoded. If a child said, “la cat,” the “la” could have been seen as an error because the Spanish noun for cat, “gato,” takes a masculine pronoun, “el.” However, “gato” can be feminine at times, such as when talking about a female cat (e.g., “la gata”). Thus, gender agreement could not be determined as correct or incorrect and therefore was not coded as an error in such instances. Words that were borrowed across languages but that were not always considered canonical in that language (e.g., borrowing “catch” from English to produce “cachar” or “cachear” in Spanish) were coded as codeswitches rather than errors.

The language proficiency measures from the language sample served the dual purposes of characterizing each child’s bilingual profile with their relative proficiency per language while also allowing for me to sample across a range of language abilities to provide a characterization of the level of spectrum bias that was present or avoided. To show relative proficiency per each language in the bilingual profile, we needed one score per language, so the MLU-W Spanish and MLU-W English were used. Measuring language proficiency to sample across the language ability spectrum was complicated by needing one score when each child had three scores (PGU, MLU-W_{Sp}, MLU-W_{Eng}).

To yield a single proficiency score, each child’s language ability was calculated via an average of *z*-scores of the PGU total and the child’s best language MLU-W (i.e., MLU-W_{Sp} or MLU-W_{Eng}). To calculate the *z*-scores, the mean and standard deviations for the TD group from the Hernandez and Castilla-Earls (2022) data were used. This method had the advantage of being relatively robust by incorporating measures that capture both languages, which, when used together, have been shown to yield high sensitivity and specificity (Hernandez & Castilla-Earls, 2022). An a priori classification of language ability was established as a *z*-score of < -1.0 = "low", -1.0 – 1.0 = "medium," and > 1.0 = "high."

Dynamic Assessment Administration

The assessors performed the dynamic assessments before any other language tests (i.e., the BESOS and language samples), and they were blinded to caregivers' responses on the caregiver questionnaire, including whether the child was or had been in speech-language or special education services (STARD #13b). Because I was privy to the caregiver questionnaire responses, I never administered the teaching component of the dynamic assessment. Each child took the dynamic assessment in English. The children were not told whether they had DLD or TD (STARD #13a). Each part of the dynamic assessment is described in the subsequent sections.

Pre-Teaching Checks. The pre-teaching phase involved three checks to ensure the child could proceed to the teaching phase: (1) a speech screen, (2) a receptive guessability check, and (3) a vocabulary check.

The ten items for the speech screen and guessability check were administered in tandem. For the speech screen, the child was shown a picture of a platform word in its whole form and asked to label it. The assessor scored whether the child produced the target speech form or not. To ensure that the assessor could reliably score the child's verbal responses in the teaching and posttest phases, the child was required to produce at least 60% of the target speech forms correctly to move on to the teaching phase, including producing three-syllable words and final stop consonants, both of which were needed to appropriately produce the target morpheme.

For the receptive guessability check, after the child produced the platform word, on an adjacent page, the assessor showed the child four images, one of the target morpheme and three distractors, and said, "Show me the __- əb," affixing the target morpheme onto the platform word. The assessor circled the number of the image to which the child pointed and scored it as correct (1) or incorrect (0); ideally the child did not guess the target morpheme correctly above chance.

After the child passed the speech screen and guessability check, the vocabulary check was administered. The child was required to correctly name ten words in to continue with the assessment. If the child did not correctly name all the first ten words, they were shown more words until they could name ten or until there were no alternate words left (ten alternate words; twenty possible items to name). Whichever ten words the child correctly labeled on the pre-teaching vocabulary check were the ten words that were then tested on the posttests. Prompts were given if a child was unsure, used a similar but non-target label (e.g., “lips” for “mouth”), or used Spanish. After prompting, if a child used a different but pragmatically appropriate label for an item (e.g., “lemonade” for “drink”), their label (e.g., “lemonade”) was used on the posttests so long as it was not a target word in the teaching phase.

Teaching Phase. After the child passed all three pre-teaching checks, they entered the teaching phase. The assessor recorded the start time just before reading the introductory instructions, “I’m going to read you a story. At first, you’ll just listen. Then later, you’ll help me make some new words for the story. Are you ready?” Then, the assessor started reading the child the scripted story with focused stimulation and enhanced recasting to teach the target morpheme, /-əb/ (partial object).

During the focused stimulation section, the child listened and was not asked to respond. During the recast portion, the child was cued to label the platform word with the target morpheme. Regardless of the child’s accuracy, as long as the child produced the platform word, the assessor first ensured they had the child’s attention, then said the correct answer as the enhanced conversational recast (Plante et al., 2014). Assessors were encouraged to use the final phrase when providing the recast rather than just the word itself. The child was not provided with any feedback other than the enhanced conversational recast.

The child was scored based on their use of the correct platform word and morpheme; see Table 13 for a list of possible child responses and scores. If the child used the correct root word and morpheme in the target language, the child received a 2. If the child used the incorrect root word but correct morpheme, the child received a 1. A score of 1 was also merited if the child attempted the target morpheme but had phonological errors (e.g., “eyebub”) or produced a different nonsense morpheme (e.g., “hairker”). If the child used an incorrect morpheme (i.e., one that already exists) or no morpheme, the child received a 0. Regardless of the correctness of the morpheme, if the child used the incorrect root word, they were cued to “Use the word ____. What is this?” The child received a score for their initial response and another score for their response post cue as applicable. Additionally, if the child spoke about the image or story generally (e.g., “she took a bite out of it!” or “broken apple”), the initial score was 0. Whenever the child spoke generally or produced only the root word without a morpheme, they were cued to “Help me make a new word. Use the word ___. What is this?” or “Your job is to make a new word. Use the word ___. What is this?” Another common cue was to ask the child to speak like the examiner had been, such as, “Can you make a new/silly word like I’ve been doing in the story?”

Table 13

Scoring and Cueing Scheme for Teaching and Posttests

Score	Target language					Non-target language		Example	Cue?
	Platform word		Target Morpheme			Platform word			
	C	I	C	I	P	C	I		
2	x		x					catəb	
		x	x					dogəb	X
			x			x		gatoəb	X
			x				x	perroəb	X
1	x				x			catbəb	X
		x			x			dogbəb	X
					x	x		əbgato	X
					x		x	perrobəb	X
	x			x				cats	
0		x		x				dogs	X
				x		x		gatos	X
				x			x	perros	X

Note. C = Correct, I = Incorrect (omitting invented morpheme, using a pre-existing bound morpheme, or no response), P = Partially correct (uses target but with incorrect placement, uses phonologically different form that is not a pre-existing bound morpheme); Examples provided as though target language was English with the platform word "cat."

Modifiability. Modifiability was rated by the assessor immediately after the teaching phase and just before the first posttest, so the assessor had easy recall of the child's behaviors during the teaching phase but was not biased by the child's performance on the posttest. The assessor rated the child on two previously developed scales (see "Modifiability" in "Dynamic Assessment Components"). Assessors were trained first through discussion about the scales and then practice with role-play administrations and with pilot participants. Each assessor also engaged in a consensus discussion and rating comparison with me after three early participants. After that, the ratings were the assessors' own. Having raters provide their own ratings (i.e., rather than engage in consensus discussion for every participant) was commensurate with typical practice in the literature (Peña et al., 2006; Peña et al., 2001; Peña et al., 2007; Petersen et al., 2017; Petersen et al., 2020; Ukrainetz et al., 2000). For inter-rater reliability, I also conducted a live rating of modifiability for every child in the field.

Posttest. To administer the posttest, the assessor showed the child a page with the whole visual form of a platform word next to its partial form. The assessor pointed to the whole form and said, "This is a ___," and then they pointed to the partial form and asked, "What is this?" If the child used the wrong word, the assessor would cue, "Use the word ___. This is a ____ (pointing to whole object). What is this?" The child was only given general feedback to encourage their continued efforts but nothing specific regarding their performance. The child's initial response and response post cue (if applicable) was scored separately but both with the same scoring scheme used during the enhanced conversational recast in the teaching phase.

Fidelity and Reliability

Using a fidelity checklist, I scored fidelity live for 100% of dynamic assessment administrations. The checklist included "yes" or "no" questions for whether the pre-teaching checks were being administered and scored according to the protocol. For the teaching phase, the assessor was scored with a "1" or "0" for whether they (1) appropriately prompted the child if a prompt was needed ("N/A" was marked if no prompt was needed and no prompt was given), (2) cued for the child's attention before the recast, and (3) provided the recast. The fidelity checklist for the posttest scored "N/A," "1," or "0" for appropriate prompting if needed. Typically, I only scored fidelity for the immediate and sometimes for the hour posttest. Fidelity was not scored for the one-day posttest because I usually administered it without research assistants present. In addition to the dynamic assessment fidelity checklist, I monitored the assessors' administration of all other tasks, and I provided live feedback as needed to maintain protocol fidelity.

For every dynamic assessment and every BESOS morphosyntax subtest, inter-rater reliability was scored by a trained assessor using the audio recordings. The guessability check on the dynamic assessment involved pointing, which could not be recorded on audio, so inter-rater reliability for the speech and guessability check (administered together) was conducted live. Similarly, the BESOS semantics subtests required the child to point, so inter-rater reliability was conducted live whenever the trained assessor was present with me (i.e., if the BESOS semantics subtests were administered on day 1 of data collection). Discrepancies were resolved by relistening to the audio and consensus discussion.

For the language sample analyses, reliability was ensured through the procedures of initial attempt, double-scoring, and a revision for each phase of the analysis: transcription, C-unit parsing, and grammar and error coding.

Training Research Assistants

Research assistants were Spanish-English bilingual undergraduate students in the Speech, Language, and Hearing Sciences department who trained to be assessors or to work with language samples. All assessors were trained on all data collection tasks by first watching training videos and then practicing with peers. Then, they had to demonstrate competence with each procedure in a fidelity check with me; their performance was scored on a checklist for each task. For any areas that needed improvement, I discussed it with the assessor and had them practice it more. An assessor passed their fidelity check when they correctly conducted each procedure and when I had confidence that they would be able to correctly conduct the procedure in an actual data collection. Once that fidelity check was passed, the assessor was allowed to conduct data collections with me present.

Other research assistants were trained on language sample transcription and analysis for bilingual children. Before working on real data, research assistants were required to pass the self-paced SALT online courses (SALT Software LLC, n.d.-b) 1301: Transcription – Getting Started, 1302: Transcription – Transcription Format, 1303: Transcription – Utterance Segmentation, 1304: Transcription – Conventions Part 1, 1305: Transcription – Conventions Part 2, 1306: Transcription – Conventions Part 3, 1308: Transcription – Practice Samples (two practice samples required), 1603: Bilingual SE – Transcribing Samples, and 1604: Bilingual SE – Transcription Practice Samples (one practice sample required). In addition, they met with me individually to discuss questions and as a group to practice chunking modified C-units and coding for SALT conventions and errors. They then were required to pass a fidelity check of C-units parsing and error coding before moving on to real data. To support reliability, I provided cheat sheets, examples, videos, and individualized feedback for the research assistants working on language sample analyses.

Analysis Plan

Analyses were run in RStudio version 2024.04.0+735 (2024.04.0+735) (R Core Team, 2021). Descriptive statistics were used for Aim 1 and, for Aim 2, were examined prior to conducting statistical tests. Statistical tests for group comparisons were run as *t*-tests in a Bayesian framework, which has the advantage of allowing for probabilistic statements to be made about the result. Bayesian statistics run a Markov Chain Monte Carlo (MCMC), an algorithmic function that uses the current data, prior assumptions (e.g., mean, standard deviation, type of distribution), and the model to generate the most likely posterior distribution using probabilities. The posterior distribution may be described by a likely range of values for the parameter estimate—the credible interval (CI). The Bayesian CI (often 95%) is not to be confused with the frequentist confidence interval as the Bayesian 95% CI can be interpreted as a 95% probability that the effect lies within the population given the observed data and the model. In contrast, the frequentist 95% confidence interval is interpreted as 95% of the time, with repeated replications, the population parameters would fall within the confidence interval, assuming all model assumptions had been met. Because the posterior distribution and CI are generated by probabilistic mechanisms, using Bayesian statistics allows me to make statements about what is likely occurring, even when what is likely occurring is close to zero, rather than a statement of uncertainty about the occurrence of a null result as is the case in frequentist approaches. This is ideal for small sample sizes like the current study because it can lend more confidence to the result. The MCMC is based off the data, so more data is still better. Nonetheless, it is more robust than traditional frequentist approaches even with small datasets.

All Bayesian models were run in the *brms* package, version 2.21.0 (Bürkner, 2017) with default, uninformative priors. Uninformative priors were used because, even though current study is a replication of previous studies, it is different enough that the analysis was exploring whether this version of the dynamic assessment yielded feasible, interpretable results and could not easily be compared to previous versions in terms of its scoring. Four chains were used with up to 6,000 iterations each (including up to 3,000 iterations for warm-up) resulting in 12,000 post-warm-up draws. Each model was checked for model convergence, which is determined by multiple factors. First, Rhats, “the ratio of the within- and between-chain variability” (van de Schoot et al., 2021, p. 14) must equal 1.00 as a value of 1.01 or greater indicates a lack of convergence. If the Rhats were greater than 1.00, the model was rerun with more iterations. The effective sample size (ESS) values of the sampled parameters were examined to ensure they were ideally greater than 1,000 and at least similar to each other across each model parameter. Trace plots, which show the parameter estimate across every iteration for each of the four chains, were visually inspected to ensure the MCMC process reached a relatively stable state (van de Schoot et al., 2021). Finally, visual inspection was used to determine how well the observed variable compared to the posterior predictive distribution datasets. When a model showed evidence of not having converged, I increased the number of iterations and reran the model.

Reporting for Bayesian statistics differs from frequentist frameworks because in the Bayesian framework, there is no *p* value. Instead, there is the mean or median, the 95% CI, and the probability of direction (*pd*) value. If the 95% CI crosses over zero, the effect estimate is positive or negative, meaning that the tested model does not seem to produce a clear effect. The 95% is an arbitrary choice, and it can be set at the author's discretion. In the current study, I chose 95% for more stringency given the purpose of the current study is related to the high stakes scenario of assessment for determining a need for SLP services. The *pd* expresses the percent of the posterior distribution that is expected to be negative or positive and is highly correlated with the frequentist p-value at a 1:1 ratio (Makowski et al., 2019a). It ranges from 50% to 100%, and a higher percent means there is a higher likelihood the effect exists whereas a smaller percent reflects more uncertainty. A $pd \leq 95\%$ corresponds to a two-sided $p > .1$ with a suggested interpretation of "uncertain" while a $pd > 99\%$ corresponds with $p < .01$, with a suggested interpretation of "probability existing" (Makowski et al., 2019a; Makowski et al., 2019b).

Whereas *pd* conveys the existence of an effect, the region of practical equivalence (ROPE) or Bayes Factors (BFs) express effect size. The ROPE is the proportion of the posterior distribution that fall within the range of values defined as the ROPE, frequently defined as the 89% CI [ROPE (*89%*)], the 95% CI [ROPE (*95%*)], or 100% of the posterior distribution [ROPE (*full*)] (Makowski et al., 2019a; Makowski et al., 2019b). Makowski et al. (2019b) recommended using ROPE (*full*) due to its greater consistency than ROPE (*95%*). The ROPE specifies an area around the mean (here -0.1 – 0.1) that is so small, that if a large portion of the posterior distribution fell within this range, the effect would be considered practically meaningless (Makowski et al., 2019b). ROPEs are affected by sample sizes when the effect is null (Makowski et al., 2019b), and they are sensitive to the scale of the predictor (Makowski et al., 2019a; Makowski et al., 2019b). BFs quantify the relative likelihood of an effect compared to a null model, with their values heavily influenced by the prior assumptions. ROPE (*full*) was used in the current study rather than BFs because BFs should be reported when informative priors are used while ROPE only reports on the posterior distribution and can be used when uninformative priors are used (Makowski et al., 2019b). The ROPE and *pd* were created using bayestestR (version 0.13.2; Makowski et al., 2019a)

To estimate the diagnostic accuracy of the dynamic assessment, I ran receiver operator characteristic curve (ROC) analyses and compared the area under the curve (AUC) across different predictors (STARD #14). An ROC provides an optimal number for a cut-off score based on sensitivity and specificity values; it is plotted as a y-axis of sensitivity (i.e., true positive rate) and an x-axis of 1 – specificity (i.e., the false positive rate) (Carter et al., 2016). The ROC was chosen over a discriminant analysis due to the small sample size. However, ROC analyses have the limitation of being highly sensitive to the variability in the current data set.

The calculation compares the child's performance on the dynamic assessment to their language status of DLD or TD as determined by the BESOS. No cases were indeterminate on the reference standard (STARD #15). The variables with the highest AUC were considered the most diagnostically informative. See Table 14 for a list of the a priori analyses for ROCs (STARD #17).

From the ROCs, sensitivity, specificity, and positive and negative likelihood ratios were calculated. Sensitivity was the proportion of children who had DLD and were accurately identified as having DLD by the dynamic assessment. Specificity was the proportion of children who were TD and identified as TD by the dynamic assessment. For a diagnostic assessment, sensitivity and specificity should be above .80 to be "fair" and above .90 to be considered "good" (Plante & Vance, 1994). A positive likelihood ratio (LR+) is the odds that a child with DLD was correctly identified as having DLD by the dynamic assessment. The negative likelihood ration (LR-) is the odds that a child who is TD was correctly identified as TD by the dynamic assessment. Likelihood ratio values range from 0 to 100 with an LR+ $\geq$ 10 and LR- $\leq$ .1 as indications that one can have high confidence that a disorder is present or not present, respectively (Dollaghan, 2007b). There was no pre-determined cut score for the dynamic assessment across any scoring option given its exploratory nature. Cut scores were determined by the variable(s) that were in the winning ROC analysis (STARD #12a).

Because there were multiple participant profiles, I compared the ROCs for the balanced and unbalanced bilinguals for the winning variable or combination of variables. If the ROC results were similar between each group and the full group, and if they were similar to each other, the analysis for the full group would be used to increase power. If they were distinct, the analyses would be run for the separate groups (E. Plante, personal communication, April 18, 2024).

Table 14

Planned ROC Analyses (STARD #17)

ROC	Compare To
Child's Performance (Scores)	
Teaching phase	
Immediate posttest	Teaching phase
1 hour posttest	IPT
1 day posttest	IPT, HPT
Combination of teaching phase and most informative posttest(s)	Best score (teaching phase or a posttest)
Modifiability Measures	
Short rating scale	
Long rating scale	Short rating scale
Rating scales combined	Best rating scale
Time in teaching phase	Best rating scale measure or combined rating scales
Criterion	Best modifiability measure so far
Combination of all best rating scale, time, and criterion	Each measure alone; if one part is less informative, remove that and compare to individual measures
[a]Combination of best rating scale and time	Best modifiability measure so far
[a]Combination of best rating scale and criterion	Best modifiability measure so far
Overall	
Combination of best child performance and modifiability measures	Best child performance scores and modifiability measures alone

Note. IPT = Immediate posttest, HPT = 1-hour posttest, DPT = 1-day posttest. Items were

[a]Comparisons were only made if component parts showed promising results.

Only one participant had missing data for the one-hour posttest due to the family needing to leave early. That case was included in every analysis other than the one-hour posttest analyses (STARD #16).

For each ROC (run with package pROC, version 1.18.2; Robin et al., 2011), I ran a Bayesian logistic regression (package brms; Bürkner, 2017) and a Random Forest analysis (package randomForest, version 4.7 – 1.1; Liaw & Wiener, 2002) for each predictor variable, and results are presented for both. Random Forest are non-parametric analyses that create multiple decision trees models based on bootstrapping from about two-thirds of the sample based on random sampling with replacement. The analysis uses the aggregate of the trees on real data (i.e., bagging) to "vote" for the most likely result for group membership (Breiman, 2001). The one-third of the dataset (i.e., out-of-bag data) that was not bootstrapped to make the decision tree models is then compared to the Random Forest's prediction. This step determines the classification accuracy of the model. The out-of-bag error is the proportion of the out-of-bag samples that were classified incorrectly. Random Forests avoid overfitting (Breiman, 2001), which is a benefit in the current analysis given that I was trying to reduce the dependence of ROC fit on the data. Bayesian logistic regression was also advantageous in its ability mitigate the ROC overfitting issue.

Because both Bayesian and Random Forest methods are based partially on real data and partially on data created as part of the modeling process, their results change slightly each time they are run. For consistent reporting and reproducibility, the results for each analysis were saved, so the exact models that are reported in this manuscript can be run in the future. Random Forest analyses were run multiple times for each predictor variable, and I chose the result that either (1) came up most frequently, or (2) came up first if the results did not yield a most frequent outcome. When the numbers were similar across both Bayesian and Random Forest analyses, confidence in the result was higher. When they were different, I interpreted the results as unstable, yielding lower confidence.

The ROC analyses provided the AUC, the threshold ("selected as the means between any two consecutive values observed in the data"; Robin et al., 2023, p. 73), sensitivity, and specificity. The threshold was in the units of the ROC input (i.e., log odds or fraction of votes for the outcome of TD). The cutoff score was determined by examining sensitivity and specificity for varying values of the optimal predictor and matching those values to the sensitivity and specificity for the predictor with the optimal AUC.

Results

Aim 1

The first aim was to create a replicable dynamic assessment procedure for evaluating the expressive morphology of Spanish-English speaking children ages 4;0 – 5;11 to classify them as having TD or DLD. I predicted that the procedures would be replicable given their scripted nature. Replicability was measured by fidelity to the dynamic assessment protocol and by inter-rater reliability scoring. Fidelity was scored for the pretest, teaching phase, and immediate posttest for the two trained assessors who administered the dynamic assessment. The pretest fidelity included administering every item, adding the morpheme correctly when administering the guessability check, scoring each item correctly, and administering the vocabulary check until the child correctly named ten items (0 = incorrectly or not done, 1 = fully correctly done). The teaching phase fidelity checklist included whether the assessor cued the child (N/A = not needed and not done, 0 = cued when not needed or not cued when needed, 1 = correctly cued when needed), whether the assessor got the child's attention before administering the enhanced recast (0 = did not get child's attention, 1 = got child's attention), and whether the assessor provided the correct response in the recast (0 = recast of correct response not provided, 1 = recast of correct response provided). The fidelity checklist for the immediate posttest monitored whether the assessor cued the child correctly and used the same scoring as the fidelity to cueing in the teaching phase. Fidelity was calculated as the total fidelity score divided by the total number of items that were not N/A.

All fidelity ratings were scored for all participants. See Table 15 for fidelity percentages. The average fidelity to the protocol was 95% and greater for all areas measured. Fidelity in the teaching phase was highest for providing a correct recast, followed by cueing appropriately, and then by getting the child's attention prior to providing the recast.

Table 15

Percent Fidelity to Dynamic Assessment Procedures

Fidelity Area	Mean	SD	Median	Mode	Min	Max
Pretest	95.56	8.56	100	100	80.00	100
Teaching phase						
Cueing	98.68	2.93	100	100	88.88	100
Attention	98.61	4.04	100	100	83.33	100
Correct recast	99.77	0.98	100	100	95.83	100
Immediate posttest						
Cueing	99.44	2.36	100	100	90.00	100
Total	98.41	3.77	100	100	80.00	100

Average interrater reliability for scoring across all components of the dynamic assessment was 99% (*SD* = 1.42%, median = 100%, mode = 100%, range = 50% -- 100%). Interrater reliability was calculated for the pretest vocabulary responses, the teaching phase, and for each posttest. For the teaching and posttests, interrater reliability was calculated for the child's initial response separately from their response post cue. The minimum of 50% was for a response post cue on the one-day posttest. The child only needed to be cued twice, and there was one disagreement, resulting in 50% accuracy for that scoring subsection.

