ASHA American Speech-Language-Hearing Association

JSLHR

Review Article

The Use of Language Sample Analysis to Differentiate Developmental Language Disorder From Typical Language in Bilingual Children: A Systematic Review and Meta-Analysis

José A. Ortiz,[a] Jessica M. Nolasco,[a] Yi Ting Huang,[a] and Jason C. Chow[b]

[a] Department of Hearing and Speech Sciences, University of Maryland, College Park [b] Department of Special Education, Vanderbilt University, Nashville, TN

ARTICLE INFO

Article History:
Received March 29, 2024
Revision received May 30, 2024
Accepted June 25, 2024

Editor-in-Chief: Julie A. Washington
Editor: Raúl Rojas

https://doi.org/10.1044/2024_JSLHR-24-00212

ABSTRACT

Purpose: Language sample analysis (LSA) is a commonly recommended method of assessment for bilingual children. This systematic review and meta-analysis provides a comprehensive overview of the literature on the use of LSA to differentiate between developmental language disorder (DLD) and typical language (TL) in bilingual children.
Method: We conducted a search of several large electronic databases along with forward and backward searches and applied abstract and full-text screening procedures to identify all relevant studies. We then estimated standardized mean differences, representing the ability of LSA to differentiate between DLD and TL, using multilevel model and subgroup and moderator analyses to identify characteristics of LSA that may be associated with differences in effect size magnitude. We conducted assessments of publication bias and risk of bias by examining quality indicators for each study.
Results: The search yielded 35 articles that met the inclusion criteria. Participants ranged in age from 2;0 (years;months) to 11;9, with over 40 languages represented. Across studies, the pooled standardized mean difference indicated that children with DLD performed 0.78 *SD* lower on LSA measures than those with TL. Measures of morphosyntactic accuracy exhibited the largest pooled effect size. Elicitation method, language of task, and age were not associated with differences in effect size.
Discussion: Results of this study provide evidence of the clinical utility of LSA in differentiating between DLD and TL in bilingual children. Further research is needed to examine classification accuracy as well as task characteristics that may improve its diagnostic utility.

There is an ongoing need for improved methods to accurately identify developmental language disorder (DLD) in bilingual children. Many existing language assessment tools do not adequately capture the linguistic ability of bilingual children, leading to inaccurate clinical impressions and potentially biased outcomes (De Lamo White & Jin, 2011). Concerns about possible misidentification have led to the development of a range of alternative assessment methods, such as dynamic assessment (Orellana et al., 2019) and nonword repetition (Ortiz, 2021), as well as the adapted applications of traditional assessment methods to better account for the unique language experiences of bilinguals, such as parent report (e.g., Paradis et al., 2010). Language sample analysis (LSA), sometimes referred to as the gold standard of language assessment (Miller et al., 2016; Ramos et al., 2022), is often recommended for bilingual children due to its flexibility and its ability to accurately

Correspondence to José A. Ortiz: jortiz5@umd.edu. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

measure the language ability of children across different linguistic contexts (e.g., Castilla-Earls et al., 2020; Ebert, 2020).

LSA demonstrates several appealing characteristics relative to other methods of assessment. First, LSA is unique in its ability to provide information with direct clinical applications that would be otherwise difficult to obtain (Ebert, 2020; Hewitt et al., 2005; Ramos et al., 2022). Unlike some forms of assessment, information from language samples can be used in the development of treatment goals. For example, LSA can be used to easily identify targets for intervention, such as difficulty with specific grammatical forms, that other measures (e.g., nonword repetition, standardized tests) may be unable to isolate. Another potential advantage of LSA is its inherent flexibility; language samples can be elicited in a variety of different ways, such as play (e.g., De Anda et al., 2023), conversation (e.g., Paradis et al., 2022), or narratives (e.g., Guiberson, 2020). Because of this flexibility, LSA can be tailored to meet the needs of individual clients and, as a result, does not suffer from the requirement to adhere to a defined protocol, unlike standardized tests. Lastly, many speech-language pathologists (SLPs) are familiar with LSA (Arias & Friberg, 2017; Pavelko et al., 2016), making it one of the most readily available assessment methods in the field. This is a major advantage considering the barriers to acquiring and learning how to administer many assessment tools.

LSA demonstrates several qualities that make it an appealing choice for language assessment broadly, but it may be particularly well suited for use with bilingual children. One of the main strengths of LSA is its ecological validity (Ebert, 2020; Hewitt et al., 2005; Ramos et al., 2022), which distinguishes it from other forms of assessment recommended for use with bilinguals, such as processing-dependent measures (e.g., nonword repetition; Ortiz, 2021). The accurate assessment of language in bilinguals requires the examination of skills across languages (Peña et al., 2016), and LSA is an effective tool for this purpose. Because of its potential to provide rich information, LSA can accurately characterize cross-linguistic skills in ways that other assessment methods cannot. For example, the flexible administration of LSA also means that it can be used in the absence of standardized tests designed for bilinguals. If there is no test available for speakers of a given language, language sample elicitation offers a means of collecting descriptive information about language ability. Because SLPs can collect samples in any language, all clinicians, with the assistance of interpreters and translators, can use LSA to measure ability in several linguistic domains using a range of metrics, such as mean length of