Modifiability reliability was also scored (Table 16) for two of the three questions on the short modifiability scale (transfer was not given interrater scores), each with four possible response options (0 – 3), and the eight questions on the long modifiability scale, each with five possible response options (1 – 5). Across the 180 total datapoints (10 modifiability questions, 18 participants), point-by-point interrater agreement was 59.44% (*SD* = 49.4%, range by question = 44.44% - 77.78%). When scored across participants, modifiability ranged from 20% - 100%, with 42.50% agreement between me and assessor 1 (range = 20% - 100%) and 70.00% agreement between me and assessor 2 (range = 40% - 100%). Agreement by language ability status was 56.67% for children with DLD and 60.83% for children in the TD group.

Table 16

Interrater Reliability for Modifiability Scoring

Variable	Total Agreement[a]	Percent Inter-rater Reliability	*SD*	Assessor 1 *M*(*SD*) Range	Assessor 2 *M*(*SD*) Range	Author *M*(*SD*) Range
By Question						
Examiner Effort	10	55.56	51	2.62(0.52) 2-3	1.9(0.57) 1-3	1.83(0.62) 1-3
Child Responsivity	10	55.56	51	2.00(0.53) 1-3	0.6(0.97) 0-3	1.00(1.03) 0-3
Motivation	10	55.56	51	2.25(0.71) 1-3	2.0(0.67) 1-3	1.94(0.73) 1-3
Anxiety	11	61.11	50	2.12(0.35) 2-3	1.9(0.32) 1-2	1.83(0.62) 1-3
Persistence	12	66.67	49	1.75(0.46) 1-2	1.4(0.70) 1-3	1.94(0.73) 1-3
Task orientation	9	50.00	51	3.50(0.76) 2-4	4.4(0.97) 2-5	3.94(1.11) 2-5
Metacognition	10	55.56	51	3.62(0.92) 2-5	4.4(0.97) 2-5	4.11(1.08) 2-5
Nonverbal Self-Reward	13	72.22	46	1.88(0.64) 1-3	1.6(0.97) 1-4	1.50(0.62) 1-3
Attention	14	77.78	43	2.25(0.46) 2-3	2.2(0.79) 1-3	2.22(0.73) 1-4
Compliance	8	44.44	51	1.62(0.52) 1-2	1.8(0.63) 1-3	1.33(0.59) 1-3
Total	107	59.44	24.85			
By Participant						
Assessor 1				Assessor 1 Score	Assessor 2 Score	Author Score
14	7	70		23		24

9	10	100	18		18
13	5	50	24		19
11	2	20	25		29
17	3	30	26		20
15	3	30	26		20
2*	4	40	23		25
4*	3	30	24		20
Total (assessor 1)	37	46.25			
Assessor 2					
6	8	80		17	17
1	7	70		21	18
7	9	90		23	22
12	4	40		21	17
5	5	50		22	23
3	10	100		18	18
8*	7	70		22	22
16*	8	80		22	22
10*	6	60		28	23
18*	6	60		28	25
Total (assessor 2)	70	70.00			
By Language Ability Status			Assessor 1 *M*(*SD*) Range	Assessor 2 *M*(*SD*) Range	Author *M*(*SD*) Range
TD	73	60.83	23.67(3.01) 18-26	20.33(2.34) 17-23	20.42(3.55) 17-29
DLD	34	56.67	23.50(0.71) 23-24	25.00(3.46) 22-28	22.83(1.94) 20-25

[a]There were 18 participants, so the maximum total agreement score was 18 for each variable. The maximum total agreement score per participant was 10. *SD* reported in percentages.

*Asterisk indicates DLD group membership.

Aim 2

The second aim was to determine which elements of administration (teaching, modifiability scoring, posttest at varying time intervals) yielded the best differentiation between children in the DLD and TD groups. Given that this dynamic assessment used a more cognitively appropriate teaching method than past iterations (e.g., Roseberry & Connell, 1991), I predicted that the dynamic assessment would reach fair to good classification accuracy per the standards by Plante and Vance (1994). Regarding modifiability, I predicted that it would be diagnostically informative given that modifiability is one of the most diagnostically informative factors in other dynamic assessments (Hunt et al., 2022; Orellana et al., 2019). I also looked at the difference across immediate, one-hour, and one-day posttests, but I did not have a directional hypothesis regarding whether the different timepoints would change classification accuracy. I ran ROC analyses for all the pairings listed in Table 14. First, I will present summary statistics and plots. Then, I will present the ROC results. No adverse events occurred from performing the dynamic assessment or the reference standard (STARD #25).

Descriptive Statistics for Demographics and ROC Variables

Table 17 has group differences by demographic variables based on descriptive statistics and Bayesian *t*-tests, which have the assumption of independent groups and normal distribution of the outcome variable by group. All groups were independent, and distributions were checked for skewness (ideal = -1.00 – 1.00), kurtosis (ideal = -2.00 – 2.00), and through visual inspection of histograms. Bayesian modeling allows for models to be fit under a skew-normal distribution to account for nonnormality when running parametric tests that require a normal distribution assumption (Martin & Williams, 2017), so Bayesian *t*-tests were run with a skew-normal distribution. In the current sample, the DLD group was slightly younger than the TD group by about one month on average, though there was high uncertainty (*pd* = 72.58%) about the existence of a true between-group difference and a small cohen's *d*. As a proxy for SES, the mother's and father's highest levels of education were converted to numeric codes corresponding to the education level (e.g.,1 = “8^{th} grade or less”, 2 = “Between 8^{th} and 12^{th} grade”, 3 = “High school diploma”), which were added to create a combined parental education variable. SES was lower in the DLD group than the TD group, but the effect was uncertain (*pd* = 91.03%) though the effect sizes amounted to a significant, medium effect. Nonetheless, the 95% CI crossed over zero, indicating that the effect could be positive or negative, decreasing the likelihood that there was a true group difference in parental education. Likely the small sample size influenced these estimates. Gender was also examined for differences by language ability group, but a χ^2 test could not be used because multiple cells had fewer than five observations, violating a required assumption for χ^2 tests. The ratio for the TD:DLD groups was 2.5 for boys and 1.75 for girls.

Table 17

Language Ability Group Differences by Demographic Variables

Variable	TD (n = 12) $M(SD)$	DLD (n = 6) $M(SD)$	Median estimate	95% CI	*pd*	ROPE (*full*)	Cohen's *d*	Interpretation
Age[a]	4;8(0;6)	4;6(0;7)	-0;2	-0;8, 0;4	72.58%	2.44%	0.24	Uncertain, probably significant, small effect
Parent Education[b]	9.92(3.42)	7.5(3.33)	-2.27	-5.67, 1.23	91.03%	1.74%	0.62	Uncertain, probably significant, medium effect
Gender	*n*	*n*						
Boy	5	2						
Girl	7	4						

[a]The *t*-test for age was calculated using age in months, and it was converted to years;months format for readability in the current table. TD age in months $M(SD)$ = 55.92(6.32), and DLD age in months $M(SD)$ = 5 4.33(6.95). In months, the median estimate = -1.61, 95%CI [-7.85, 4.31].

[b]Parent Education was the combination of mother's and father's highest level of education and so was treated as a continuous variable. Each category was given a number and subsequently added. In the two cases when the father's level of education was not reported, the mother's level of education was used.

Table 18 reports group differences on the language measures. As expected, the DLD group had significantly lower scores on the reference standard (i.e., BESOS) than the TD group. Of the three language sample measures, the MLU-W_{Eng} was the most likely candidate for demonstrating group differences with a probably existing, significant, large effect when codeswitched utterances were included. Because not every participant spoke enough Spanish for their Spanish language sample to be analyzed, the MLU-W_{Sp} analyses were only run with ten participants (codeswitching included) and seven participants' data (codeswitching excluded), so confidence in the result for MLU-W_{Sp} was low. The PGU did not show differences between the TD and DLD groups.

For the child scores, none of the variables could be run in a *t*-test because there was no variance in the DLD group (all scored 0 for teaching phase and posttests). Thus, a Bayesian Chi-square analysis was attempted to look at the likelihood of scoring 0 versus scoring 1 or higher. However, the assumption of 5 or more participants per cell was violated, so the test could not be run. For modifiability, the long scale alone and the combined long and short scales (modifiability—full) both yielded effects that probably existed (i.e., means were different for DLD and TD groups) and were considered significant, large effects. For the modifiability scores, in these descriptive statistics and in the ROCs, the consensus scores were used when available (i.e., three initial participants per assessor), and assessors' scores were used for all other participants.

Table 18

Language Ability Group Differences on Language Measures Based on Bayesian t tests

Language Measures	TD (*n* = 12)				DLD (*n* = 6)				Median estimate	95% CI	*pd*	ROPE (*full*)	Cohen's *d*	Interpretation
	M	***SD***	**Skew**	**Kurtosis**	***M***	***SD***	**Skew**	**Kurtosis**						
BESOS score	99.42	9.07	-0.38	-1.58	76.83	6.79	-0.17	-2.09	-23.61	-30.81, -14.91	100%	0%	2.63	Certainly existing, significant, large
Language Samples[a] Including CS														
PGU	55.43	11.66	0.04	-1.70	47.75	21.68	0.72	-1.27	-10.71	-24.80, 5.82	92.40%	0.38%	0.67	Uncertain existence, significant, medium effect
MLU-W_{Eng}	6.03	0.86	-1.72	2.41	4.65	0.94	0.04	-2.13	-1.18	-2.12, -0.31	99.12%	0.55%	1.30	Probably existing, significant, large effect
MLU-W_{Sp}	5.70	1.66	0.85	-0.86	4.52	0.36	-0.38	-2.33	-0.82	-3.34, 0.96	82.61%	6.14%	0.61	Uncertain existence, undecided significance, medium effect
Excluding CS														
PGU	51.59	14.59	-0.21	-0.91	47.94	22.72	0.26	0.56	-1.61	-22.09, 15.76	67.17%	0.85%	0.200	Uncertain of effect existing, significant, small effect
MLU-W_{Eng}	5.54	1.06	-1.06	-0.42	3.99	1.20	0.26	-1.76	-1.24	-2.54, -0.07	97.78%	1.10%	1.08	Likely existing effect, probably significant, large effect
MLU-W_{Sp}	6.09	1.94	0.27	-1.80	4.43	0.38	0.00	-2.75	-1.54	-5.39, 1.80	82.66%	3.25%	0.77	Uncertain of effect existing,

Language Measures	TD (n = 12)				DLD (n = 6)				Median estimate	95% CI	*pd*	ROPE (*full*)	Cohen's *d*	Interpretation
	M	*SD*	Skew	Kurtosis	*M*	*SD*	Skew	Kurtosis						
														undecided significance, medium effect
Dynamic Assessment														
Child scores[b]														
Teaching phase score	8.17	12.23	1.70	2.01	0	0	N/A	N/A						
Immediate posttest	4.25	6.18	1.32	0.72	0	0	N/A	N/A						
One-hour posttest	3.55	6.65	1.40	0.58	0	0	N/A	N/A						
One-day posttest	3.08	6.29	1.67	1.54	0	0	N/A	N/A						
Modifiability Measures[c]														
Time in teaching phase	18.33	2.10	0.08	-1.19	18.67	1.75	-0.14	-1.52	0.33	-1.87, 2.44	62.05%	7.00%	0.14	Uncertain existence, undecided significance, small effect
Modifiability scale (short)	4.33	1.92	0.48	-1.12	2.33	1.51	0.71	-1.15	-1.70	-3.60, -0.13	98.29%	1.04%	0.94	Likely existing, probably significant, large effect
Modifiability scale (long)	18.25	3.57	-0.93	-0.59	22.67	3.14	0.32	-1.85	4.91	1.81, 7.90	99.81%q	0.03%	1.44	Probably existing, significant, large effect
Modifiability scale (full)	24.92	4.85	-1.32	0.20	31.33	3.93	-0.29	-1.42	6.85	2.99, 10.66	99.87%	0.05%	1.61	Probably existing, significant, large effect

Note. *M* = Mean, *SD* = Standard Deviation, TD = typically developing language, DLD = Developmental Language Disorder, 95% CI = 95% credible interval, *pd* = probability of direction, ROPE (*full*) = region of practical equivalence for the full posterior distribution. Interpretation based on *pd*, ROPE (*full*), and cohen's *d* based on Makowski et al. (2019a), Makowski et al. (2019b), and Cohen (1977b), respectively.

[a]The language sample measure *M* and *SD* reported here are the raw values (i.e., not z-scores). "Including CS" = analyses included utterances with codeswitches. "Excluding CS" = analyses excluded code-switched utterances. MLU-W_{Eng} analyses were run on the TD group (n = 11) and DLD group (n = 5) for those who had an English language sample. These numbers were the same regardless of including or excluding codeswitching. For MLU-W_{Sp} including codeswitching, TD group n = 7, DLD group n = 3. For the MLU-W_{Sp} excluding codeswitching, TD group n = 5, DLD group n = 2.

[b]Because the child scores (teaching and posttests) did not have a distribution for the DLD groups, no *t*-tests could be run.

[c]The modifiability scales in this table represent the scores based on consensus (three for each of the assessors) and then the assessors' scores subsequently as this was the data used in the original ROC models.

ROC Results

See Table 19 for a summary of the ROC results. The child score Bayesian logistic regression models (i.e., teaching phase and posttests) had some difficulty with convergence, likely because there was no variance for the DLD group's scores. Every child with DLD scored 0 on the teaching phase and posttests. Those models were rerun with increased iterations until they demonstrated convergence. Of the child scores (teaching, immediate posttest, one-hour posttest, one-day posttest), the teaching phase yielded the same diagnostic accuracy as the combination of the teaching phase and immediate posttest scores. For parsimony, the teaching phase alone was favored (Figure 3). The AUC was different for the Bayesian logistic regression (AUC = .833) and Random Forest (AUC = .667) AUCs, but their sensitivity (100%) and specificity (66.67%) scores were the same. For modifiability, the criterion analysis could not be run because only one child reached the predetermined criterion. The best modifiability AUC was for the combination of the two modifiability scales (Figure 4), which was superior to the combination of those and the teaching time. To combine the modifiability scales, the short scale was reverse scored to run from 1 – 4 (rather than 0 – 3) in the same direction as the large modifiability scale. The Bayesian logistic regression and Random Forest AUCs for the full modifiability scale were similar, and their sensitivity measures were the same (83.3%) with differing specificity results (Bayesian specificity = 100%, Random Forest specificity = 83.3%).

Table 19

ROC Results for All Variables Tested

Variable for ROC	Group	Bayesian ROC						Random Forest ROC						
		AUC	Threshold	Sensitivity	Specificity	LR+	LR-	AUC	Threshold	Sensitivity	Specificity	LR+	LR-	OOB Error Rate
Child Scores														
Teaching phase score	full group	.83	0.40	1.00	.67	3.00	0.00	.67	0.60	1.00	.67	3.00	0.00	22.22
IPT	full group	.75	0.33	1.00	.50	2.00	0.00	.58	0.60	1.00	.58	2.40	0.00	66.67
HPT	full group	.63	0.24	1.00	.25	1.33	0.00	.73	0.80	1.00	.67	3.00	0.00	35.29
DPT	full group	.63	0.23	1.00	.25	1.33	0.00	.75	0.77	1.00	.75	4.00	0.00	33.33
Combo of teaching + IPT	full group	.83	0.43	1.00	.67	3.00	0.00	.67	0.60	1.00	.67	3.00	0.00	22.22
Modifiability Measures														
Modifiability scale 1 (short w/o transfer)	full group	.80	0.47	.67	.83	4.00	0.40	.57	0.20	.33	1.00	N/A	0.67	33.33
Modifiability scale 1 (short w/ transfer)	full group	.81	0.48	.67	.83	4.00	0.40	.53	0.23	.33	1.00	N/A	0.67	33.33
Modifiability scale 2 (long)	full group	.83	0.34	.83	.67	2.50	0.25	.58	0.90	.67	.75	2.67	0.44	38.89
Time in teaching phase	full group	.55	0.31	.83	.33	1.25	0.50	1.00	0.80	1.00	1.00	N/A	0.00	55.56
Modifiability (full)	full group	.88	0.55	.83	1.00	N/A	0.17	.83	0.90	.83	.83	5.00	0.20	11.11
Modifiability (full) + time (in teaching phase)	full group	.83	0.52	.67	.92	8.00	0.36	.82	0.95	.83	.75	3.33	0.22	16.67
Modifiability (full) + teaching scores	full group	.94	0.60	.83	1.00	N/A	0.17	.87	0.20	.83	1.00	N/A	0.17	11.11
Modifiability (full) + teaching scores	Balanced group	1.00	0.64	1.00	1.00	N/A	0.00	.88	0.10	.75	1.00	N/A	0.25	25.00
Modifiability (full) + teaching scores	Un-balanced group	1.00	0.57	1.00	1.00	N/A	0.00	1.00	0.50	1.00	1.00	N/A	0.00	0.00
Modifiability (full) + teaching scores	Assessor ratings[a]	.88	0.65	.67	1.00	N/A	0.33	.86	0.70	.83	.92	10.00	0.18	16.67
Modifiability (full) + teaching scores	My ratings[a]	.65	0.51	.33	1.00	N/A	0.67	.85	0.90	.83	1.00	N/A	0.17	44.44

Note. ROC = receiver operator characteristic (curve), AUC = Area Under the Curve, LR+ = positive likelihood ratio, LR- = negative likelihood ratio, OOB error rate = out-of-bag error rate, expressed as a percentage. N/A ratings for LR+ occurred when the denominator was zero due to a specificity of 1.00. IPT = immediate posttest, HPT = 1-hour posttest, DPT = 1-day posttest.

[a]Assessor ratings and my ratings denote the respective group's ratings of modifiability scores. These ROCs were run with the full dataset. When unspecified (i.e., for all models above these), the ratings are the first three scores for each assessor, which were rated independently by me and the assessor and then given a consensus rating, and then the assessors' scores after those.

Figure 3

ROCs for Teaching Scores

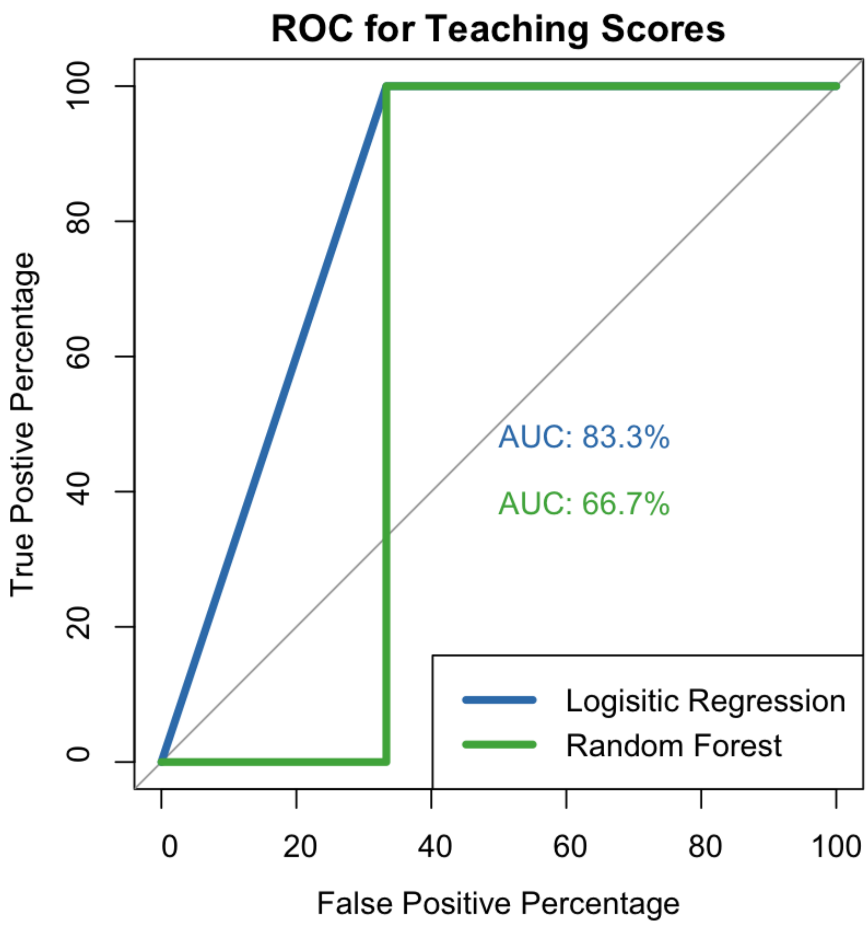


Figure 4

ROCs for Combined Modifiability Scales

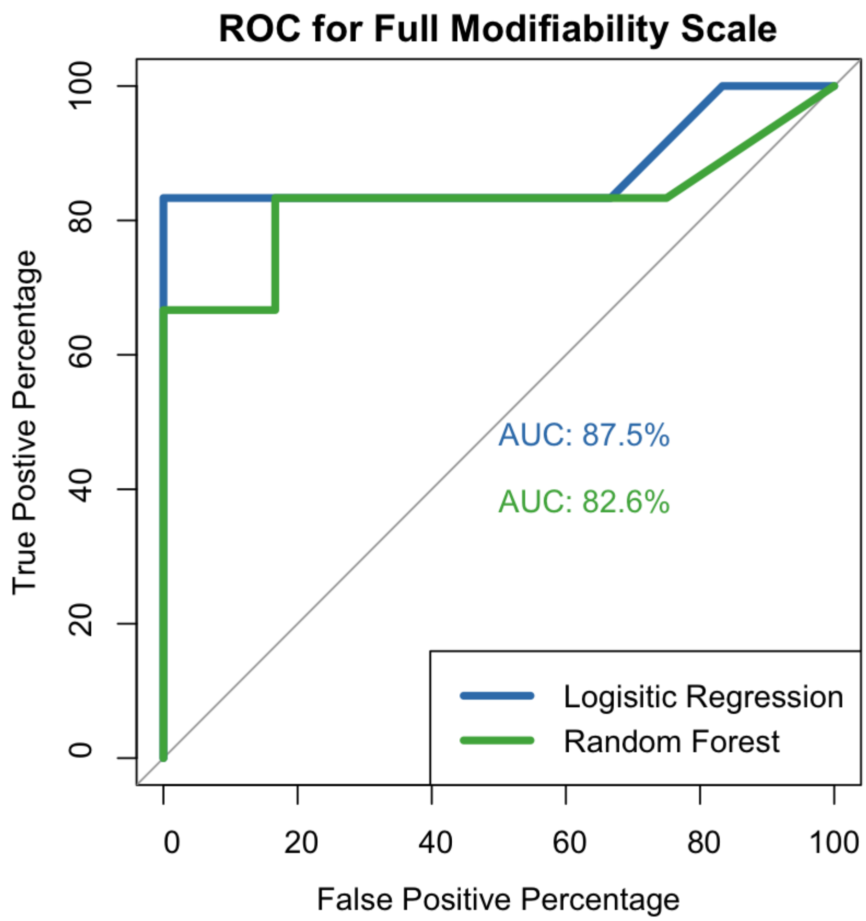
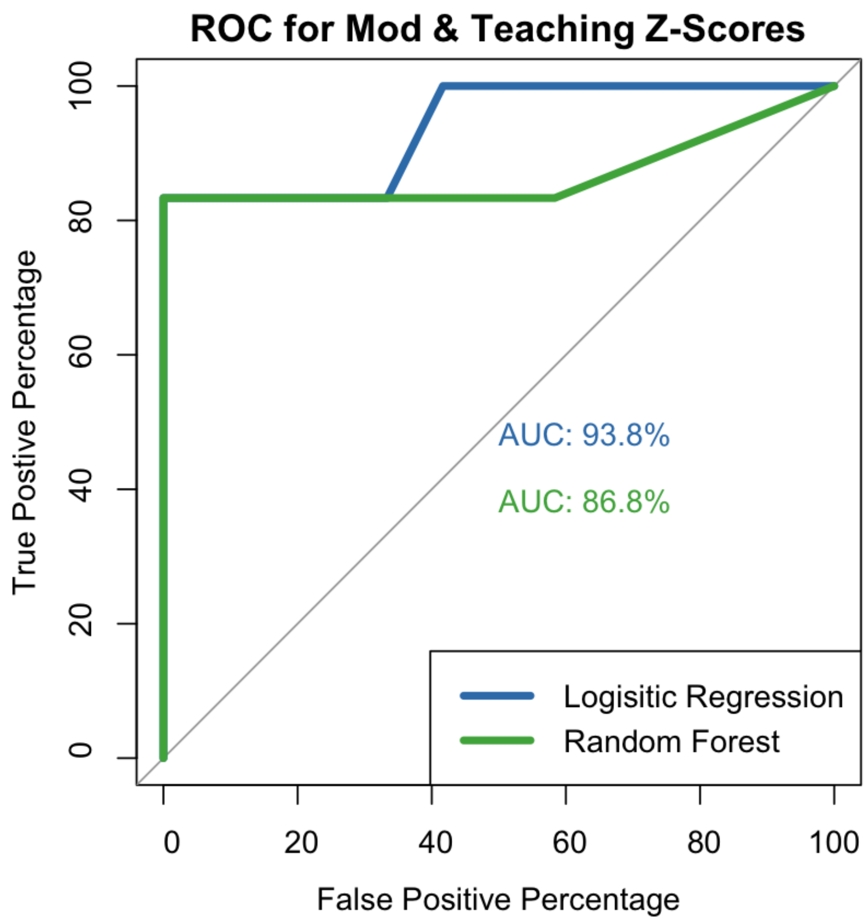


Figure 5

ROCs for Combined Modifiability and Teaching Scores



Finally, the best variables of the child scores and of the modifiability measures were combined (i.e., teaching phase + combined modifiability scales). Because the teaching scores and modifiability scales were on different scales, I converted both variables into z-scores. Additionally, the variables went in opposite directions. As teaching phase scores increased, the child was more likely to be TD. Originally, as modifiability scores increased, the child was more likely to be DLD. To address this difference in directionality, I multiplied all the modifiability scores by -1 to ensure they were going in the same direction as the teaching phase scores. Thus, as z-scores increased, the child was more likely to be classified as TD. The analysis revealed the most favorable AUCs of all the analyses, and they were similar across both ROC types (Bayesian AUC = .938, Random Forest AUC = .868; Figure 5). Additionally, the sensitivity (83.3%) and specificity (100%) were the same across the Bayesian and Random Forest ROCs. To ensure this analysis was not biased against either of the language use profiles (i.e., balanced or unbalanced Spanish-English use), the winning analysis (i.e., z-scores of combined teaching phase and full modifiability scales) was run on each group separately and compared to the full group. Results were comparable across balanced and unbalanced bilinguals and between each bilingual use type and the full group (see Table 19, Figure 6 and Figure 7). These results should be viewed as preliminary because there were small samples within each subgroup increasing the likelihood of the ROC overfitting to these groups.

Figure 6

ROCs for Combined Modifiability and Teaching Scores of Balanced Bilingual Group

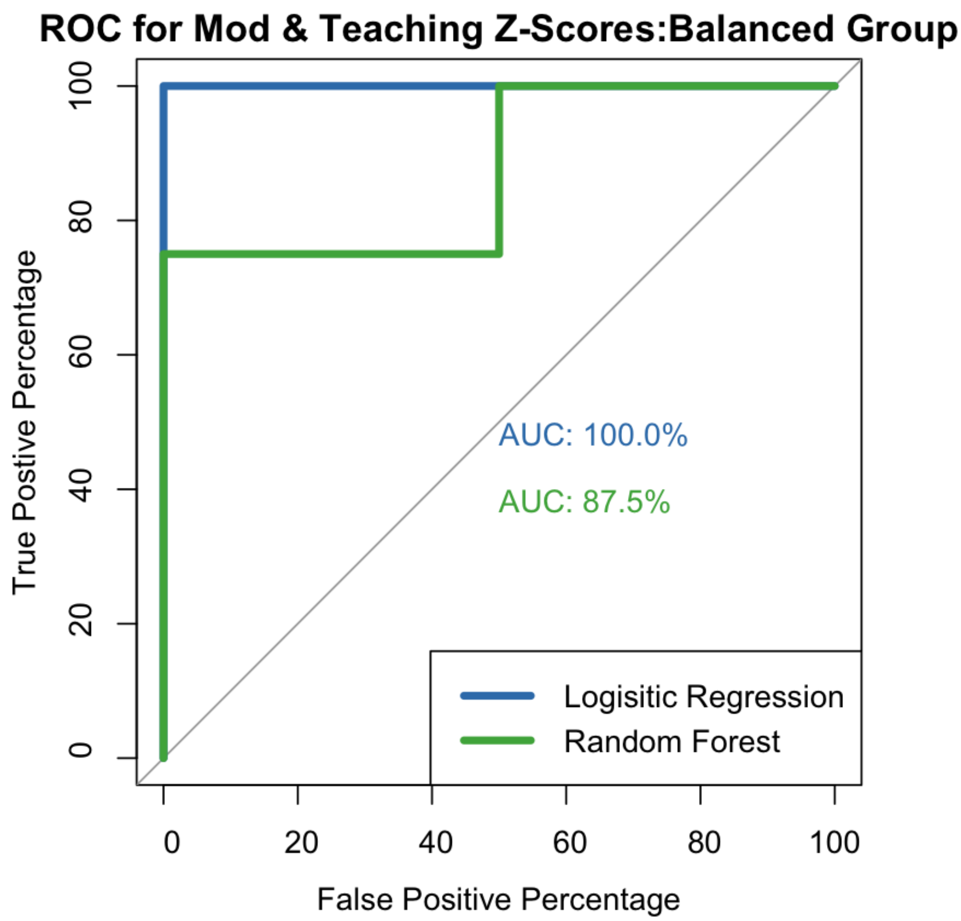
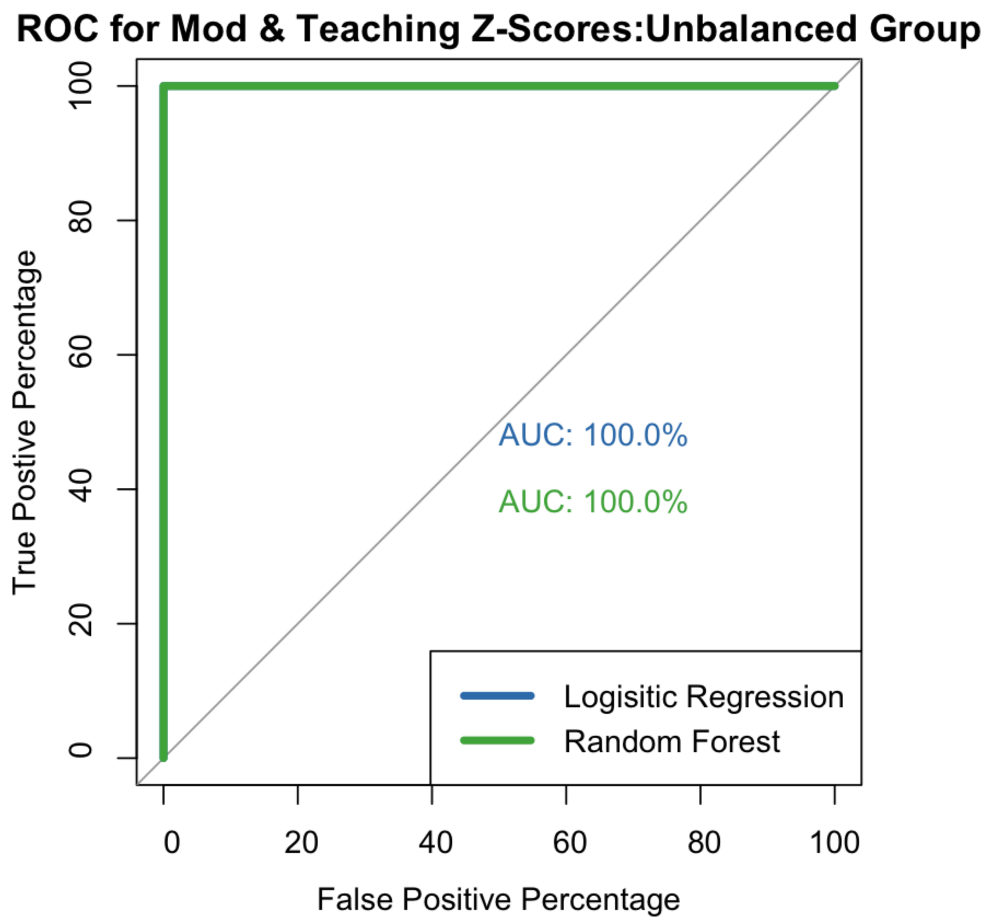


Figure 7

ROCs for Combined Modifiability and Teaching Scores for Unbalanced Bilingual Group



Group Differences by Winning ROC Variables

A Bayesian *t*-test determined the effect size of the language ability group differences for the winning combination of the modifiability and teaching scores.

Figure 9 shows that the two groups were not normally distributed per visual analysis, and Table 20 shows the TD group had a positive skew. Thus, the skew-normal distribution was used to run the *t*-tests. The beta estimate for the intercept (the TD group) was 0.70, and the beta estimate for the DLD group was -1.58. Because the units were in z-scores, this can be interpreted as the DLD group scoring approximately 1.5 standard deviations below the mean score. The CI did not cross zero for the DLD group (Median = -1.54, 95%CI [-3.02, -0.37]), and can be interpreted as a 95% probability that the DLD group would score between -3 and -.3 SDs below the mean. The DLD group mean had a *pd* of 99.30% for a negative effect (i.e., DLD < TD). Thus, the current difference between TD and DLD groups on the combination of the teaching scores and modifiability scales can be interpreted as probably existing (Figure 9). Additionally, I wanted to run a chi-square analysis to determine the likelihood of getting a teaching score > 0 by language ability group, but the frequency table had multiple cells with frequencies less than 5, violating an assumption of the test.

Figure 8

Histogram of Combined Modifiability & Teaching Scores

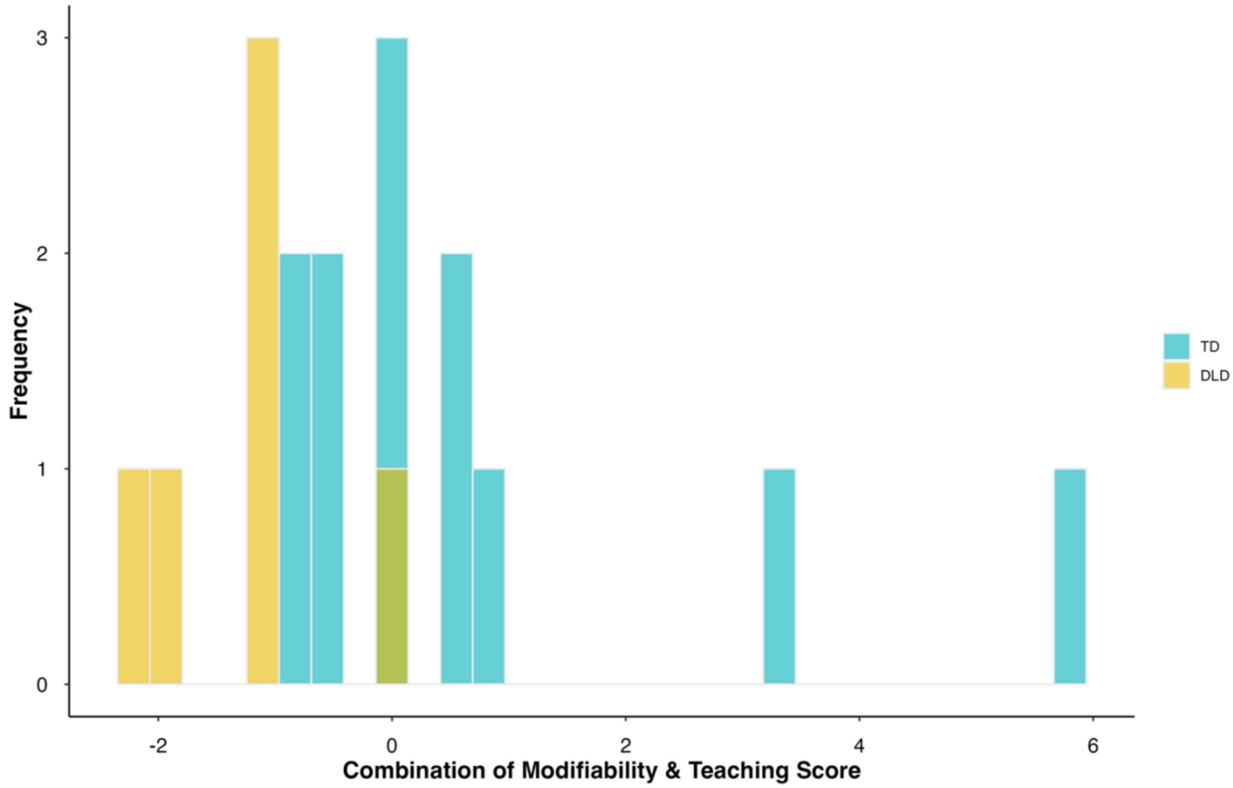
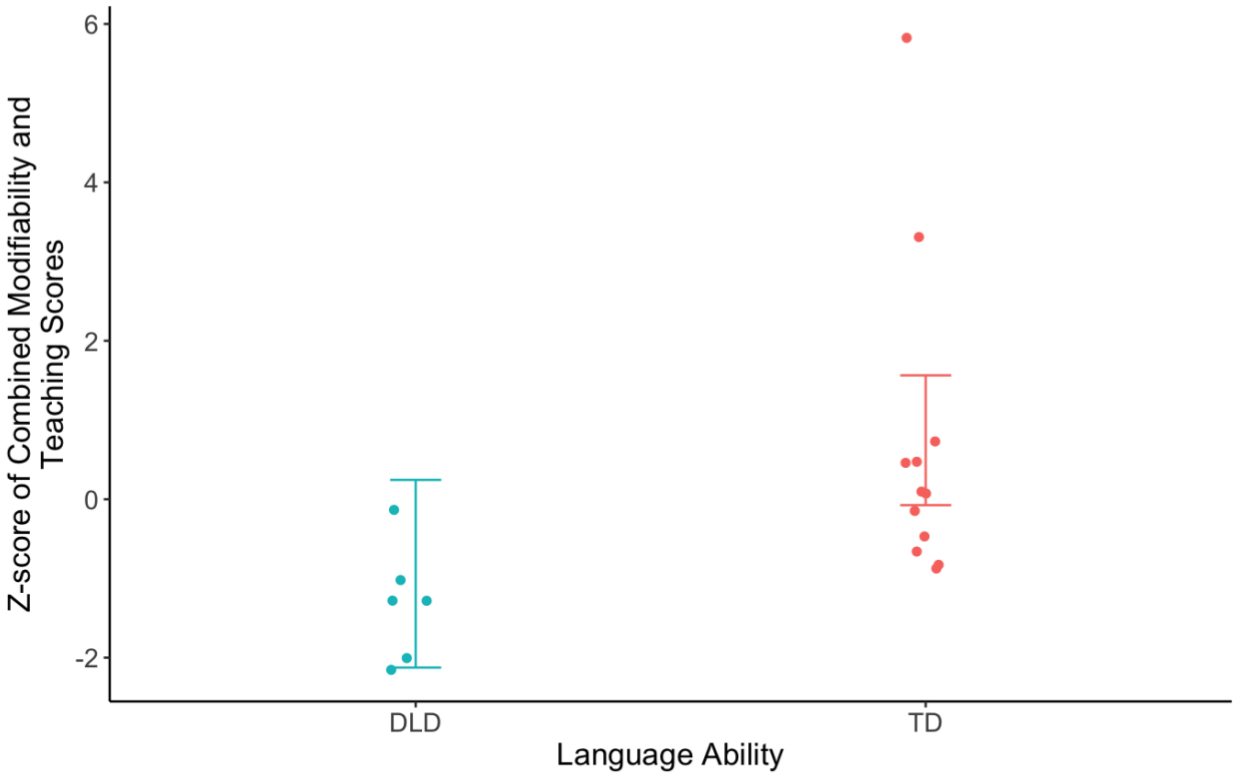


Figure 9

Differences Between Predicted Mean Scores by Language Ability Group



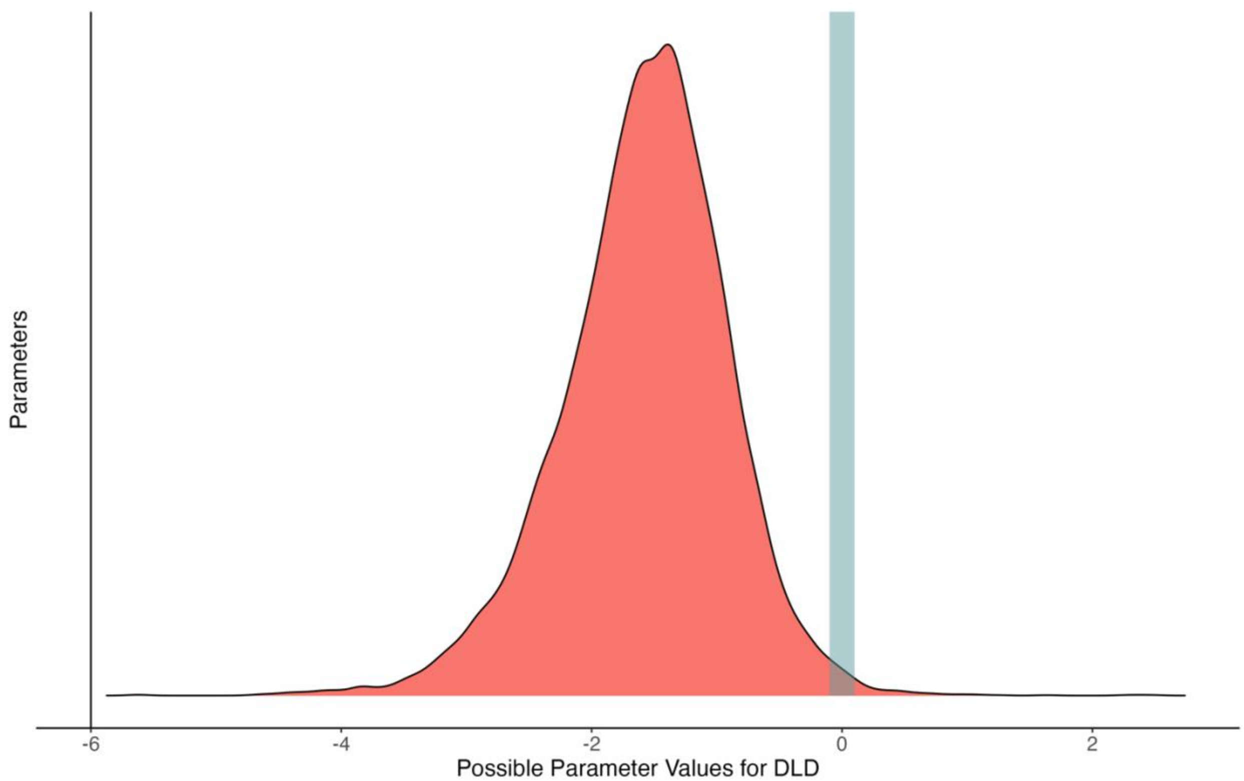
Note. The dots indicate actual data points from the sample. The error bars are the estimates for the posterior distribution from the lower to upper bounds of the uncertainty interval based on the Bayesian *t*-test.

In the current analysis, the ROPE (*full*) is on a standardized scale because the combined outcome of teaching scores and modifiability scale scores were converted to *z*-scores when combined. Based on the suggested interpretation that a ROPE <1% is significant (Makowski et al., 2019a; Makowski et al., 2019b), the current effect was considered practically significant (0.57% in the ROPE (*full*); Figure 10). The cohen's *d* effect size was calculated as the difference between group means divided by the standard deviation, and *d* = 1.04, which is a large effect size

(Cohen, 1977a). Additionally, $R^2 = 0.1811$, meaning approximately 18% of the variance was accounted for by the model.

Figure 10

Region of Practice Equivalence (ROPE) for 100% of Posterior Distribution for Combined Modifiability and Teaching Scores



Note. The vertical blue bar represents the area tested for the ROPE (-01. – 0.1) while the pink distribution represents the predicted posterior distribution for children with DLD given the model and the data.

Table 20

Group and Demographic Differences on the Winning ROC Variable (Full Modifiability and Teaching Scores Combined)

TD (*n* = 12)				**DLD (*n* = 6)**				**Mdn Estimate**	**95% CI**	***pd***	**ROPE (*full*)**	**Cohen's *d***	**Interpretation**
M	***SD***	**Skew**	**Kurtosis**	***M***	***SD***	**Skew**	**Kurtosis**						
0.65	1.98	1.58	1.29	-1.30	0.72	0.29	-1.42	-1.54	-3.01, -0.37	99.30	0.57%	1.04	Probably existing, significant, large effect
Balanced (*n* = 8)				Unbalanced (*n* =10)									
0.78	2.54	0.94	-0.75	-0.63	0.9	-0.45	-1.22	-0.82	-2.55, 0.59	89.00	5.40%	0.48	Uncertain, undecided significance, medium effect
Boy (*n* = 7)				Girl (*n* = 11)									
0.46	2.71	1.52	0.60	-0.29	1.54	0.88	0.09	-0.76	-2.10, 0.78	85.86	5.42%	0.41	Uncertain, undecided significance, medium effect

Note. *M* = Mean, *SD* = Standard Deviation, TD = typically developing language, DLD = Developmental Language Disorder, 95% CI = 95% credible interval, *pd* = probability of direction expressed as a percentage, ROPE (*full*) = region of practical equivalence for the full posterior distribution. Interpretation based on *pd*, ROPE (*full*), and cohen's *d* based on Makowski et al. (2019a), Makowski et al. (2019b), and Cohen (1977b), respectively. Balanced = balanced bilingual, Unbalanced = unbalanced bilingual based on current language use survey.

Other variables were examined to determine their relationship to the winning combination variable, particularly to see if any variables revealed group differences or trends that could bias their scores (see Table 20) .Based on a Bayesian *t*-test, balanced and unbalanced language use groups had a CI that crossed over zero and a *pd* and ROPE (*full*) that could be interpreted as uncertain of the effect (but see the medium effect cohen's *d*). Gender had a similar result, revealing uncertainty about group differences. Other comparisons were between two continuous variables, so Pearson's correlation coefficients were computed using a Frequentist approach (Table 21). There were no significant correlations between the combined modifiability and teaching scores and parent level of education (SES proxy), child age in months, and the MLU-W English. The other language sample measures (MLU-W Spanish, PGU) were excluded from the current correlations because they did not seem to reflect group differences (see Table 18) and were therefore less informative as a continuous language measure.

Table 21

Means, Standard Deviations, and Correlations with Confidence Intervals

Variable	*M*	*SD*	1	2	3
1.Modifiability (full) + teaching scores	0.00	1.90			
2. Parent education	9.11	3.50	.10 [-.39, .54]		
3. Child age in months	55.39	6.37	.12 [-.37, .56]	.12 [-.37, .55]	
4. MLU-W_{Eng}[a]	5.60	1.08	.38 [-.15, .73]	.09 [-.43, .56]	.41 [-.11, .75]

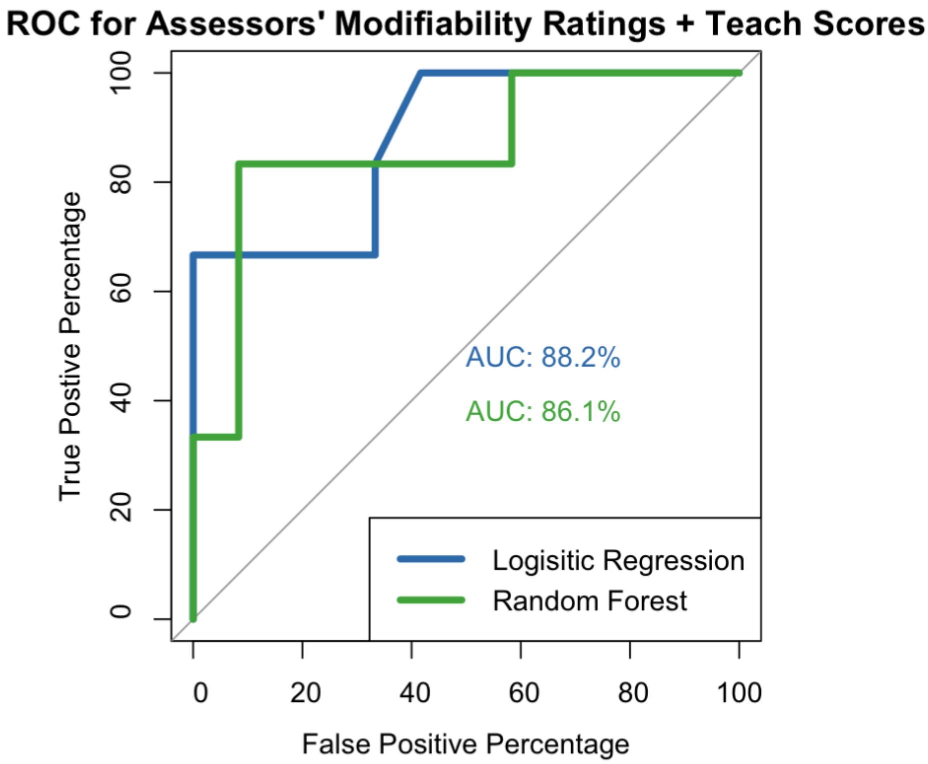
Note. M and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). * indicates $p < .05$. ** indicates $p < .01$.
[a]The language sample measures that are reported have codeswitched utterances included.

Winning ROC Reliability by Modifiability Rater

Because the modifiability interrater reliability was low and one of the important components of the final ROC was modifiability, the ROCs were run for the combined teaching and modifiability z-score variable across the different raters (last two rows of Table 19). The first ROC included the initial scores from both assessors. This differed from the original combined variable because, in the original, the consensus score for the first three participants was used along with the assessors' scores for the rest of the participants who did not receive consensus scoring. The second ROC was run with just my modifiability scores. Using the modifiability scores from the assessors (Figure 11) resulted in a decrease in AUC (.882) in sensitivity (66.67%) for the Bayesian analysis. For the Random Forest analysis, the AUC was similar to the original ROC, but its specificity declined to 91.7% while sensitivity remained at 83.3%. This discrepancy across the two types of modeling reduces confidence in the results.

Figure 11

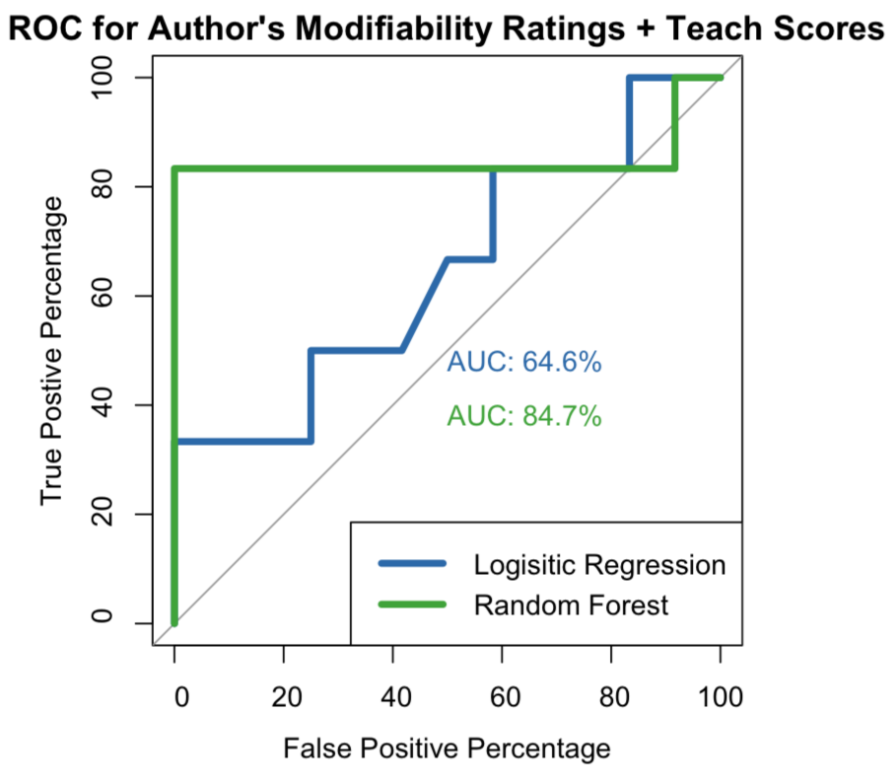
ROC for Assessors' Modifiability Ratings Combined with Teaching Phase Scores



My scores (Figure 12) performed differently than the assessors' scores and the original model on the Bayesian logistic regression model. While specificity was maintained at 100%, my sensitivity was 33.33% (AUC = .65, original model AUC = .94). The Random Forest ROC, though, maintained the same sensitivity (83.33%) and specificity (100%) as the original model with a similar AUC (.85, original model AUC = .87). This large discrepancy across the Bayesian and Random Forest ROCs reduces confidence in the results.

Figure 12

ROC for My Modifiability Ratings Combined with Teaching Phase Scores



Cut Scores

The threshold scores suggested by the ROCs were presented in the units of the logistic regression or Random Forest, which were not clinically meaningful. I based the clinically meaningful cut score on the winning ROC of the combined teaching scores and full modifiability scale scores (original model) with the goal of matching its sensitivity (83%) and specificity (100%) to a real score. Because all children with DLD scored a 0 on the teaching phase (sensitivity = 100%, specificity = 67% at score of 0), I derived the cut score from an assumed teaching phase score of 0 and a full modifiability scale score. See Table 22 for the list that was considered. The cut score with sensitivity of 83% and specificity of 100% was 29, meaning that children with a teaching score of 0 and a full modifiability score of 29 and below would be considered TD. Children with a teaching score of 0 and a full modifiability score of 30 or above would be considered DLD.

Table 22

Sensitivity and Specificity by Cut Score from the Full Modifiability Measure

Cut Score	True Negative	False Positive	True Positive	False Negative	Sensitivity	Specificity
14	1	11	6	0	1.00	0.08
15	1	11	6	0	1.00	0.08
16	2	10	6	0	1.00	0.17
17	2	10	6	0	1.00	0.17
18	2	10	6	0	1.00	0.17
19	2	10	6	0	1.00	0.17
20	2	10	6	0	1.00	0.17
21	2	10	6	0	1.00	0.17
22	2	10	6	0	1.00	0.17
23	2	10	6	0	1.00	0.17
24	2	10	6	0	1.00	0.17
25	4	8	5	1	0.83	0.33
26	7	5	5	1	0.83	0.58
27	8	4	5	1	0.83	0.67
28	10	2	5	1	0.83	0.83
29	**12**	**0**	**5**	**1**	**0.83**	**1.00**
30	12	0	4	2	0.67	1.00
31	12	0	2	4	0.33	1.00
32	12	0	2	4	0.33	1.00
33	12	0	2	4	0.33	1.00
34	12	0	2	4	0.33	1.00
35	12	0	1	5	0.17	1.00
36	12	0	1	5	0.17	1.00
≥ 37	12	0	0	6	0.00	1.00

Note. The cut scores were considered the given score or lower would indicate TD and any score above the cut score would indicate DLD. True positives were children with DLD correctly classified as DLD, and true negatives were children with TD correctly classified as TD. False positives were children with TD incorrectly classified as DLD, and false negatives were children with DLD incorrectly classified as TD. The bolded row is the ideal cut score that matches the sensitivity and specificity from the winning AUCs.

Discussion

The purpose of the current study was to create to create a replicable dynamic assessment procedure for evaluating the expressive morphology of Spanish-English speaking preschool age children suspected of having DLD (Aim 1). This was done by modifying a previously created dynamic assessment to assess morpheme learning in English with the invented morpheme /-əb/ (Roseberry & Connell, 1991). The original was modified to have the teaching phase incorporate statistical learning principles by increasing the regularity (and dose) of the target morpheme and the variability of the linguistic and visual stimuli regarding the target morpheme. Additionally, the teaching phase was made more interactive by adding enhanced conversational recasts (ECR), which allowed for the assessor to evaluate the child's modifiability on two modifiability rating scales. Objective modifiability components—criterion and time on the teaching phase—were also evaluated. Finally, the posttest was administered at multiple timepoints to ascertain the effect of timing on classification accuracy given that a posttest one hour or one day after the teaching phase would place additional constraints on the child's memory.

Aim 1

For aim 1, the hypothesis was that the procedures would be replicable given their scripted nature. This was borne out in the results. The mean percent fidelity to the dynamic administration was 98.41% (*SD* = 3.77, range = 80.00% - 100%), and interrater reliability was also high (*M* = 99%, *SD* = 1.42%). While Roseberry and Connell (1991) did not report fidelity to the procedures, they did report high interrater reliability for scoring (99.76%). Other replications had similarly high interrater reliability scoring rates of 99%, 100%, and 99.04% for Boyer and Martin (2012), Hwa-Froelich and Matsuo (2005), and Restrepo (1998), respectively. Kohnert and Danahy (2007) did not report reliability. Additionally, Boyer and Martin (2012) reported

100% fidelity to their procedures, which was scored for one-fifth of the sessions. No others reported fidelity. This high reliability and fidelity to the scoring and procedure demonstrate that the dynamic assessment is replicable, likely due to the level of scriptedness and operationalization of scoring criteria.

For modifiability, reliability between me and the assessors was low, 59.44%, when scored live for all participants. Of Roseberry and Connell (1991) and its replications, only Hwa-Froelich and Matsuo (2005) had a modifiability measure. They reported 93% interrater reliability, which was scored for one-fifth of the samples (i.e., 15 data points total). Differences between the current sample's reliability and their reliability interrater agreement may be due to the scales themselves and the nature of the tasks. Both Hwa-Froelich and Matsuo (2005) and the current study rated child responsivity, examiner effort, and transfer based on the scales in Peña (2000), and the current study had the addition of most of the items from the scales in Peña et al. (2007). The addition of more items created more possibility for disagreement, potentially resulting in the lower reliability. Additionally, Hwa-Froelich and Matsuo (2005) scored the child's modifiability after the child had completed multiple different dynamic assessment tasks, including the invented morpheme and learning nonwords. Their nonword-learning task may have provided more process information than the invented morpheme rule in their (or the current) study as it involved having the child repeat the nonword labels and providing cues to the child if they could not label the nonword object. Thus, they may have had qualitatively different opportunities to assess modifiability, which may have leant themselves to higher reliability in modifiability scoring.

Other types of dynamic assessment studies have seen higher reliability than in the current study, as well. For example, on a dynamic assessment of word learning, Petersen et al. (2020)

had an interrater reliability of 86% (range = 75% - 93%) on their full modifiability index (i.e., 7 questions each that could be scored 0, 1, or 2) for 30% of the dynamic assessment recordings. Their scale was created by their research team and was similar to the one used for a narrative-based dynamic assessment in Petersen et al. (2017), which had a mean interrater agreement of 85% (range = 14% = 100%, median and mode = 100%) for 38% of the dynamic assessment recordings. The modifiability scale from Petersen et al. (2017) was researcher-created and drew upon previously created modifiability scales (Peña et al., 2006, 2007; Petersen & Spencer, 2014; Ukrainetz, et al., 2000 as cited by Petersen et al., 2017). All questions could be scored as a 0, 1, or 2 with operationalized definitions for each level. In these studies, though, it was unclear whether the second raters were blinded to the child's diagnostic status. If they were not blinded at the time of conducting the interrater checks, their judgments may have been biased. Regardless, each study had higher interrater reliability than the current study, suggesting that higher agreement is possible. Potential variables contributing to this discrepancy in reliability include the nature of the task administration, the scales themselves—their score ranges, their items, and their operationalization, as well as the assessors—their training and scoring biases. In the rest of this section, I will explore each of these variables.

The nature of the dynamic assessment task may impact how much process information is gleaned and therefore how reliably modifiability is scored. For a dynamic assessment of narrative ability, Peña et al. (2006) did not report inter-rater reliability, but Peña et al. (2007), which used data from Peña et al. (2006), reported high consistency across the child's first and second session (r^2 = .99). Similarly, Petersen et al. (2017) had high reliability on their modifiability scores for a dynamic assessment targeting narrative ability. Word-learning tasks show similarly high reliability on modifiability measures (Hwa-Froelich & Matsuo, 2005; Peña, 2000; Petersen et al., 2020). It may be that explicit teaching procedures increase the reliability of modifiability because they often involve multiple prompts after which the child can respond and improve (or not) as opposed to the single, minimally informative, ECR for each item provided in the current study after which the child may repeat (some did) but is not required to do anything until the next item. Alternatively, the explanatory variable may be the number of sessions—each of these had at least two sessions (either of the dynamic assessment or of data collection overall), so they may have had more information about the child in general before rating the child's modifiability, increasing the accuracy of the ratings. In contrast, the current study's dynamic assessment always occurred on day 1 to reduce the likelihood of prior information influencing modifiability judgments.

In Peña (2000), which used a word-learning dynamic assessment, modifiability reliability was not reported because consensus scoring was used, and there was high internal consistency of the modifiability measure's scores (alpha coefficient = .83, .85 for sessions 1 and 2, respectively). The scoring range was four possible score choices per question on 2/3 items on the modifiability scale and three possible scores on the learning strategies checklist and on the third item on the modifiability scale. Similarly, Petersen et al. (2017) and Petersen et al. (2020) used only three scoring options, and Hwa-Froelich and Matsuo (2005) only used a 4-point span scoring range and saw high reliability. Interrater reliability may increase for the dynamic assessment in the current study if the scoring range is reduced (e.g., only three possible response choices versus five). In the current study, the use of typically only three or four of the scores on the five-point scale supports this notion that assessors may not need such fine-grained detail as a five-point scale when scoring modifiability. Future research would need to explore this possibility.

Operationalization may have played a role in the low modifiability in the current study, as well. The scoring operationalization was taken directly from the definitions for each score in Peña (2000) and Peña et al. (2007). Some definitions may have benefitted from modification to suit the dynamic assessment in the current study more aptly, which may have resulted in higher interrater reliability. For example, the scoring definitions for the question of "compliance" were "1 = cooperative," "2 = insecure," "3 = hesitant," "4 = uncooperative," and "5 = refusing." This question had the lowest level of interrater agreement, potentially due to the difficulty applying "insecure" or "hesitant" labels to the current dynamic assessment situation. Anecdotal evidence from during data collection supports this idea that further operationalization specific to the study may have improved interrater agreement, because the assessors often spoke with me about what to score for compliance, commenting, for example, that "insecure" did not fit the participant, but a score of "2" did. Another example comes from the awareness of errors ("metacognition"), which was difficult to score unless the child engaged in spontaneous self-talk or gave verbal or gestural indications that they understood they had been in error after an ECR was provided. The modifiability categories were not always the best fit for the task, and future research can consider tailoring them to potentially improve interrater reliability.

Training may have also contributed to lower reliability on the modifiability scales. We did practice multiple example scenarios (at least three each) and a few practice participants as a team and discussed the modifiability ratings for each, and we came to a consensus after independently scoring modifiability for each of the first three participants with whom each assessor worked. This was similar to other dynamic assessment study trainings who had graduate students administer the dynamic assessments. Petersen et al. (2017), for instance, trained their graduate and undergraduate research assistants by having them administer five practice sessions with their peers of the dynamic assessment and modifiability scoring, receiving feedback from Petersen (modifiability interrater agreement = 85%). Peña (2000) trained five bilingual graduate students in Speech, Language, and Hearing Sciences by having them watch video examples and receiving online feedback from Peña. Additionally, consensus discussion was used for any discrepancies on the rating scales until 90% agreement was met. They reported an internal consistency alpha of .83 and .85 for the two administration time points. Peña et al. (2006) also trained graduate clinicians by video tapes and having them practice "until the clinicians were familiar and confident with the procedure" (p. 11). Although the training procedures in the current study were similar to other studies that had high inter-rater reliability for modifiability scores, it may have been that the assessors in the current study needed more practice administering the dynamic assessment and scoring modifiability with more children before working with the children in the current sample. In the future, training may need to include checking practice modifiability ratings for reliability with my own judgments and discussing the results, and the assessors may be required to reach a certain level of interrater reliability on modifiability scoring before going into the field.

Lower modifiability may also have resulted from the difference in experience. The assessors were undergraduate seniors in the Department of Speech, Language, and Hearing Sciences while I am a certified SLP with four years of work experience (and additional research experience) with bilingual preschoolers as a bilingually trained SLP. It may be that our perceptions of modifiability differed by our experience levels. One assessor had much lower interrater reliability on modifiability scoring than the other, and there was not a substantial difference between the interrater reliability agreement for the TD and DLD groups, underscoring the potential variability in scoring due to experience and perspective. It may also be that the lived experience of being the assessor is different from being the live observer, resulting in different ratings. For example, if a child is having difficulty paying attention, the assessor would have been working hard to keep their attention while I would have been observing, taking notes, and at times helping with behavior management as needed. The assessor did the bulk of the behavior management in that interaction and therefore might have rated the child as being more distractible than I would have. This explanation was somewhat supported by the data as the assessors provided higher modifiability ratings by group (TD M = 23.50, SD = 3.10; DLD M = 24.50, SD = 2.81) than I did (TD M = 20.42, SD = 3.55; DLD M = 22.83, SD = 1.94), but we did have overlapping score ranges. It may also be that I had some bias in my ratings because, while the assessors were blinded to the child's DLD status, I had some information about the child from their caregiver questionnaire. However, none of the children in the DLD group had a prior diagnosis of DLD, though some were in SLP services for speech difficulties. In future studies, this limitation could be fixed by having a blinded second rater.

The reliability of the scoring on the teaching phase and the high fidelity to administration procedures bodes well for translating this dynamic assessment to real-world environments. However, the low reliability on the modifiability scoring is an area of limitation. It is of particular concern because modifiability was an important component to establishing the desired classification accuracy. Future research should engage more assessors at varying levels of experience, varying levels of bilingual training, and for monolingual and bilingual clinicians to help establish which factors are most important for improving reliability in modifiability judgements. Additionally, tailoring the scale to the task with well-constructed operational definitions for each score choice and reducing the scoring range may improve interrater reliability in future iterations. Item Response Theory (IRT) analyses could be used to simultaneously scale the raters' responses and gauge the stability of the items on the assessment. Ideally, I would have conducted an IRT analysis for the current dynamic assessment, but that would have required a much larger sample size. A different tack would be to explore other modifiability scales, including potentially creating a fully new one to fit this dynamic assessment. Again, it would need to be validated through IRT before claims could be made about its utility. Finally, it will be important to establish what should be considered acceptable training requirements and reliability thresholds for modifiability that both research assistants and clinicians should meet before administering the current dynamic assessment to children.

Aim 2

Aim 2 was to determine which elements of administration (i.e., teaching scores, posttest at varying time intervals, modifiability scoring) yielded the best differentiation between children with DLD and typical language. Across all child scores (teaching phase scores, immediate posttest, 1-hour posttest, and 1-day posttest), the teaching phase score had the highest AUC and sensitivity (100%) and specificity (67%). Among the modifiability measures (time in teaching phase, criterion of three correct items in a row in the teaching phase, short modifiability scale, long modifiability scale), the combination of the short and long modifiability scales (i.e., full modifiability) had the best classification accuracy (sensitivity = 83%, specificity = 100% and 83% for Bayesian logistic regression and Random Forest, respectively). The highest classification accuracy was for the combination of the modifiability and teaching phase scores (sensitivity = 83%, specificity = 100%). This sensitivity and specificity outperformed the classification accuracy of the original Roseberry and Connell (1991; sensitivity = 77%, specificity = 92%) and its replications that tested classification accuracy--Restrepo (1998;

sensitivity = 65.22%, specificity = 60.87%) and Boyer and Martin (2012; sensitivity = 71%, specificity = 57%).

Hypothesis 1 predicted that modifiability measures would increase the classification accuracy of the dynamic assessment, which was confirmed by the modifiability full scale alone and when the modifiability full scale was combined with the teaching phase scores. Hypothesis 2 was non-directional, exploring whether adding a posttest at varying time intervals would differentiate better than an immediate posttest alone. None of the posttests had sufficient classification accuracy alone, and the immediate posttest, which had the best classification accuracy of the posttests, did not improve upon the classification accuracy of the teaching phase when they were combined. This demonstrated that the posttests were poor indicators of learning potential, such that future iterations of this dynamic assessment could be run with the teaching phase and no posttests. However, without any posttests, the question of transfer would have to be removed from the modifiability scale because there would be no opportunity to demonstrate transfer of the skill to a new set of items. This could change the classification accuracy of the modifiability scale, which future research would need to assess.

Classification accuracy reached acceptable levels for the whole participant pool, and group differences were negligible. The lack of correlation by age and parent education and the lack of group difference by gender and language use (balanced or unbalanced) showed that this dynamic assessment has promise to be used across the range of Spanish-English bilinguals in the 4;0 – 5;11 age range across SES groups. Testing with a larger sample and more sophisticated analyses (e.g., IRT) are needed to confirm a lack of bias. The fact that there were no differences by age was surprising given that some of the past iterations did see improved performance for older children (Boyer & Martin, 2012; Hwa-Froelich & Matsuo, 2005) but see Kohnert and Danahy

(2007). The relationship between age and dynamic assessment scores was not reported in Restrepo (1998) or Roseberry and Connell (1991). However, Roseberry and Connell (1991) had a higher mean age and range (5;7, range = 4;5 – 6;4) than all other studies (including the current one M = 4;7, range = 4;0 – 5;7) except for Restrepo (1998), and they had some of the highest rates of learning. Thus, it may be that there is an effect of age, but the current study's range was too restricted to demonstrate it.

Zooming in on the four TD children who did not learn the invented morpheme, I looked for relationships among the children and for variables that may have explained the reason for why they did not learn the morpheme like the rest of the TD group. The first evidence to examine was the modifiability ratings as they may have tapped into some factor that could explain why these children did not productively use /-əb/. For example, if all these children had high anxiety and low motivation, those traits could have impeded their readiness to learn. Recall that a higher modifiability score is more likely to be in DLD group, and the cutoff for DLD was > 29. Three of these TD children had scores near the modifiability cutoff (scores of 28 and 29), and the fourth child had a score just below the median for the TD group (median = 26, child's score = 25). Their relatively high modifiability scores may have been in part due to a redundancy. The lack of learning was demonstrated in two of the modifiability questions—task orientation (Completely understands task = 1, Doesn't understand tasks = 5) and metacognition (Aware of all errors = 1, Unaware of any errors = 5). These two questions had the highest sets of scores (4's and 5's) while most other questions had 1's and 2's for this group. The question of motivation also trended higher with two children scoring 3's. Without a clear trend in the modifiability scores, the scale provides little evidence as to why these children did not learn the invented morpheme.

Another hypothesis would be that the children who did not learn the invented morpheme had relatively low morphosyntactic skills but not low enough to be classified as DLD. The BESOS combined scores did not pattern with their modifiability scores, nor were they the lowest of the TD group for three of the children (one child scored 85, which is the BESOS cutoff score). Two of the four children were at or above the median PGU (including codeswitched utterances) for the TD group (median = 55.35). Language scores do not show a clear explanation for why these children did not demonstrate knowledge of /-əb/.

Absent language ability discrepancies, perhaps demographic variables revealed a certain profile of child struggled with this dynamic assessment. Demographic variables also failed to show an explanatory pattern for why these children did not learn the invented morpheme—two were boys and two girls. Only one child was younger than the median age of the TD group, and three of the children had higher levels of parent education than the median for the TD group.

The only variable that related to the low performance on the teaching phase for the four TD children was their modifiability scores, and even still, it was not a perfect relationship as one child's modifiability score was below the median. Given the absence of within-child factors to explain their scores of zero in the teaching phase, the reason remains unclear. It is possible that the parameters of the dynamic assessment could be improved (e.g., dose rate, variability) to facilitate learning for TD children.

Similarly, it is important to focus on the child with DLD who did not learn the invented morpheme but whose modifiability score was 25, which was lower than the TD median of 26 (lower = more TD-like, higher = more DLD-like). Looking at their demographic variables, this child's parent education score (3; mean = 7.5, range = 4 - 12) might predict a higher modifiability score if modifiability ratings were subject to an SES bias, but their low (i.e., TD-like) modifiability score contradicted this prediction. This child did have a slightly higher BESOS combined score and substantially higher PGU than the median for the DLD group (BESOS score = 82, median = 78.5, range = 69 – 84; PGU score = 85.70%, PGU median = 39.94%, range = 30.4% - 85.7%), but their MLU-W_{Sp} was at the median for the DLD group (MLU-W_{Sp} score = 4.73 , median = 4.73, range = 4.11 - 4.73).Thus, this child's language abilities were on the higher end of the DLD group, which may have explained their low, TD-like, modifiability score. Another possibility is that this child was misclassified as DLD by the reference standard. Although the BESOS has high classification accuracy (sensitivity = 90%, specificity = 91%), there is potential for error. The PGU score alone is an anomaly suggesting TD language as it was the highest of the entire sample (TD PGU median = 55.35%). However, the child used relatively short utterances, so there were fewer opportunities for errors within each utterance. Nonetheless, their MLU-W_{Sp} was within the MLU-W_{Sp} range for the TD group (range = 4.17 - 8.89), suggesting the possibility of misclassification. This child had also been in SLP services in the past, so they may have learned the routine of therapeutic interactions and therefore been scored relatively low on the modifiability scale. Finally, this participant was older than the DLD group median (child's age = 5;0, median = 4;5), which may have influenced a lower rating on the modifiability scale (e.g., longer attention, higher persistence). Multiple potential reasons could have led to this child having a modifiability score that put them in the TD range. Future research should examine similar cases to find converging evidence.

Classification Accuracy and Dynamic Assessment Features

The potential reasons for the current dynamic assessment's improvement on classification accuracy over and above Roseberry and Connell (1991), Boyer and Martin (2012), and Restrepo

(1998) may be due to several changes in the current dynamic assessment including its change to the morpheme itself and its inclusion of statistical learning.

First, the morpheme was potentially more salient with the addition of the consonant at the end. It could be that /-əb/ patterns after other suffixes in English better as a /-VC/ structure (e.g., /- ɪŋ/, /- əd/) so it was more likely to be recognized as a bound morpheme and added into existing schemas for bound morpheme use. The task in the current study provided the participants a relatively highly phonetically salient morpheme if salience is operationalized as length (i.e., a syllable was added) and with a concrete, imageable concept—a partial object. Per mapping theory—the theory that children with DLD most easily learn structures high phonetic and perceptual salience (Chiat, 2001)—this should have been a relatively easier skill for the children with DLD to learn. However, every child with DLD in the current study failed to demonstrate learning of the invented morpheme. The phonological and perceptual salience of the invented morpheme in the current study may have boosted learning for the TD group, but it did not support the DLD group's learning enough to interfere with differentiating across the groups. This was a benefit in the current study. Future studies of invented morpheme learning dynamic assessments should continue to have mapping theory in mind when designing stimuli as the current study cannot provide evidence as to the relative importance of phonological and perceptual salience for learning by children with DLD.

Statistical Learning. The second major change in the dynamic assessment was the introduction of statistical learning principles—regularity and variability. Statistical Learning Theory leads to the hypothesis that the invented morpheme in the current dynamic assessment would be learned well by TD children because there was (1) high regularity of the target (always NOUN+əb) and (2) high variability around the target (Plante & Gómez, 2018). Additionally, this dynamic assessment was different than many artificial grammar learning tasks of statistical learning because the morpheme was given semantic information through the images and script, which may have facilitated learning. None of the children in the DLD group learned the target morpheme while eight children in the TD group did learn it. In the following subsections, I will explore how statistical learning may have contributed to the TD group learning and the DLD group's and four participants in the TD group's failure to learn.

Statistical Learning: Regularity. The high regularity of the target morpheme was primarily manifested through the dose. In the current study, each child heard 24 doses before they started on the enhanced recasting portion of the teaching phase, which included another 24 doses. Having 48 total doses more than doubled the 10 -- 20 doses in the teaching phases used by Roseberry and Connell (1991) and its replications. Although increasing the dose may have improved learning opportunities, even still, four TD child did not demonstrate learning of the invented morpheme.

One might question whether dose rate (also called dose density), the number of doses per a given unit of time, greatly differed between the current study and previous iterations. Alt et al. (2020) summarized the dose densities for multiple late-talker interventions, which ranged from 1.55 – 5.16 words/minute, increasing with their own study to approximately 9 words/minute. The high dose rate, as opposed to dose number per target, was the key variable that spurred learning in their statistical learning-based word-learning intervention for late-talkers (Alt et al., 2020). Although dose rate may not need to be 9 doses/minute for preschoolers learning morphemes, especially TD preschoolers, the principle of high dose rate likely remains necessary to spark learning.

The optimal dose rate for TD children in this context, though, has not been explored. Only general timing was documented in Roseberry and Connell (1991). Their sessions (for pre-test, teaching, and posttest) were 15 to 20 minutes long. It is likely that the pretest and posttest took up at least five minutes of that time, so teaching time may have been up to 10 - 15 minutes, though likely shorter given the short script. Dose density would have been 4 doses/minute, 2 doses/minute, or 1.33 doses/minute at 5 minutes, 10 minutes, or 15 minutes of teaching time, respectively. The current study averaged about 18 minutes of teaching time, yielding a dose density of 2.67 doses/minute.

Although this is relatively low, the focused stimulation phase had a much higher dose density than the overall dose rate. Anecdotally, the focused stimulation phases took approximately 4 minutes, which would be a dose density of 6 doses/minute, which was likely similar to or higher than the dose density in Roseberry and Connell (1991). It may be that 6 doses/minute is too low for all TD children to learn the target morpheme. However, the target dose rate of 9 words/minute comes from literature for younger children (i.e., late-talking toddlers) who presumably have poorer attention and thereby might need more doses to counteract the potential lack of attention to some doses. Preschoolers may have better attention than toddlers and therefore can learn from a lower dose density. They also have more language than late-talking toddlers, so they have existing schema to draw upon and add to, which might make language learning more efficient at this age, again negating the need for a dose rate of 9 doses/minute.

Children with DLD also need a higher dose number than their TD peers to learn linguistic targets (e.g., Gray, 2004; Rice et al., 1994), so TD children may need only about half the target dose rate for children with DLD (e.g., only 4.5 doses/minute). If this were the case, the dose density in the focused stimulation phase (i.e., 6 doses/minute) would have satisfied that criteria. Even still, 4.5 doses/minute may not be needed—at least in the ECR phase—because studies of ECR intervention with children with DLD present 24 doses in a 30-minute session (i.e., 0.8 doses/minute; (i.e., 0.8 doses/minute; e.g., Plante et al., 2019; Plante et al., 2014). This may be because the ECR procedure requires the preschooler to engage in effortful retrieval whereas the toddlers in studies by Alt et al. (e.g., 2021; 2020) put minimal pressure on the child to respond. The enhanced recasting portion of the current dynamic assessment (14 minutes on average) had a dose density of 1.71 doses/minute, higher than in Plante's ECR studies. Because the dose density surpassed typical dose densities used for children with delayed and disordered language, dose rate seems a less likely candidate for why some TD children failed to learn the morpheme. It may be that they indeed did learn, but the measure was not sensitive enough to capture their learning.

One step the current study took to show emergent understanding of the invented morpheme was to allow variants of the invented morpheme to be counted for partial credit as long as the child's suffix was not a bound morpheme that occurred in English (e.g., -er was not given partial credit). Many children put their own version of the morpheme on, which may have signaled their awareness of the presence of a unique bound morpheme but that they had not learned the phonological sequence of it yet. Another potential explanation is that they were trying to follow the instruction of creating a "new word," perhaps with the thought that they had to come up with their own new word rather than repeat the assessor's new word. Others placed the invented morpheme at the beginning of the word (e.g., "ebcup"). One guess for why some children used a word-initial placement for /-eb/ was that they were so focused on the morpheme that they did not want to forget it, so the added it in first. The children were also asked if they noticed anything about how the assessor said the words in the story, but responses were minimally informative about their understanding or lack thereof. Most children did not know how to respond, providing a non-sequitur or vague answer. The lack of awareness of learning the invented morpheme is relatively unsurprising given that the participants are of an age when metalinguistic awareness is still developing and given that many participants in statistical learning tasks do not demonstrate explicit knowledge of what they learned (e.g., Reber, 1967). The partial credit option and final question about what they noticed gave children multiple opportunities to demonstrate their learning of the morpheme in different ways. There may be other options to explore in future studies.

Children in Hwa-Froelich and Matsuo (2005) demonstrated learning of the invented morpheme at a much higher rate on their receptive posttest than on their expressive posttest. One consideration would be to add a receptive posttest to the current dynamic assessment in the future to better capture any knowledge children might have of the invented morpheme as it might be easier for children to demonstrate receptive understanding than to have to produce the invented morpheme expressively. The caveat is that receptive tests may be easily guessable. If the child is asked to point to a picture of a NOUN+əb in an array of four images, the child has a 25% probability of being correct by chance. Moreover, children—likely both the TD and DLD groups—might pick the partial object image by virtue of having seen 48 partial objects immediately prior in the teaching phase. In other words, there would be a confound in the receptive test, such that responses may not reflect phonological knowledge of the invented morpheme. Moreover, it seems that the receptive learning that Hwa-Froelich and Matsuo (2005) demonstrated may have reflected vocabulary knowledge more than morpheme knowledge as the children were shown twenty picture pairs that had not been trained “and the children were asked in English to point to the word-po [their invented morpheme was -po] for each pair” (p. 236). Thus, while a receptive posttest cannot be ruled out as a potential way for children to demonstrate their knowledge, there is reason to be skeptical that its inclusion would greatly improve the classification accuracy of the current dynamic assessment.

A different measure might be more sensitive to capturing emerging knowledge about the invented morpheme—spontaneous repetition. In an enhanced conversational recast intervention for children with DLD, Nicholas et al. (2021) found that children who spontaneously repeated the clinician’s recast were likely to show treatment gains about three days after beginning the repetition. Although this is promising evidence that repetitions demonstrate some morpheme learning, the study was with children with DLD in an extended intervention. The question is if TD children would exhibit a similar behavior in the one-session learning context of the dynamic assessment. Some children did repeat the clinician after the recasts in the current dynamic assessment study. A future analysis could be to examine whether and to what extent children repeated the clinician and whether that can improve classification accuracy when combined with the teaching phase score. If there is initial promise from the current data, it could be worth exploring adding the presence of a repetition to the score sheet in future studies.

Beside a high dose number, the regularity principle of statistical learning also dictates that counterexamples should be minimal if at all included. Although the script, pictures, and procedures in the current study attempted to preclude any counterexamples, it may be that unintentional counterexamples were presented. The participants saw the whole and partial form of each platform word side-by-side in the ECR portion of the teaching phase. The assessor was supposed to point to the whole form when saying the whole-form word and then to the partial visual form when cueing the child for a response and when providing the ECR. Although I did not take data on the assessors' pointing accuracy per the procedure, I observed that they pointed to the whole image when providing the recast on a few occasions. This may have functioned as a counterexample, potentially disrupting learning. Alternatively, in the guessability check, the child heard the invented morpheme ten times, and each time they saw four possible visual correlates. If the child paired the invented morpheme with an incorrect visual correlate in the guessability check, they may have had to overcome that self-created counterexample when the teaching phase began.

Statistical Learning: Variability. Variability of the non-target stimuli (linguistic and visual) may have been another key to promoting learning by the TD children in this study. Previous iterations of the invented morpheme dynamic assessment had low variability by always presenting the same script across doses (e.g., “This is a NOUN. This is a NOUNə.”) and presenting the same type of cut at the same angle on for all visual stimuli. Only about 55 – 65% of TD children learned the morpheme in this context (Boyer & Martin, 2012; Hwa-Froelich & Matsuo, 2005; Kohnert & Danahy, 2007) with the exception of the 92% of children in the original study by Roseberry and Connell (1991). In the current study, 67% of participants in the TD group learned the invented morpheme. While this was better than some of the original studies, all of which had low variation, it was only slightly better than the others and not better than Roseberry and Connell (1991). It is likely that the high dose was helpful. However, there may have been too much variability, which could have prevented some TD children from learning the invented morpheme.

Having both the varied language around the target morpheme and the varied visual stimuli may have overloaded the TD learners. This hypothesis may be less likely because a body of statistical learning literature has demonstrated the facilitative effect of variability in basic research (e.g., Apfelbaum & McMurray, 2011; Gómez, 2002; Twomey et al., 2017) and in intervention conditions (e.g., Alt et al., 2021; Alt et al., 2020; Meyers-Denman & Plante, 2016; Plante et al., 2019; Plante et al., 2014). Other studies have shown a lack of facilitative effect of variability, though, in certain conditions. Broedelet et al. (2023) found no difference in learning for low or high contextual diversity conditions for their TD or DLD groups of 7 – 9-year-olds. In a study comparing monolingual and bilingual cross-situational word-learning, Crespo et al. (2023) found that bilinguals showed an advantage over monolinguals in conditions of two sources of variability—multiple exemplars and multiple speakers. However, they showed weaker word-learning when there were multiple exemplars alone, which the authors could not explain. Additionally, all children performed similarly in the no-variability condition as in the variability conditions. A study by Höhle et al. (2020) showed that 14-month-olds did not benefit from visual variability of exemplars when learning novel minimal pair words, though they did benefit from multiple speakers. Finally, Nicholas et al. (2019) found that monolingual, English-speaking TD 3- to 5-year-olds did not learn prepositions better in a high variability condition (object and linguistic label) than in a low variability condition (only object variability). The current study provided both exemplar variability and visual variability, which cannot be directly compared to any of these studies. However, there is a possibility that the high visual variability or the interplay of high visual and linguistic variability might have resulted in difficulty learning the morpheme for some TD children.

Modifiability and Classification Accuracy

The biggest roadblock to continuing this dynamic assessment is the lack of interrater reliability on the modifiability scales. When the consensus scores were not used, and it was the assessors' alone, the classification accuracy dropped, specifically the sensitivity on the Bayesian logistic regression ROC (67%) and the specificity of the Random Forest ROC (92%) decreased compared to the winning model. The Random Forest ROC model, though, had good and fair sensitivity and specificity, respectively (Plante & Vance, 1994). When my modifiability scores were used, sensitivity also plummeted on the Bayesian logistic regression ROCs (33%), though the Random Forest ROC model for my scores yielded the same sensitivity and specificity as the original model. Notably, the out-of-bag (OOB) error for the Random Forest ROC for my scores was 44.44% while the original model only had 11.11% OOB error. This demonstrated that consensus process helped establish the classification accuracy to acceptable levels in the current study per the criteria from Plante and Vance (1994) and stabilized the estimates (e.g., Bayesian and Random Forest ROCs were similar). Thus, consensus may be necessary in future studies. The challenge is that a consensus process is less feasible in a clinical situation as clinicians typically evaluate children without others present. One solution might be to leave out modifiability scales. However, because there was higher classification accuracy for the modifiability and teaching scores combined than for teaching scores alone, modifiability is an important feature to include in future studies. I refer the reader to the discussion for Aim 1 for an in-depth discussion of solutions to potentially improve reliability for modifiability scoring.

Derivational Versus Inflectional Morphemes

Although it is likely that derivational and inflectional morphemes are acquired in a similar way, their use differs. Derivational morphemes that do not change word class can be avoided, and the grammaticality of the utterance does not change. For example, "I will reheat this" is just as grammatical as, "I will heat this again," and the meaning remains equivalent. In contrast, an inflectional morpheme is obligatorily produced when the dialect requires it because the absence of a required inflectional morpheme would result in an ungrammatical utterance. The current dynamic assessment used a derivational morpheme, which was not obligatory. Thus, it is most likely that this task tapped into underlying morphological skills but not necessarily underlying morphosyntactic skills. This nuanced distinction will be important for SLPs as they interpret the results. It is also important to note that good classification accuracy was reached in the current study. If those results can continue in future studies, it will be shown that morphology alone (rather than morphosyntax) is sufficient to distinguish between DLD and TD groups.

Operationalizing Bilingualism

While current percentage input and output of both languages was required to be greater than 20% for at least one language to be considered bilingual in the current study, there are other ways to operationalize bilingualism. One way might be to consider output (i.e., production) alone. It could be argued that output is a better variable to use in this type of study because the current study measures children's language skills by their expressive language production. Several children in the current study did not produce a language sample in both their languages, indicating weaker productive use of one language. The concern becomes generalizability—how bilingual are these children if they cannot converse in both languages? If current use were the method to measure bilingualism, the sample theoretically would be restricted to only those who were functionally bilingual. That restriction has the trade-off of ensuring that each child in the sample uses both languages at least somewhat, but this could limit generalizability in the future, potentially excluding a class of emergent bilinguals who have low use of one language that they are newly exposed to. Another way to define bilingualism for the sample would be to ask for functional proficiency, often operationalized as whether each child can hold a conversation in each language. This further restricts the sample to truly ensure bilingual use, and it may also reduce the number of unbalanced bilinguals who can participate. Nonetheless, it is a more rigorous way to ensure the participants are dual-language users. No matter where the cutoff is drawn, there is gatekeeping of who is considered bilingual. It is important to have criteria for research studies, so the results can be interpreted appropriately. However, it is important to remember that in life, the bilingual identity is not the researcher's to define.

Feasibility and Clinical Applicability of the Dynamic Assessment

As this dynamic assessment was designed with implementation in mind, it is important to consider the implementation factors along with classification accuracy to determine if it merits further study. Given that this research is at the feasibility stage, the dynamic assessment cannot be recommended for widespread clinical use yet (STARD #27). The lack of bias at this stage, the scoring reliability, and the classification accuracy are all promising. Some of the biggest considerations for clinicians are the time and burden of administration. The teaching phase was administered in about 18 minutes on average (range = 15 – 22 minutes). If a typical SLP has 45 – 60 minutes for an evaluation, this could fit within that time span and allow for several other assessments to be administered to get a full profile of the child's ability. The drawback is that this dynamic assessment does not provide information about the child's language use (e.g., types of errors, typical sentence structures, vocabulary diversity), so other assessments would be required to gather that information. The burden of assessment would be moderate at first because the clinician would have to print and organize the visual stimuli, script, and scoring pages (if accessed online), but beyond that, set-up would be minimal. Additionally, the current dynamic assessment has the benefit of accurately classifying Spanish-English bilingual children (who have at least 20% English use) without the need for an interpreter, which can minimize the time and money, needed for the interpreter during the evaluation.

The primary concern for implementation is the lack of reliability in modifiability scoring. Future studies should focus on achieving reliable modifiability scores both with trained research assistants and in actual clinical practice (see "Future Directions"). Other implementation science studies will be needed to understand potential barriers and facilitators to implementation.

Limitations

I incorporated limitations throughout the discussion. Those that did not fit into previous sections are discussed here (STARD #26). One limitation of the current study was its case-control design, which can introduce spectrum bias (Rutjes et al., 2005). To counteract that, I attempted to recruit across the spectrum of language ability by using the language sample scores. Additionally, the vocabulary check before the dynamic assessment served as a bit of a control because children with the lowest language could not label ten items in English. This is a true extreme and rare, but one potential participant had to be withdrawn because they could not label ten vocabulary items in English (or in Spanish, though that was not required). A few other children with DLD could not name ten out of twenty vocabulary items in English with the target word, so their synonyms were allowed and were used on the posttests (e.g., "oval" for "circle"). In clinical use, difficulties with this pretest vocabulary check may serve as an early indicator that DLD may be present. However, clinicians would need to be cautious about interpretation because difficulty labeling in English may be due to higher Spanish use rather than a language disorder. The simple solution would be to test the child in Spanish, as well.

One area of note in the dynamic assessment, though not inherently a limitation, was the utility of the guessability check. Although children who guessed too many items correctly on the guessability check were not included in this study, three such children continued with data collection to provide a sense of whether the assumption of the guessability check was true. The assumption was that children who pointed to the invented morpheme correctly at least five times, which had a very low probability based on chance alone, knew something about the invented morpheme. In other words, they would be predicted to respond correctly to most, if not all, items starting with item one. Of the three who guessed above chance, the teaching scores were 30 (TD), 1 (DLD), and 0 (TD). The child who scored 30 did not demonstrate immediate knowledge of /-əb/, not responding correctly until enhanced recast item seven. Additionally, the exclusion of these children reduced the sample size and variability, which may have affected the conclusions. For instance, the child with DLD in this subsample scored a 1, something that none of the other children with DLD in the final sample had done. That provides a potential counter-case to the idea that all children with DLD score 0 that should be examined further. This child had BESOS morphosyntax scores in the typical range and semantics scores below average, which led to a combined score in the DLD range. Because this is an unusual profile for DLD, it may have been that this child had low exposure to the vocabulary that had been tested and did not actually have DLD. This child’s modifiability score, though, would have classified them as having DLD had they scored a 0. Anecdotally, based on clinical experience, this child exhibited some signs of difficulties not characteristic of DLD that may have influenced the modifiability score (e.g., refusing to work and then immediately giving a hug and saying, “I love you,” to the assessor, difficulty sitting still, difficulty engaging in more than one item without breaks). Thus, the guessability check may not effectively capture inherent knowledge of the invented morpheme, so it may be better to exclude in future versions of this study.

As a feasibility study, the current study was designed to examine how the dynamic assessment functioned in a small portion of the population as a precursor to larger scale studies. Though the small sample size was a feature of the design, it did limit the confidence in and generalizability of the conclusions (Bujang & Adnan, 2016). Larger samples provide more information about bias within the assessment and across demographic groups, and they allow for more sophisticated statistical analyses. Because ROC curves are subject to variation in the sample and show low reliability in small sample sizes (Carter et al., 2016; Hanczar et al., 2010), their classification accuracy should not be considered applicable to external populations. Discriminant analysis requires a larger sample size and can provide more robust, generalizable information about classification accuracy.

Although I attempted to have a culturally responsive assessment battery, some elements may not have been less culturally responsive. The Frog stories used for the language samples included anachronistic elements (e.g., old-fashioned carriage, a magazine) and elements that are relatively uncommon in the Southwest US (e.g., a deer, a grassy park with a stream). Some of this lack of familiarity likely resulted in errors that children with other backgrounds may not have made. For instance, many children did not know the word for “deer” (e.g., “un moo”, “giraffe”, “horse”). However, these stories have been widely used with Spanish-English bilingual children and were used in the current study due to the availability of comparative data.

Another limitation was the inability to test in both languages. The initial set of children did not speak enough Spanish to assess them in both languages on the dynamic assessment as determined by their inability to label ten words in Spanish on the vocabulary check. Children may have performed the same across both languages, or it may be that some children would have learned the invented morpheme in Spanish or to a higher level of accuracy in Spanish compared to their English performance. Another possibility is that the modifiability scores would have differed across languages because a child might have shown more frustration or fatigue in English if it was their less proficient language. Although the theory is that the child's general language learning ability was engaged in this dynamic assessment, it cannot be known whether that was true without assessing in both languages.

Future Directions

The current study was a feasibility study to determine whether classification accuracy of the dynamic assessment from Roseberry and Connell (1991) could be improved by using statistical learning in the teaching phase, including posttests at multiple timepoints, and adding a modifiability scale. In future iterations, I would keep the morpheme /-əb/, the speech and vocabulary pre-checks, and the teaching phase with its adherence to statistical learning principles. However, I would eliminate the posttests and update the modifiability scale. For the guessability check, I would continue to administer it for another iteration but would continue administering the dynamic assessment to children who guessed above chance. If results showed an association between the guessability check and knowledge of the morpheme, I would include it in future iterations. Otherwise, I would exclude it from future studies. The next step in researching this dynamic assessment is to administer this assessment to a larger sample of children, so I can examine teaching parameters in detail. Additionally, a larger sample will support more robust statistics, including item response theory, using the omega coefficient for internal reliability and discriminant analysis for classification accuracy (Peña et al., 2013). The sample size may be built up over several smaller studies as long as core elements (i.e., a subset of the items, modifiability questions) are kept consistent to allow for comparison across studies. The benefit to this approach is that each study can make systematic improvements to the dynamic assessment while efficiently expending resources.

Study one would focus on improving the reliability of modifiability ratings. There are several possible approaches to this. One method would be to systematically assess the modifiability scale itself to decide how many response choices each question should have, which questions should be kept, and if any questions should be added. Borrowing from scoping and systematic review methods, I would look for all modifiability scales for dynamic assessments of language that are intended to identify children with DLD. Based on my experience choosing scales for the current assessment, many studies use the same or similar scales, so it is likely the results will not be exceedingly variant from the current scales. Then, I would remove duplicate items and pair similarly phrased items. Next, I and a group of trained raters familiar with the dynamic assessment would select all questions that might be relevant. An additional method would be to see how reliability would change by transforming the data from the current study to a three-choice scale. Following that, I could either go directly into testing those scales with participants, or I could workshop it by asking focus groups of SLPs what they think about the wording and operationalization of each item. In the focus groups, I could have the SLPs practice rating on the scale to get a sense of which items might have higher reliability. This method incorporates some principles of implementation science, which would be a first step to minimizing barriers to implementation later.

When testing the updated modifiability scale with participants in research, one approach would be to have two groups of assessors, each of which would have undergone different training. One group might have a more intensive training with a lot of discussion and practice items while the other group might have a more succinct training that would mirror how SLPs would most likely be trained to administer this dynamic assessment in clinical practice. Interrater reliability would be compared within and between groups. With IRT for scales, there should be at least 10 ratings at each response choice (Bond et al., 2021), so for a question with three response choices, there would need to be at least thirty participants. However, it is unlikely that the responses would be split by thirds exactly (i.e., 10 ratings for response choice 1, 10 ratings for response choice 2, 10 ratings for response choice 3), so likely at least double the number of participants would be needed. Even still, an *n* of 60 is relatively low for IRT, but it may support an initial look at the items' reliability. To make the most user-friendly modifiability scale, I would do a focus group with the assessors at the end of the project to hear about how the modifiability scale fit their experiences of working with the children. From this project, the modifiability rating scale with the highest interrater reliability and classification accuracy would be chosen to move forward with in future studies.

The next elements to be compared would be the dose number, variability, and languages of administration. Additionally, it might be important to expand the age range in future studies to determine if there is an age effect and to increase clinical utility. Study two would compare dose number across five conditions (see Table 23). Two of the conditions would have the original 48 doses, one matching the original study and the second increasing the focused stimulation doses to increase the dose rate. The second set of conditions would halve the total dose number, which would have the benefit of reduced time of administration if it were to demonstrate good classification accuracy. In condition three, there would be an equal split across the focused stimulation and enhanced recast doses for parity with condition one. Condition four, though, would have enhanced recasts only to see if the focused stimulation would even be needed. This enhanced recast only condition would have the same number of recasts as the original, but it would reduce the time of administration by excluding the focused stimulation. A fifth condition of only focused stimulation, a parallel with the original studies (Boyer & Martin, 2012; Hwa-Froelich & Matsuo, 2005; Kohnert & Danahy, 2007; Restrepo, 1998; Roseberry & Connell, 1991), would be included to determine whether variability was critical to the success of learning in this dynamic assessment. The dose number of 24 is similar to the dose number of 20 in many of the original studies of the invented morpheme dynamic assessment, so primarily, the difference would be the variability of the visual and linguistic stimuli. Because focused stimulation does not allow for interaction during the teaching phase, there would be no modifiability rating, and this condition would require an immediate posttest.

Table 23

Conditions for Comparison in Future Studies of the Dynamic Assessment

Condition	Focused Stimulation Doses	Enhanced Recast Doses	Total Dose Number	Descriptor
1	24	24	48	Original
2	36	12	48	High dose rate
3	12	12	24	Half the original
4	0	24	24	Recasts only
5	24	0	24	Focused stimulation only

Based on the results of study two, a study to examine the classification accuracy by variability condition may or may not be needed. It would only be needed if classification accuracy were not at the ideal range yet (i.e., sensitivity and specificity at .90 and above). A variability-focused study could include multiple conditions, similar to study two, to compare items across the number of angles that were cut (e.g., vary cuts across all four corners, just two, or just one) and the number of cut types (e.g., all four: straight, big curve, cloud-like cut, and zigzag; just two; or just one). Linguistic variability could be assessed systematically if needed, but condition five from the previous study would address whether it would need to be tested systematically or if high variability of the platform utterances should be maintained.

Finally, it is important to study the dynamic assessment in both languages. Spanish and English versions could be incorporated into each of the previously described studies, or the language comparison could be its own study. I would create a Spanish equivalent to the winning English version and explore how children's scores are impacted by working in both languages. The Spanish version would use a morpheme appropriate for Spanish morpho-phonology: -ebo/a (e.g., "gatebo" or "caseba" for "partial cat" or "partial house," respectively). Assessments would be counter-balanced by language. When studying the classification accuracy of both languages compared to only one language, the spotlight would be on whether a different language or both languages were best for each bilingual profile. For instance, a child who has only been exposed to English for a month might need the Spanish dynamic assessment whereas a child who hears both languages equally might do best with English alone (as the current study suggests) or with both the English and Spanish versions of the dynamic assessment.

Once the most accurately-classifying, time-efficient, and linguistically appropriate version was found, the next step would be to validate this dynamic assessment with a large sample across multiple regions of the country and across multiple evaluators. That study could also have an eye on implementation, examining how clinicians might utilize this test in real-world situations. Adaptations could be made based on the facilitators and challenges they encountered. If a Spanish version were utilized, implementation studies would also need to explore how to best administer this dynamic assessment with interpreters. With sufficient data pointing to the assessment's accuracy, the final step would be disseminating the dynamic assessment.

Conclusions

The current study put a modern twist on an old dynamic assessment by Roseberry and Connell (1991). The goal was to understand which elements of the assessment yielded high classification accuracy for Spanish-English preschoolers with DLD for this version of the dynamic assessment. Learning potential was evaluated by seeing how well children learned the invented morpheme, /-əb/, which was added to nouns to mean a partial object. Children were taught using statistical learning principles of high regularity and variability through the strategies of focused stimulation and enhanced recasting. Their performance was measured by their accuracy at producing the invented morpheme during the enhanced recasting and on immediate and delayed posttests. Their modifiability was recorded on rating scales and measured by the time needed to complete the teaching phase. Through ROC curves, the combination of the teaching phase score and modifiability rating scales were determined to yield acceptable classification accuracy (sensitivity = 83%, specificity = 100%). However, while interrater reliability and fidelity to administration was high for the dynamic assessment, the modifiability rating scales had low interrater reliability. Additionally, ROC curves predicate their analysis on the variability of the sample. Thus, future studies will be needed to improve the reliability of the modifiability scales and allow for more robust statistical analyses to yield confidence in the ability of this dynamic assessment to distinguish between TD and DLD bilingual children.

Appendix A. The Unused Spanish Version of The Dynamic Assessment

Although the dynamic assessment was originally planned to be administered once in English and once in a parallel Spanish version, the Spanish version had to be omitted in the current study. The first few participants did not produce enough Spanish on the Spanish vocabulary check to participate in the Spanish dynamic assessment, so it was decided to maintain consistency across participants and only administer the English version. The Spanish version mirrored the English version's format and included the exact same modifiability scoring. The only difference was that the Spanish version was entirely in Spanish, which obligated the selection of a different invented morpheme given the difference between English and Spanish phonotactics.

Originally, the Spanish morpheme /-be(i)/ was selected given its use in Kohnert and Danahy (2007) and Boyer and Martin (2012), its dissimilarity to any free or bound morphemes in Spanish, and its parity with the English morpheme in that it was a single syllable and had the phoneme /b/. However, it was unclear which syllable would need to be stressed with /-be(i)/ affixed. There was an argument to retain the word's original stress to keep the word as a whole, understandable form (e.g., /ˈkasaβei/; G. Arizmendi, personal communication, September 19, 2023; L. Hiebert, personal communication, September 15, 2023). Another argument could be made that the stress of the word should shift to ensure the penultimate syllable would be stressed, which would be in accordance with the general Spanish rules (e.g., /kaˈsaβei/; Bakovic, 2016). A third argument suggested that the morpheme should be stressed if it ended in the diphthong (e.g., /kasaˈβei/; S. Colina, personal communication, September 18, 2023). It is also relatively uncommon for Spanish words to end in the vowel /-e/ or the diphthong /-ei/. The lack of clarity along with the theoretical linguistics suggested that /-bei/ would be an odd construction for

Spanish and perhaps a harder task than what children would encounter in ordinary life (S. Colina, personal communication, September 18, 2023), so /-bei/ was exchanged for /-ebo/ or /-eba/.

The /-eβo/ or /-eβa/ morpheme was patterned after the Spanish diminutive construction of /-ito/ or /-ita/. The final letter /-o/ or /-a/ matched the root word's gender of masculine or feminine, respectively. This morpheme met the criteria of being relatively easy to produce articulatorily and not resembling any currently existing free or bound morphemes in Spanish. It also fit better with Spanish morphophonological rules because the stress predictably fell on the first syllable of the morpheme (/-ˈeβa/). Even though it was a multisyllabic morpheme, it only added one syllable to each word. The terminal element (the /-o/ or /-a/) was removed from the end of root word, and the final element of the invented morpheme acquired the /-o/ or /-a/ terminal element to match the root word's gender (e.g., "casa" as /kas-ˈeβ-a/).

With this change, I maintained high regularity, a statistical learning principle, by ensuring that the root words had canonical constructions (i.e., gender matched an ending of /-o/ or /-a/). Words with non-canonical constructions were removed from the root word list and replaced with canonical words (e.g., "l**a** luz" replaced with "l**a** bols***a***"). Thus, counter-examples were avoided. In addition, I distributed both genders evenly across each phase of the dynamic assessment.

For the Spanish word list (see Table A1), I used the EsPal database for Latin American Spanish using the search criteria for written and web tokens (Duchon et al., 2013). Words were first found on EsPal, then were cross-checked for age of acquisition to reduce the list. When the morpheme was switched, I added more root words, so I had to expand the word frequency range to 3.00 and above. However, I ensured the words were still relatively commonly known by seeking stimuli with similar item trajectories to the Stanford Wordbank (Frank et al., n.d.).

Table A1

Psycholinguistic Properties of Dynamic Assessment Stimuli

Psycholinguistic Characteristic	Value Range	Original Database
Word Frequency (Zipf scale)	4.00 – 7.62	EsPal: (Duchon et al., 2013)
Number of Syllables	1 – 3	EsPal: (Duchon et al., 2013)
Number of Phonological Neighbors	3 - 75	EsPal: (Duchon et al., 2013)
Category of Speech	Noun	EsPal: (Duchon et al., 2013)
Age of Acquisition	1.5 – 3.8	SANDchild database: (Sabater et al., 2020)

Appendix B. Spanish Translations of Instructions for Tasks

Hearing Screening

"Vas a escuchar algunos sonidos. Cada vez que escuches un sonido, levanta la mano. A veces, los sonidos serán muy suaves/bajitos. Levanta la mano tan pronto como escuches el sonido. ¿Lo entiendes?"

Language Sample

Before Story

Aquí tengo un libro. Vamos a escuchar a este cuento mientras miramos el libro juntos. Cuando terminemos, quiero que me vuelvas a contar el cuento en español. ¿Ok? Vamos a mirar el primer libro.

After Story

Ahora, cuéntame lo que pasó en este cuento.

Appendix C. Language Sample Codes

Feature Marked	Code	English example	Spanish example
SALT Conventions			
Plurals	/s	Frog/s	Rana/s
Possessives	/z	Boy/z head	N/A
Third Person Singular	/3s	He call/3s	N/A
Past Tense	/ed	He call/ed	N/A
Past Participle	/en	He had eat/en	N/A
Progressive	/ing	He was call/ing	N/A
Contractions	/n't, /'ll, /'ve, etc.	He could/n/t	N/A
Bound Pronomial Clitics	+clitic	N/A	Va a buscar+lo.
Mazes	(maze)	(He he um) he went.	
Root Forms	verb|infinitive or diminutive|noun or shortened word|word	Doggie|dog	Voy|ir Osito|oso Pa|para
Codeswitch	[CS]	The boy had a rana[CS].	El niño tenía una frog[CS].
Reflexive Pronoun	[x]	N/A	Se[x] cayó|caer.
Modified C-unit	[F]	C The frog jump/ed [F]. C And land/ed in the water [F].	N/A
Errors			
Omissions			
Preposition	*preposition[EW:omission_prep]	A squirrel came out *of[EW:omission_prep] the hole.	Se subió *en[EW:omission_prep] una piedra.
Verb	*verb[EW:omission_verb]	I don't know what that *is[EW:omission_verb] called.	Y el niño estaba *diciendo[EW:omission_verb], "¿Rana, dónde estás?"
Article	*article[EW:omission_article]	He looked in *the[EW:omission_article] hole.	Uno brincó en *el[EW:omission_article] niño.
Clitic	*clitic[EW:omission_clitic]	N/A	Y *lo[EW:omission_clitic] vio.

Word errors			
Tense	[EW:tense]	The dog ran as fast as he can[EW:tense].	Por mientras el perro estaba molesto[EW:tense] a las abejas.
Number	[EW:number]	There was[EW:number] baby frogs.	Buscó en su[EW:number] botas.
Number and Tense	[EW:number_tense][a]	The boy come[EW:number_tense] downstairs.	Rare[b]
Gender	[EW:gender]	N/A	y vio un[EW:gender] rata.
Reflexive	[EW:reflexive]	N/A	Y se[EW:reflexive] estaban enojados con el perro.
Pronoun	[EP:correct pronoun]	Him[EP:he] went.	Rare[b]
Word Choice			
Preposition	[EW:wordchoice_prep]	The boy looked on[EW:wordchoice_prep] his boots.	Se cayó a[EW:wordchoice_prep] su bote.
Other	[EW:wordchoice]	The bug[EW:wordchoice] fall down.	Cuando ellos se despertaron, estaba el jarrón vacído[EW:wordchoice].
Utterance-Level Error			
General	[EU]	The gonna be something else [EU].	Y que no estaban más una un camino [EU].
Word order	[WO]	And then fell down, the dog and the boy [WO].	y luego el perro seguía de lado a él corriendo [WO].

[a]The code [EW:number_tense] was used when a verb had a number error in the present tense, but past tense would have been more appropriate.

[b]Rare indicates codes that were possible but never or almost never applied.

References

Ahufinger, N., Guerra, E., Ferinu, L., Andreu, L., & Sanz-Torrent, M. (2021). Cross-situational statistical learning in children with Developmental Language Disorder. *Language, Cognition and Neuroscience*, *36*(9), 1180-1200. https://doi.org/10.1080/23273798.2021.1922723

Alt, M. (2011). Phonological working memory impairments in children with specific language impairment: Where does the problem lie? *Journal of Communication Disorders*, *44*(2), 173-185. https://doi.org/10.1016/j.jcomdis.2010.09.003

Alt, M., Figueroa, C. R., Mettler, H. M., Evans-Reitz, N., & Erikson, J. A. (2021). A Vocabulary Acquisition and Usage for Late Talkers Treatment efficacy study: The effect of input utterance length and identification of responder profiles. *Journal of Speech, Language, and Hearing Research*, *64*(4), 1235-1255. https://doi.org/10.1044/2020_JSLHR-20-00525

Alt, M., Mettler, H. M., Erikson, J. A., Figueroa, C. R., Etters-Thomas, S. E., Arizmendi, G. D., & Oglivie, T. (2020). Exploring input parameters in an expressive vocabulary treatment with late talkers. *Journal of Speech, Language, and Hearing Research*, *63*(1), 216-233. https://doi.org/10.1044/2019_JSLHR-19-00219

American Academy of Audiology. (2011). *Childhood Hearing Screening Guidelines*.

American Speech-Language-Hearing Association. (2022). *Profile of ASHA multilingual service providers, year-end 2022*. https://www.asha.org/practice-portal/professional-issues/multilingual-service-delivery/

American Speech-Language-Hearing Association. (n.d.-a). *Childhood hearing screening (Practice portal)*. Retrieved August 15, 2023 from www.asha.org/Practice-Portal/Professional-Issues/Childhood-Hearing-Screening/

American Speech-Language-Hearing Association. (n.d.-b). *Dynamic Assessment*. Retrieved June 13, 2022 from https://www.asha.org/practice/multicultural/dynamic-assessment/

Anderson, R. T. (2022). Spanish child heritage speakers: Patterns of change and implications for clinical practice. In B. A. Goldstein (Ed.), *Bilingual language development and disorders in Spanish-English speakers* (3rd ed., pp. 193-214). Brookes Publishing Co.

Antovich, D. M., & Graf Estes, K. (2018). Learning across languages: Bilingual experience supports dual language statistical word segmentation. *Developmental Science*, *21*(2). https://doi.org/10.1111/desc.12548

Apfelbaum, K. S., & McMurray, B. (2011). Using variability to guide dimensional weighting: Associative mechanisms in early word learning. *Cognitive Science*, *35*(6), 1105-1138. https://doi.org/10.1111/j.1551-6709.2011.01181.x

Arizona Revised Statutes, Title 15, Article 3.1, §15-751–17.755 (2000).

Aslin, R. N., & Newport, E. L. (2012). Statistical learning: From acquiring specific items to forming general rules. *Current Directions in Psychological Science*, *21*(3), 170-176. https://doi.org/10.1177/0963721412436806

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321-324. https://doi.org/10.1111/1467-9280.00063

Axelsson, E. L., Williams, S. E., & Horst, J. S. (2016). The effect of sleep on children's word retention and generalization. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01192

Babad, E. Y., & Budoff, M. (1974). Sensitivity and validity of learning-potential measurement in three levels of ability. *Journal of Educational Psychology*, *66*(3), 439-447. https://doi.org/10.1037/h0036510

Bakovic, E. (2016). Exceptionality in Spanish stress. *Catalan Journal of Linguistics*, *15*, 9-25. https://doi.org/10.5565/rev/catjl.182

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchinson, K. A., Kessler, B., Loftis, B., Neely, J. H., Neslon, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445-459. https://doi.org/10.3758/BF03193014

Bamford, C. K., Masso, S., Baker, E., & Ballard, K. J. (2022). Dynamic assessment for children with communication disorders: A systematic scoping review and framework. *American Journal of Speech-Language Pathology*, *31*(4), 1878-1893. https://doi.org/10.1044/2022_AJSLP-21-00349

Baron, A., Bedore, L. M., Pena, E. D., Lovgren-Uribe, S. D., Lopez, A. A., & Villagran, E. (2018). Production of Spanish grammatical forms in U.S. bilingual children. *American Journal of Speech-Language Pathology*, *27*(3), 975-987. https://doi.org/10.1044/2018_AJSLP-17-0074

Barragán, B., Castilla-Earls, A., Martinez-Nieto, L., Restrepo, M. A., & Gray, S. (2018). Performance of low-income dual language learners attending English-only schools on the Clinical Evaluation of Language Fundamentals-Fourth Edition, Spanish. *Language*, *Speech, and Hearing Services in Schools*, *49*(2), 292-305. https://doi.org/10.1044/2017_LSHSS-17-0013

Bartolotti, J., Marian, V., Schroeder, S. R., & Shook, A. (2011). Bilingualism and inhibitory control influence statistical learning of novel word forms. *Frontiers in Psychology*, *2,* 324. https://doi.org/10.3389/fpsyg.2011.00324

Bedore, L. M., Pena, E. D., Gillam, R. B., & Ho, T. H. (2010). Language sample measures and language ability in Spanish-English bilingual kindergarteners. *Journal of Communication Disorders*, *43*(6), 498-510. https://doi.org/10.1016/j.jcomdis.2010.05.002

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

Benitez, V. L., Yurovsky, D., & Smith, L. B. (2016). Competition between multiple words for a referent in cross-situational word learning. *Journal of Memory and Language*, *90*, 31-48. https://doi.org/10.1016/j.jml.2016.03.004

Berman, H. B. (n.d.). *Binomial probability calculator*. https://stattrek.com/online-calculator/binomial

Bishop, D. V., Snowling, M. J., Thompson, P. A., Greenhalgh, T., & consortium, C. (2016). CATALISE: A multinational and multidisciplinary delphi consensus study. Identifying language impairments in children. *PLoS One*, *11*(7), e0158753. https://doi.org/10.1371/journal.pone.0158753

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

Bitetti, D., & Hammer, C. S. (2021). English narrative macrostructure development of Spanish–English bilingual children from preschool to first grade. *American Journal of Speech-Language Pathology*, *30*(3), 1100-1115. https://doi.org/10.1044/2021_ajslp-20-00046

Bond, T., Yan, Z., & Heene, M. (2021). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Taylor & Francis.

Born, J., Rasch, B., & Gais, S. (2006). Sleep to remember. *Neuroscientist*, *12*(5), 410-424. https://doi.org/10.1177/1073858406292647

Boyer, V. E., & Martin, K. Y. (2012). Invented rule with English language learners. *Clinical Linguistics & Phonetics*, *26*(7), 613-627. https://doi.org/10.3109/02699206.2012.682694

Bracken, B. A. (1988). Ten psychometric reasons why similar tests produce dissimilar results. *Journal of School Psychology*, *26*(2), 155-166. https://doi.org/10.1016/0022-4405(88)90017-9

Braine, M. D. S. (1987). What is learned in acquiring word classes—A step toward an acquisition theory. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (1st ed., pp. 65-87). Psychology Press. https://doi.org/10.4324/9781315798721

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Broedelet, I., Boersma, P., & Rispens, J. (2023). Implicit cross-situational word learning in children with and without Developmental Language Disorder and its relation to lexical-semantic knowledge. *Frontiers in Communication*, *8*. https://doi.org/10.3389/fcomm.2023.1021654

Broer, M., Bai, Y., & Fonseca, F. (2019). *A review of the literature on socioeconomic status and educational achievement: Evidence from twenty years of TIMSS* (Vol. 5). Springer Open. https://doi.org/10.1007/978-3-030-11991-1

Brown, A. L., & French, L. A. (1979). The zone of potential development: Implications for intelligence testing in the year 2000. *Intelligence*, *3*(3), 255-273. https://doi.org/10.1016/0160-2896(79)90021-7

Brown, R. (1976). *First language: The early stages* (1st ed.). Harvard University Press.

Bryant, N. R. (1982). *Preschool children's learning and transfer of matrices problems: A study of proximal development* [Master's Thesis, University of Illinois at Urbana-Champaign].

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977-990. https://doi.org/10.3758/BRM.41.4.977

Budoff, M., & Pagell, W. (1968). Learning potential and rigidity in the adolescent mentally retarded. *Journal of Abnormal Psychology*, *73*(5), 479-486. https://doi.org/10.1037/h0026219

Bujang, M. A., & Adnan, T. H. (2016). Requirements for minimum sample size for sensitivity and specificity analysis. *Journal of Clinical Diagnostic Research*, *10*(10), YE01-YE06. https://doi.org/10.7860/JCDR/2016/18129.8744

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1-28. https://doi.org/10.18637/jss.v080.i01

Camilleri, B., Hasson, N., & Dodd, B. (2014). Dynamic assessment of bilingual children's language at the point of referral. *Educational & Child Psychology*, *31*(2), 57-72. https://openaccess.city.ac.uk/id/eprint/4224/

Camilleri, B., & Law, J. (2007). Assessing children referred to speech and language therapy: Static and dynamic assessment of receptive vocabulary. *Advances in Speech Language Pathology*, *9*(4), 312-322. https://doi.org/10.1080/14417040701624474

Campione, J. C., Brown, A. L., Ferrara, R. A., & Bryant, N. R. (1984). The zone of proximal development: Implications for individual differences and learning. *New Directions for Child and Adolescent Development*, *23*, 77-91. https://doi.org/10.1002/cd.23219842308

Campione, J. C., Brown, A. L., Ferrara, R. A., Jones, R. S., & Steinberg, E. (1985). Breakdowns in flexible use of information: Intelligence-related differences in transfer following equivalent learning performance. *Intelligence*, *9*(4), 297-315. https://doi.org/10.1016/0160-2896(85)90017-0

Carter, J. V., Pan, J., Rai, S. N., & Galandiuk, S. (2016). ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. *Surgery*, *159*(6), 1638-1645. https://doi.org/10.1016/j.surg.2015.12.029

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. https://doi.org/10.1044/2020_AJSLP-19-00179

Castilla-Earls, A., Perez-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., Pérez-Leroux, A. T., Martinez-Nieto, L., Restrepo, M. A., & Barr, C. (2019). Vulnerability of clitics and articles to bilingual effects in typically developing Spanish-English bilingual children. *Bilingualism: Language and Cognition*, *23*, 825-835. https://doi.org/10.1017/S1366728919000610

Castilla-Earls, A., Ronderos, J., & Fitton, L. (2023). Spanish bilingual morphosyntactic development in bilingual children with and without Developmental Language Disorder: Articles, clitics, verbs, and the subjunctive mood. *Journal of Speech, Language, and Hearing Research*, *66*(12), 4678-4698. https://doi.org/10.1044/2023_JSLHR-23-00091

Chiat, S. (2001). Mapping theories of developmental language impairment: Premises, predictions and evidence. *Language and Cognitive Processes*, *16*(2-3), 113-142. https://doi.org/10.1080/01690960042000012

Cohen, J. (1977a). *Statistical power analysis for the behavioral sciences* (Revised ed.). Elsevier Science & Technology.

Cohen, J. (1977b). The t test for means. In *Statistical power analysis for the behavioral sciences: Revised edition* (pp. 19-74). Academic Press.

Cohen, J. F., Korevaar, D. A., Altman, D. G., Bruns, D. E., Gatsonis, C. A., Hooft, L., Irwig, L., Levine, D., Reitsma, J. B., de Vet, H. C., & Bossuyt, P. M. (2016). STARD 2015 guidelines for reporting diagnostic accuracy studies: Explanation and elaboration. *BMJ Open*, *6*(11), e012799. https://doi.org/10.1136/bmjopen-2016-012799

Cohen, J. F., Korevaar, D. A., Gatsonis, C. A., Glasziou, P. P., Hooft, L., Moher, D., Reitsma, J. B., de Vet, H. C. W., & Bossuyt, P. M. (2017). STARD for Abstracts: Essential items for reporting diagnostic accuracy studies in journal or conference abstracts. *BMJ*, *358*, j3751-j3751. https://doi.org/10.1136/bmj.j3751

Connell, P. J. (1987). An effect of modeling and imitation teaching procedures on children with and without Specific Language Impairment. *Journal of Speech and Hearing Research*, *30*(1), 105-113. https://doi.org/10.1044/jshr.3001.105

Cooperson, S. J., Bedore, L. M., & Pena, E. D. (2013). The relationship of phonological skills to language skills in Spanish-English-speaking bilingual children. *Clin Linguist Phon*, *27*(5), 371-389. https://doi.org/10.3109/02699206.2013.782568

Crespo, K., & Kaushanskaya, M. (2021). Is 10 better than 1? The effect of speaker variability on children's cross-situational word learning. *Language Learning and Development*, *17*(4), 397-410. https://doi.org/10.1080/15475441.2021.1906680

Crespo, K., & Kaushanskaya, M. (2022). The role of attention, language ability, and language experience in children's artificial grammar learning. *Journal of Speech, Language, and Hearing Research*, *65*(4), 1574-1591. https://doi.org/10.1044/2021_JSLHR-21-00112

Crespo, K., Vlach, H., & Kaushanskaya, M. (2023). The effects of bilingualism on children's cross-situational word learning under different variability conditions. *Journal of Experimental Child Psychology*, *229*, 105621. https://doi.org/10.1016/j.jecp.2022.105621

Crowe, K., & McLeod, S. (2020). Children's English consonant acquisition in the United States: A review. *American Journal of Speech Language Pathology*, *29*(4), 2155-2169. https://doi.org/10.1044/2020_AJSLP-19-00168

Dautriche, I., & Chemla, E. (2014). Cross-situational word learning in the right situations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(3), 892-903. https://doi.org/10.1037/a0035657

Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: Neural and behavioural evidence. *Philosophical Transactions of the Royal Society B: Biological Sciences,*, *364*(1536), 3773-3800. https://doi.org/10.1098/rstb.2009.0111

De Cat, C. (2021). Socioeconomic status as a proxy for input quality in bilingual children? *Applied Psycholinguistics*, *42*(2), 301-324. https://doi.org/10.1017/s014271642000079x

Denman, D., Cordier, R., Kim, J. H., Munro, N., & Speyer, R. (2021). What influences Speech-Language Pathologists' use of different types of language assessments for elementary school-age children? *Language, Speech, and Hearing Services in Schools*, *52*(3), 776-793. https://doi.org/10.1044/2021_LSHSS-20-00053

Desmottes, L., Maillart, C., & Meulemans, T. (2017). Memory consolidation in children with Specific Language Impairment: Delayed gains and susceptibility to interference in implicit sequence learning. *Journal of Clinical and Experimental Neuropsychology*, *39*(3), 265-285. https://doi.org/10.1080/13803395.2016.1223279

Desmottes, L., Meulemans, T., & Maillart, C. (2016). Later learning stages in procedural memory are impaired in children with Specific Language Impairment. *Research in Developmental Disabilities*, *48*, 53-68. https://doi.org/10.1016/j.ridd.2015.10.010

Dollaghan, C. A. (2007). *The handbook for evidence-based practice in communication disorders*. Paul H. Brookes Publishing Co.

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)

Dowling, J., & Hugo Lopez, M. (2022). Understanding and measuring Latino racial and ethnic identity [Webinar]. National Research Center on Hispanic Children and Families.

https://www.hispanicresearchcenter.org/research-resources/understanding-and-measuring-latino-racial-and-ethnic-identity/

Duchon, A., Perea, M., Sebastian-Galles, N., Marti, A., & Carreiras, M. (2013). EsPal: One-stop shopping for Spanish word properties. *Behavior Research Methods*, *45*(4), 1246-1258. https://doi.org/10.3758/s13428-013-0326-1

Dudai, Y. (2004). The neurobiology of consolidations, or, how stable is the engram? *Annual Review of Psychology*, *55*, 51-86. https://doi.org/10.1146/annurev.psych.55.090902.142050

Dunn Davison, M., & Scheffner Hammer, C. (2012). Development of 14 English grammatical morphemes in Spanish-English preschoolers. *Clinical Linguistics and Phonetics*, *26*(8), 728-742. https://doi.org/10.3109/02699206.2012.700679

Ebbers, S. (2010). Master morpheme list. In *Vocabulary through morphemes: Suffixes, prefixes, and roots for intermediate grades* (2nd ed.). SoprisWest.

Ehrler, D. J., & McGhee, R. L. (2008). *Primary Test of Nonverbal Intelligence: Examiner's Manual*. Pro-Ed.

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. https://doi.org/10.1044/0161-1461(2012/11-0089)

Eisenberg, S. L., Guo, L. Y., & Germezia, M. (2012). How grammatical are 3-year-olds? *Language, Speech, and Hearing Services in Schools*, *43*(1), 36-52. https://doi.org/10.1044/0161-1461(2011/10-0093)

Embretson, S. E. (1987). Toward development of a psychometric approach. In C. S. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 141-170). The Guilford Press.

Ensminger, M. E., & Fothergill, K. E. (2002). A decade of measuring SES: What it tells us and where to go from here. In R. H. B. Marc H. Bornstein, *Socioeconomic status, parenting, and child development* (pp. 13-27). Taylor & Francis Group.

Erickson, L. C., & Thiessen, E. D. (2015). Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition. *Developmental Review*, *37*, 66-108. https://doi.org/10.1016/j.dr.2015.05.002

Escobar, K., Smith, J., Scheffner Hammer, C., & Rodríguez, B. (2022). Bilingual language acquisition and the child socialization process. In B. Goldstein (3rd Ed.), *Bilingual language development and disorders in Spanish-English speakers* (pp. 66-85). Brookes Publishing.

Escudero, P., Mulak, K. E., Fu, C. S., & Singh, L. (2016). More limitations to monolingualism: Bilinguals outperform monolinguals in implicit word learning. *Frontiers in Psychology*, *7*, 1218. https://doi.org/10.3389/fpsyg.2016.01218

Evans, J. L., Saffran, J. R., & Robe-Torres, K. (2009). Statistical learning in children with Specific Language Impairment. *Journal of Speech, Language, & Hearing Research*, *52*(2), 321-335. https://doi.org/10.1044/1092-4388(2009/07-0189)

Ferrara, R. A., Brown, A. L., & Campione, J. C. (1986). Children's learning and transfer of inductive reasoning rules: Studies of proximal development. *Child Development*, *57*(5), 1087-1099. https://doi.org/10.1111/j.1467-8624.1986.tb00438.x

Feuerstein, R. (1979). *The dynamic assessment of retarded performers*. University Park Press.

Fey, M. E., & Finestack, L. H. (2009). Research and development in child language intervention: A five-phase model. In R. G. Schwartz (1st Ed.), *Handbook of child language disorders* (pp. 513-531). Routledge Handbooks Online. https://doi.org/10.4324/9780203837764-34

Fillmore, L. W. (1991). When learning a second language means losing the first. *Early Childhood Research Quarterly*, *6*(3), 323-346. https://doi.org/10.1016/S0885-2006(05)80059-6

Frahm, A. E. (2021). *Norming a dynamic assessment of narrative language for diverse school-age children with and without language disorder: A preliminary psychometric study* (Publication Number 8935) [Master Thesis, Brigham Young University]. BYU ScholarsArchive.

Frank, M. C., Braginsky, M., Yurovsky, D., Marchman, V., & Keller, D. (n.d.). *Wordbank*. http://wordbank.stanford.edu/

Fulcher-Rood, K., & Castilla-Earls, A. (2023). Differences in child language assessment practices between school-based and non-school-based Speech-Language Pathologists: Results from a nationwide survey. *Language, Speech, and Hearing Services in Schools*, *54*(4), 1-19. https://doi.org/10.1044/2023_LSHSS-22-00185

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*

SpeechpLanguage Pathology, *27*(2), 796-812. https://doi.org/10.1044/2018_AJSLP-16-0121

Gaeta, L., & Brydges, C. R. (2020). An examination of effect sizes and statistical power in speech, language, and hearing research. *Journal of Speech, Language, and Hearing Research*, *63*(5), 1572-1580. https://doi.org/10.1044/2020_JSLHR-19-00299

Gao, C., Shinkareva, S. V., & Desai, R. H. (2022). SCOPE: The South Carolina psycholinguistic metabase. *Behavior Research Methods*, 1-32. https://doi.org/10.3758/s13428-022-01934-0

Gatt, D., Baldacchino, R., & Dodd, B. (2020). Which measure of socioeconomic status best predicts bilingual lexical abilities and how? A focus on four-year-olds exposed to two majority languages. *Journal of Child Language*, *47*(4), 737-765. https://doi.org/10.1017/S0305000919000886

Gillis, J. U., Gul, A., Fox, A., Parikh, A., & Arbel, Y. (2022). Artificial grammar learning in children with Developmental Language Disorder. *Journal of Speech, Language, and Hearing Research*, *65*(1), 292-302. https://doi.org/10.1044/2021_JSLHR-21-00050

Goffman, L., & Gerken, L. (2023). A developmental account of the role of sequential dependencies in typical and atypical language learners. *Cognitive Neuropsychology*, *40*(5-6) 1-22. https://doi.org/10.1080/02643294.2023.2275837

Gómez, R. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 431-436. https://doi.org/10.1111/1467-9280.00476

Gómez, R. L., & Lakusta, L. (2004). A first step in form-based category abstraction by 12-month-old infants. *Developmental Science*, *7*(5), 567-580. https://doi.org/10.1111/j.1467-7687.2004.00381.x

Gordon, K. R. (2020). The advantages of retrieval-based and spaced practice: Implications for word learning in clinical and educational contexts. *Language, Speech, and Hearing Services in Schools, 51*(4), 955-965. https://doi.org/10.1044/2020_LSHSS-19i-00001

Gordon, K. R., Storkel, H. L., Lowry, S. L., & Ohlmann, N. B. (2021). Word learning by preschool-age children with Developmental Language Disorder: Impaired encoding and robust consolidation during slow mapping. *Journal of Speech, Language, and Hearing Research, 64*(11), 4250-4270. https://doi.org/10.1044/2021_JSLHR-21-00046

Graham, L. J., & Tancredi, H. (2019). In search of a middle ground: The dangers and affordances of diagnosis in relation to Attention Deficit Hyperactivity Disorder and Developmental Language Disorder. *Emotional and Behavioural Difficulties*, *24*(3), 287-300. https://doi.org/10.1080/13632752.2019.1609248

Gray, S. (2004). Word learning by preschoolers with Specific Language Impairment predictors and poor learners. *Journal of Speech, Language, and Hearing Research*, *48*(6), 1452-1467. https://doi.org/10.1044/1092-4388(2005/101)

Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and Language*, *36*(1), 3-15. https://doi.org/10.1016/0093-934X(89)90048-5

Grosjean, F. (2014). Bicultural bilinguals. *International Journal of Bilingualism*, *19*(5), 572-586. https://doi.org/10.1177/1367006914526297

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

Haahr, M. (n.d.). *RANDOM.ORG: True random number service*. https://www.random.org/

Hall, J., Owen Van Horne, A. J., McGregor, K. K., & Farmer, T. A. (2018). Individual and developmental differences in distributional learning. *Language, Speech, and Hearing Services in Schools*, *49*(3S), 694-709. https://doi.org/10.1044/2018_LSHSS-STLT1-17-0134

Hammer, C. S., Hoff, E., Uchikoshi, Y., Gillanders, C., Castro, D., & Sandilos, L. E. (2014). The language and literacy development of young Dual Language Learners: A critical review. *Early Childhood Research Quarterly*, *29*(4), 715-733. https://doi.org/10.1016/j.ecresq.2014.05.008

Hanczar, B., Hua, J., Sima, C., Weinstein, J., Bittner, M., & Dougherty, E. R. (2010). Small-sample precision of ROC-related estimates. *Bioinformatics*, *26*(6), 822-830. https://doi.org/10.1093/bioinformatics/btq037

Hasson, N., Camilleri, B., Jones, C., Smith, J., & Dodd, B. (2012). Discriminating disorder from difference using dynamic assessment with bilingual children. *Child Language Teaching and Therapy*, *29*(1), 57-75. https://doi.org/10.1177/0265659012459526

Hasson, N., Dodd, B., & Botting, N. (2012). Dynamic Assessment of Sentence Structure (DASS): Design and evaluation of a novel procedure for the assessment of syntax in children with language impairments. *Internationa Journal of Language and Communication Disorders*, *47*(3), 285-299. https://doi.org/10.1111/j.1460-6984.2011.00108.x

Hedenius, M., Persson, J., Tremblay, A., Adi-Japha, E., Verissimo, J., Dye, C. D., Alm, P., Jennische, M., Bruce Tomblin, J., & Ullman, M. T. (2011). Grammar predicts procedural learning and consolidation deficits in children with Specific Language Impairment.

Research in Developmental Disabilities, *32*(6), 2362-2375. https://doi.org/10.1016/j.ridd.2011.07.026

Henderson, D. E., Restrepo, M. A., & Aiken, L. S. (2018). Dynamic assessment of narratives among Navajo preschoolers. *Journal of Speech, Language, and Hearing Research*, *61*(10), 2547-2560. https://doi.org/10.1044/2018_JSLHR-L-17-0313

Henderson, L. M., Weighall, A. R., Brown, H., & Gareth Gaskell, M. (2012). Consolidation of vocabulary is associated with sleep in children. *Developmental Science*, *15*(5), 674-687. https://doi.org/10.1111/j.1467-7687.2012.01172.x

Hernandez, M., & Castilla-Earls, A. P. (2022). *The diagnostic accuracy of bilingual PGU in bilingual Spanish-speaking children with and without Developmental Language Disorders* [Poster presentation]. The Symposium on Research in Child Language Disorders, Madison, WI, United States.

Hernandez, M., Ronderos, J., & Castilla-Earls, A. P. (2024). Diagnostic accuracy of grammaticality and utterance length in bilingual children. *Language, Speech, and Hearing Services in Schools*, *55*(2), 1-21. https://doi.org/10.1044/2024_LSHSS-23-00100

Hiebert, L., & Rojas, R. (2021). A longitudinal study of Spanish language growth and loss in young Spanish-English bilingual children. *Journal of communication Disorders*, *92*, 106110. https://doi.org/10.1016/j.jcomdis.2021.106110

Hoff, E. (2006). How social contexts support and shape language development. *Developmental Review*, *26*(1), 55-88. https://doi.org/10.1016/j.dr.2005.11.002

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

Höhle, B., Fritzsche, T., Meß, K., Philipp, M., & Gafos, A. (2020). Only the right noise? Effects of phonetic and visual input variability on 14-month-olds' minimal pair word learning. *Developmental Science*, *23*(5). https://doi.org/10.1111/desc.12950

Hunt, E., Nang, C., Meldrum, S., & Armstrong, E. (2022). Can dynamic assessment identify language disorder in multilingual children? Clinical applications from a systematic review. *Language, Speech, and Hearing Services in Schools*, *53*(2), 598-625. https://doi.org/10.1044/2021_LSHSS-21-00094

Huurneman, B., & Boonstra, F. N. (2016). Assessment of near visual acuity in 0-13 year olds with normal and low vision: A systematic review. *BMC Ophthalmology*, *16*(1), 215. https://doi.org/10.1186/s12886-016-0386-y

Hwa-Froelich, D. A., & Matsuo, H. (2005). Vietnamese children and language-based processing tasks. *Language, Speech, and Hearing Services in the Schools*, *36*(3), 230-243. https://doi.org/10.1044/0161-1461(2005/023)

Hyvarinen, L., Nasanen, R., & Laurinen, P. (1980). New visual acuity test for pre-school children. *Acta Ophthalmologica*, *58*(4), 507-511. https://doi.org/10.1111/j.1755-3768.1980.tb08291.x

Individuals with Disabilities Education Improvement Act, § U.S.C. § 300.309 (2004).

Jackson, E., Leitão, S., Claessen, M., & Boyes, M. (2021). Word learning and verbal working memory in children with Developmental Language Disorder. *Autism & Developmental Language Impairments*, *6*, 1-20. https://doi.org/10.1177/23969415211004109

Jacobs, E. L. (1998). KIDTALK: A computerized language screening test. *Journal of Computing in Childhood Education*, *9*(2), 113-131.

Jensen, E., Jones, N., Rabe, M., Pratt, B., Medina, L., Orozco, K., & Spell, L. (2021). *The chance that two people chosen at random are of different race or ethnicity groups has increased since 2010*. United States Census Bureau. https://www.census.gov/library/stories/2021/08/2020-united-states-population-more-racially-ethnically-diverse-than-2010.html

Kapantzoglou, M., Brown, J. E., Cycyk, L. M., & Fergadiotis, G. (2021). Code-switching and language proficiency in bilingual children with and without Developmental Language Disorder. *Journal of Speech, Language, and Hearing Research*, *64*(5), 1605-1620. https://doi.org/10.1044/2020_JSLHR-20-00182

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

Kapantzoglou, M., Restrepo, M. A., Gray, S., & Thompson, M. S. (2015). Language ability groups in bilingual children: A latent profile analysis. *Journal of Speech, Language, and Hearing Research*, *58*(5), 1549-1562. https://doi.org/10.1044/2015_JSLHR-L-14-0290

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

Kohnert, K., & Danahy, K. (2007). Young L2 learners' performance on a novel morpheme task. *Clinical Linguistics & Phonetics*, *21*(7), 557-569. https://doi.org/10.1080/02699200701374231

Kovács, Á. M., & Mehler, J. (2009). Flexible learning of multiple speech structures in bilingual infants. *Science*, *325*(5940), 611-612. https://doi.org/10.1126/science.1173947

Kramer, K., Mallett, P., Schneider, P., & Hayward, D. (2009). Dynamic assessment of narratives with grade 3 children in a First Nations community. *Canadian Journal of Speech-Language Pathology and Audiology*, *33*(3), 119-128. https://link.gale.com/apps/doc/A220469499/AONE?u=uarizona_main&sid=bookmark-AONE&xid=cc1bf2ba

Kremin, L. V., & Byers-Heinlein, K. (2021). Why not both? Rethinking categorical and continuous approaches to bilingualism. *International Journal of Bilingualism*, *25*(6), 1560-1575. https://doi.org/10.1177/13670069211031986

Krogstad, J. M., Passel, J. S., Moslimani, M., & Noe-Bustamante, L. (2023). *Key facts about U.S. Latinos for national Hispanic heritage month*. Pew Research Center. https://www.pewresearch.org/short-reads/2023/09/22/key-facts-about-us-latinos-for-national-hispanic-heritage-month/#:~:text=Hispanics%20have%20played%20a%20major,other%20racial%20or%20ethnic%20group.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978-990. https://doi.org/10.3758/s13428-012-0210-4

Laketa, A., Studenica, A., Chrysochoou, E., Blakey, E., & Vivas, A. B. (2021). Biculturalism, linguistic distance, and bilingual profile effects on the bilingual influence on cognition: A comprehensive multipopulation approach. *Journal of Experimnetal Psychology: General*, *150*(11), 2273-2292. https://doi.org/10.1037/xge0000794

Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (2017). Statistical learning in Specific Language Impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research*, *60*(12), 3474-3486. https://doi.org/10.1044/2017_JSLHR-L-16-0439

Language, & Reading Research Consortium. (2017). Oral language and listening comprehension: Same or different constructs? *Journal of Speech, Language, and Hearing Research*, *60*(5), 1273-1284. https://doi.org/10.1044/2017_JSLHR-L-16-0039

Language and Reading Research Consortium, Yeomans-Maldonado, G., Bengochea, A., & Mesa, C. (2018). The dimensionality of oral language in kindergarten Spanish-English dual language learners. *Journal of Speech, Language, and Hearing Research*, *61*(11), 2779-2795. https://doi.org/10.1044/2018_JSLHR-L-17-0320

Language Reading Research Consortium. (2015). The dimensionality of language ability in young children. *Child Development*, *86*(6), 1948-1965. https://doi.org/10.1111/cdev.12450

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

Lee, Y.-Y. (2022). A conceptual analysis of typological distance and its potential consequences on the bilingual brain. *International Journal of Bilingual Education and Bilingualism*, *25*(9), 3333-3346. https://doi.org/10.1080/13670050.2022.2052790

Liaw, A., & Wiener, M. (2002). Classificiation and regression by randomForest. *R News*, *2*(3), 18-22.

Lidz, C. S. (1987). *Dynamic assessment: An interactional approach to evaluating learning potential*. Guilford Press. https://arizona-primo.hosted.exlibrisgroup.com/permalink/f/6ljalh/01UA_ALMA21436432710003843

Lidz, C. S. (1991). *Practitioner's guide to dynamic assessment*. Guilford Press. https://arizona-primo.hosted.exlibrisgroup.com/permalink/f/evot53/01UA_ALMA51781781120003843

Lidz, C. S. (2014). Leaning toward a consensus about dynamic assessment: Can we? Do we want to? *Journal of Cognitive Education and Psychology*, *13*(3), 292-307. https://doi.org/10.1891/1945-8959.13.3.292

Lucero, A. (2018). The development of bilingual narrative retelling among Spanish-English dual language learners over two years. *Language, Speech, and Hearing Services in Schools*, *49*(3), 607-621. https://doi.org/10.1044/2018_LSHSS-17-0152

Lugo-Neris, M. J., Pena, E. D., Bedore, L. M., & Gillam, R. B. (2015). Utility of a language screening measure for predicting risk for language impairment in bilinguals. *American Journal of Speech-Language Pathology*, *24*(3), 426-437. https://doi.org/10.1044/2015_AJSLP-14-0061

Luk, G., & Bialystok, E. (2013). Bilingualism is not a categorical variable: Interaction between language proficiency and usage. *Journal of Cognitive Psychology*, *25*(5), 605-621. https://doi.org/10.1080/20445911.2013.795574

Makowski, D., Ben-Shachar, M., & Lüdecke, D. (2019a). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software, 4*(40). https://doi.org/10.21105/joss.01541

Makowski, D., Ben-Shachar, M. S., Chen, S. H. A., & Ludecke, D. (2019b). Indices of effect existence and significance in the Bayesian framework. *Frontiers in Psychology*, *10*, 2767. https://doi.org/10.3389/fpsyg.2019.02767

Malins, J. G., Landi, N., Ryherd, K., Frijters, J. C., Magnuson, J. S., Rueckl, J. G., Pugh, K. R., Sevcik, R., & Morris, R. (2021). Is that a pibu or a pibo? Children with reading and language deficits show difficulties in learning and overnight consolidation of phonologically similar pseudowords. *Developmental Science*, *24*(2), e13023. https://doi.org/10.1111/desc.13023

Maragkaki, I., & Hessels, M. (2016). A pilot study of dynamic assessment of vocabulary in German for bilingual preschoolers in Switzerland. *Journal of Studies in Education, 7*(1), 33-49. https://doi.org/10.5296/jse.v7i1.10392

Marian, V., & Hayakawa, S. (2021). Measuring bilingualism: The quest for a "Bilingualism Quotient". *Applied Psycholinguistics*, *42*(Suppl 2), 527-548. https://doi.org/10.1017/s0142716420000533

Martin, S. R., & Williams, D. R. (2017). Outgrowing the procrustean bed of normality: The utility of Bayesian modeling for asymetrical data analysis. *PsyArXiv preprint*. https://doi.org/https://doi.org/10.31234/osf.io/26m49

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101-B111. https://doi.org/10.1016/S0010-0277(01)00157-3

Mayer, M. (1969). *Frog, Where Are You?* Dial.

Mayer, M. (1975). *Frog on His Own*. Dial.

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

McGregor, K. K., Arbisi-Kelm, T., Eden, N., & Oleson, J. (2020). The word learning profile of adults with Developmental Language Disorder. *Autism & Developmental Language Impairments*, *5,* 1-19. https://doi.org/10.1177/2396941519899311

McGregor, K. K., Smolak, E., Jones, M., Oleson, J., Eden, N., Arbisi-Kelm, T., & Pomper, R. (2022). What children with Developmental Language Disorder teach us about cross-situational word learning. *Cognitive Science*, *46*(2), e13094. https://doi.org/10.1111/cogs.13094

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessent. *Educational Researcher, 18*(2), 5-11. https://doi.org/10.3102/0013189X018002005

Meyers-Denman, C. N., & Plante, E. (2016). Dose schedule and enhanced conversational recast treatment for children with Specific Language Impairment. *Language, Speech, and Hearing Services in Schools*, *47*(4), 334-346. https://doi.org/10.1044/2016_LSHSS-15-0064

Miller, F., & Iglesias, A. (2020). *Systematic Analysis of Language Transcripts (SALT)* [Computer Software]. (Version 20) SALT Software, LLC.

Minick, N. (1987). Implications of Vygotsky's theories for dynamic assessment. In C. S. Lidz (1st Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 116-140). The Guilford Press.

Mora, L., & Lopez, M. H. (2023). *Latinos' views of and experiences with the Spanish language*. https://www.pewresearch.org/race-ethnicity/2023/09/20/latinos-views-of-and-experiences-with-the-spanish-language/

Morgan, P. L. (2021). Unmeasured confounding and racial or ethnic disparities in disability identification. *Educational Evaluation and Policy Analysis*, *43*(2), 351-361. https://doi.org/10.3102/0162373721991575

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*, *84*(1), 27-41. https://doi.org/10.1177/0014402917718341

Morgan, P. L., Farkas, G., Hillemeier, M. M., Mattison, R., Maczuga, S., Li, H., & Cook, M. (2015). Minorities are disproportionately underrepresented in special education: Longitudinal evidence across five disability conditions. *Educational Researcher*, *44*(5), 278-292. https://doi.org/10.3102/0013189X15591157

Mullerhein, S. A., & Miller, W. C. (2002). Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. *Annals of Internal Medicine*, *137*(7), 598-602. https://doi.org/10.7326/0003-4819-137-7-200210010-00011

Narayanan, T. L., & Ramsdell, H. L. (2022). Self-reported confidence and knowledge-based differences between multilingual and monolingual Speech-Language Pathologists when serving culturally and linguistically diverse populations. *Perspectives of the ASHA Special Interest Groups*, *7*(1), 209-228. https://doi.org/10.1044/2021_persp-21-00169

Nelson, N. W. (2000). Basing eligibility on discrepancy criteria: A bad idea whose time has passed. *Language Learning and Education*, *7*(1), 8-12. https://doi.org/10.1044/lle7.1.8

Nicholas, K., Alt, M., & Hauwiller, E. (2019). Variability of input in preposition learning by preschoolers with Developmental Language Disorder and typically-developing language. *Child Language Teaching and Therapy, 35*(1), 55-74. https://doi.org/10.1177/0265659019830455

Nicholas, K., Plante, E., Gomez, R., & Vance, R. (2021). The role of spontaneous repetitions during treatment of morphosyntactic forms for children with Developmental Language Disorder. *Journal of Speech, Language, and Hearing Research, 64*(10), 3995-4003. https://doi.org/10.1044/2021_JSLHR-20-00367

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. https://doi.org/10.1044/1058-0360(2008/07-0049)

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, 57*(11), 1247-1257. https://doi.org/10.1111/jcpp.12573

Obeid, R., Brooks, P. J., Powers, K. L., Gillespie-Lynch, K., & Lum, J. A. (2016). Statistical learning in Specific Language Impairment and Autism Spectrum Disorder: A meta-analysis. *Frontiers in Psychology, 7*, 1245. https://doi.org/10.3389/fpsyg.2016.01245

Ogiela, D. A., & Montzka, J. L. (2021). Norm-referenced language test selection practices for elementary school children with suspected Developmental Language Disorder. *Language, Speech, and Hearing Services in Schools, 52*(1), 288-303. https://doi.org/10.1044/2020_LSHSS-19-00067

Onnis, L., Chun, W. E., & Lou-Magnuson, M. (2017). Improved statistical learning abilities in adult bilinguals. *Bilingualism: Language and Cognition*, *21*(2), 427-433. https://doi.org/10.1017/s1366728917000529

Orellana, C. I., Wada, R., & Gillam, R. B. (2019). The use of dynamic assessment for the diagnosis of language disorders in bilingual children: A meta-analysis. *American Journal of Speech-Language Pathology*, *28*(3), 1298-1317. https://doi.org/10.1044/2019_AJSLP-18-0202

Owen Van Horne, A. J., Fey, M., & Curran, M. (2017). Do the hard things first: A randomized controlled trial testing the effects of exemplar selection on generalization following therapy for grammatical morphology. *Journal of Speech, Language, and Hearing Research*, *60*(9), 2569-2588. https://doi.org/10.1044/2017_JSLHR-L-17-0001

Paradis, J., & Genesee, F. (1996). Syntactic acquisition in bilingual children: Autonomous or interdependent? *Studies in Second Language Acquisition*, *18*(1), 1-25. https://doi.org/10.1017/S0272263100014662

Paradis, J., & Navarro, S. (2003). Subject realization and crosslinguistic interference in the bilingual acquisition of Spanish and English: What is the role of the input? *Journal of Child Language*, *30*(2), 371-393. https://doi.org/10.1017/s0305000903005609

Parveen, S., & Santhanam, S. P. (2021). Speech-Language Pathologists' perceived competence in working with culturally and linguistically diverse clients in the United States. *Communication Disorders Quarterly*, *42*(3), 166-176. https://doi.org/10.1177/1525740120915205

Paul, R., & Alforde, S. (1993). Grammatical morpheme acquisition in 4-year-olds with normal, impaired, and late-developing language. *Journal of Speech and Hearing Research*, *36*(6), 1271-1275. https://doi.org/10.1044/jshr.3606.1271

Pavlidou, E. V., Williams, J. M., & Kelly, L. M. (2009). Artificial grammar learning in primary school children with and without developmental dyslexia. *Annals of Dyslexia*, *59*(1), 55-77. https://doi.org/10.1007/s11881-009-0023-z

Peña, E. D. (2000). Measurement of modifiability in children from culturally and linguistically diverse backgrounds. *Communication Disorders Quarterly*, *21*(2), 87-97. https://doi.org/10.1177/152574010002100203

Peña, E. D., Bedore, L., Gutiérrez-Clellen, V. F., Iglesias, A., & Goldstein, B. (2008, in development). *Bilingual English-Spanish Oral Screener (BESOS)*.

Peña, E. D., Bedore, L. M., & Vargas, A. G. (2023). Exploring assumptions of the bilingual delay in children with and without Developmental Language Disorder. *Journal of Speech, Language, and Hearing Research*, *66*(12), 4739-4755. https://doi.org/10.1044/2023_JSLHR-23-00117

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

Peña, E. D., Gillam, R. B., Malek, M., Ruiz-Felter, R., Resendiz, M., Fiestas, C., & Sabel, T. (2006). Dynamic assessment of school-age chidlren's narrative ability: An experimental investigation of classification accuracy. *Journal of Speech, Language, and Hearing Research*, *49*(5), 1037-1057. https://doi.org/10.1044/1092-4388(2006/074)

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

Peña, E. D., Iglesias, A., Gutierrez-Clellen, V. F., Goldstein, B., & Bedore, L. M. (2013). *Development of a bilingual test for Spanish-English children: A long and winding road* [Powerpoint presentation]. American Speech-Language Hearing Association [ASHA]. https://academy.pubs.asha.org/2013/11/steps-in-assessment-development/

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

Peña, E. D., Reséndiz, M., & Gillam, R. B. (2007). The role of clinical judgements of modifiability in the diagnosis of language impairment. *Advances in Speech Language Pathology*, *9*(4), 332-345. https://doi.org/10.1080/14417040701413738

Petersen, D. B., Chanthongthip, H., Ukrainetz, T. A., Spencer, T. D., & Steeve, R. W. (2017). Dynamic assessment of narratives: Efficient, accurate identification of language impairment in bilingual students. *Journal of Speech, Language, and Hearing Research*, *60*(4), 983-998. https://doi.org/10.1044/2016_JSLHR-L-15-0426

Petersen, D. B., Tonn, P., Spencer, T. D., & Foster, M. E. (2020). The classification accuracy of a dynamic assessment of inferential word learning for bilingual English/Spanish-speaking school-age children. *Language, Speech, and Hearing Services in Schools*, *51*(1), 144-164. https://doi.org/10.1044/2019_LSHSS-18-0129

Plante, E., & Gómez, R. L. (2018). Learning without trying: The clinical relevance of statistical learning. *Language, Speech, and Hearing Services in Schools*, *49*(3S), 710-722. https://doi.org/10.1044/2018_LSHSS-STLT1-17-0131

Plante, E., Mettler, H. M., Tucci, A., & Vance, R. (2019). Maximizing treatment efficiency in Developmental Language Disorder: Positive effects in half the time. *American Journal of Speech-Language Pathology*, *28*(3), 1233-1247. https://doi.org/10.1044/2019_AJSLP-18-0285

Plante, E., Ogilvie, T., Vance, R., Aguilar, J. M., Dailey, N. S., Meyers, C., Lieser, A. M., & Burton, R. (2014). Variability in the language input to children enhances learning in a treatment context. *American Journal of Speech-Language Pathology*, *23*(4), 530-545. https://doi.org/10.1044/2014_AJSLP-13-0038

Plante, E., Tucci, A., Nicholas, K., Arizmendi, G. D., & Vance, R. (2018). Effective use of auditory bombardment as a therapy adjunct for children with Developmental Language Disorders. *Language, Speech, and Hearing Services in Schools*, *49*(2), 320-333. https://doi.org/10.1044/2017_LSHSS-17-0077

Plante, E., & Vance, R. (1994). Selection of preschool language tests: A data-based approach. *Language Speech and Hearing Services in Schools*, *25*(1), 15-24. https://doi.org/10.1044/0161-1461.2501.15

Poepsel, T. J., & Weiss, D. J. (2016). The influence of bilingualism on statistical word learning. *Cognition*, *152*, 9-19. https://doi.org/10.1016/j.cognition.2016.03.001

R Core Team. (2021). *R: A language and environment for statistical computing*. In R Foundation for Statistical Computing. https://www.R-project.org/

Ransohoff, D. F., & Feinstein, A. R. (1978). Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *The New England Journal of Medicine*, *299*(17), 926-930. https://doi.org/10.1056/NEJM197810262991705

Reber, A. S. (1967). Implicit learning of artificial grammar. *Journal of Verbal Learning and Verbal Behavior*, *6*, 855-863. https://doi.org/10.1016/s0022-5371(67)80149-x

Reeder, P. A., Newport, E. L., & Aslin, R. N. (2013). From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive Psychology*, *66*(1), 30-54. https://doi.org/10.1016/j.cogpsych.2012.09.001

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

Reynolds, C. R. (1981). The fallacy of "two years below grade level for age" as a diagnostic criterion for reading disorders. *Journal of School Psychology*, *19*(4), 350-358. https://doi.org/10.1016/0022-4405(81)90029-7

Rice, M. L., Oetting, J. B., Marquis, J., Bode, J., & Pae, S. (1994). Frequency of input effects on word comprehension of children with Specific Language Impairment. *Journal of Speech and Hearing Research*, *37*(1), 106-122. https://doi.org/10.1044/jshr.3701.106

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(77), 1-8. https://doi.org/10.1186/1471-2105-12-77

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., Müller, M., Siegert, S., Doering, M., & Billings, Z. (2023). *Package pROC* [R documentation]. https://cran.r-project.org/web/packages/pROC/pROC.pdf

Roseberry, C. A., & Connell, P. J. (1991). The use of an invented language rule in the differentiation of normal and language-impaired Spanish-speaking children. *Journal of Speech and Hearing Research*, *34*(3), 596-603. https://doi.org/10.1044/jshr.3403.596

Rutjes, A. W., Reitsma, J. B., Vandenbroucke, J. P., Glas, A. S., & Bossuyt, P. M. (2005). Case-control and two-gate designs in diagnostic accuracy studies. *Clinical Chemistry*, *51*(8), 1335-1341. https://doi.org/10.1373/clinchem.2005.048595

Sabater, L., Guasch, M., Ferre, P., Fraga, I., & Hinojosa, J. A. (2020). Spanish affective normative data for 1,406 words rated by children and adolescents (SANDchild). *Behavior Reseach Methods*, *52*(5), 1939-1950. https://doi.org/10.3758/s13428-020-01377-5

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*(4), 606-621. https://doi.org/10.1006/jmla.1996.0032

SALT Software LLC. (n.d.-a, January 2017). *Frog story task: Frog stories*. Retrieved August 11, 2023 from https://www.saltsoftware.com/resources/elicaids/frogstories

SALT Software LLC. (n.d.-b). *Self-paced online courses*. https://www.saltsoftware.com/training/self-paced-online-training

Sandoval, M., Patterson, D., Dai, H., Vance, C. J., & Plante, E. (2017). Neural correlates of morphology acquisition through a statistical learning paradigm. *Frontiers in Psychology*, *8*, 1234. https://doi.org/10.3389/fpsyg.2017.01234

Santhanam, S. P., & Parveen, S. (2018). Serving culturally and linguistically diverse clients: A review of changing trends in speech-language pathologists' self-efficacy and implications for stakeholders. *Clinical Archives of Communication Disorders, 3*(3), 165-177. https://doi.org/10.21849/cacd.2018.00395

Scharff Rethfeldt, W., McNeilly, L., Laasonen, M., Meir, N., Abutbul-Oz, H., Smolander, S., Niegia Garcia Goulart, B., & Frances Hunt, E. (2023). Assessment of Developmental

Language Disorder in multilingual children: Results from an international survey. *Folia Phoniatria et Logoppaedica*, *76*(2), 127-150. https://doi.org/10.1159/000533139

Semel, E., Wiig, E. H., & Secord, W. A. (2006). Clinical Evaluation of Language Fundamentals–4th edition, Spanish. San Antonio, TX: The Psychological Corporation.

Shriberg, L. (1993). Four new speech and prosody-voice measures for genetics research and other studies in developmental phonological disorders. *Journal of Speech and Hearing Research*, *36*(1), 105-140. https://doi.org/10.1044/jshr.3601.105

Siegelman, N. (2019). Statistical learning abilities and their relation to language. *Language and Linguistics Compass*, *14*(3). https://doi.org/10.1111/lnc3.12365

Smith, K., Smith, A. D. M., & Blythe, R. A. (2011). Cross-situational learning: An experimental study of word-learning mechanisms. *Cognitive Science*, *35*(3), 480-498. https://doi.org/10.1111/j.1551-6709.2010.01158.x

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558-1568. https://doi.org/10.1016/j.cognition.2007.06.010

Snow, P., & Powell, M. (2004). Developmental Language Disorders and adolescent risk: A public-health advocacy role for speech pathologists? *Advances in Speech Language Pathology*, *6*(4), 221-229. https://doi.org/10.1080/14417040400010132

Subcommittee on Attention-Deficit/Hyperactivity, D., Steering Committee on Quality, I., Management, Wolraich, M., Brown, L., Brown, R. T., DuPaul, G., Earls, M., Feldman, H. M., Ganiats, T. G., Kaplanek, B., Meyer, B., Perrin, J., Pierce, K., Reiff, M., Stein, M. T., & Visser, S. (2011). ADHD: Clinical practice guideline for the diagnosis, evaluation, and treatment of attention-deficit/hyperactivity disorder in children and adolescents. *Pediatrics*, *128*(5), 1007-1022. https://doi.org/10.1542/peds.2011-2654

Sullivan, A. L., & Bal, A. (2013). Disproportionality in special education: Effects of individual and school variables on disability risk. *Exceptional Children*, *79*(4), 475-494. https://doi-org.ezproxy2.library.arizona.edu/10.1177/001440291307900406

Suswaram, S., Brady, N. C., & Gillispie, M. (2023). The role of service providers' linguistic backgrounds on assessment of multilingual children. *Journal of Communication Disorders*, *102*, 106302. https://doi.org/10.1016/j.jcomdis.2023.106302

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology*, *39*(4), 706-716. https://doi.org/10.1037/0012-1649.39.4.706

Thompson, S. P., & Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development*, *3*(1), 1-42. https://doi.org/10.1080/15475440709336999

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, & Hearing Research*, *40*(6), 1245-1260. https://doi.org/10.1044/jslhr.4006.1245

Torkildsen, J., Dailey, N. S., Aguilar, J. M., Gómez, R., & Plante, E. (2013). Exemplar variability facilitates rapid learning of an otherwise unlearnable grammar by individuals with language-based learning disability. *Journal of Speech, Language, and Hearing Research*, *56*(2), 618-629. https://doi.org/10.1044/1092-4388(2012/11-0125)

Twomey, K. E., Ma, L., & Westermann, G. (2017). All the right noises: Background variability helps early word learning. *Cognitive Science, 42*(S2), 413-438. https://doi.org/10.1111/cogs.12539

Ukrainetz, T. A., Harpell, S., Walsh, C., & Coyle, C. (2000). A preliminary investigation of dynamic assessment with Native American kindergartners. *Language, Speech, and Hearing Services in Schools*, *31*(2), 142-154. https://doi.org/10.1044/0161-1461.3102.142

United States Census Bureau. (n.d.). *QuickFacts: Tucson city, Arizona*. Retrieved April 15, 2024 from https://www.census.gov/quickfacts/fact/table/tucsoncityarizona/RHI725222#RHI725222

van de Schoot, R., Depaoli, S., Gelman, A., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Willemsen, J., & Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, *1*(3). https://doi.org/10.1038/s43586-020-00003-0

Volterra, V., & Taeschner, T. (1978). The acquisition and development of language by bilingual children. *Journal of Child Language*, *5*(2), 311–326. https://doi.org/10.1017/S0305000900007492

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (V. J.-S. Michael Cole, Sylvia Scribner, Ellen Souberman, Ed.). Harvard University Press.

Vygotsky, L. S. (1998). The problem of age (M. J. Hall, Trans.). R. W. Rieber (Ed.), *The collected works of L. S. Vygotsky: Child psychology* (Vol. 5, pp. 187-205). Springer.

West, G., Melby-Lervag, M., & Hulme, C. (2021). Is a procedural learning deficit a causal risk factor for Developmental Language Disorder or dyslexia? A meta-analytic review. *Developmental Psychology*, *57*(5), 749-770. https://doi.org/10.1037/dev0001172

Whitehouse, A. J., Line, E. A., Watt, H. J., & Bishop, D. V. (2009). Qualitative aspects of developmental language impairment relate to language and literacy outcome in adulthood. *International Journal of Language and Communication Disorders*, *44*(4), 489-510. https://doi.org/10.1080/13682820802708080

Whitehouse, A. J., Watt, H. J., Line, E. A., & Bishop, D. V. (2009). Adult psychosocial outcomes of children with specific language impairment, pragmatic language impairment and autism. *International Journal of Language and Communication Disorders*, *44*(4), 511-528. https://doi.org/10.1080/13682820802708098

Whiting, P. F., Rutjes, A. W. S., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., Leeflang, M. M. G., Sterne, J. A. C., & Bossuyt, P. M. M. (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine*, *155*(8), 529-536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009

Witt, A., & Vinter, A. (2012). Artificial grammar learning in children: Abstraction of rules or sensitivity to perceptual features? *Psychological Research, 76*(1), 97-110. https://doi.org/10.1007/s00426-011-0328-5

Yow, W. Q., Tan, J. S. H., & Flynn, S. (2017). Code-switching as a marker of linguistic competence in bilingual children. *Bilingualism: Language and Cognition*, *21*(5), 1075-1090. https://doi.org/10.1017/s1366728917000335

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 369-468. https://doi.org/10.1111/j.1467-9280.2007.01915.x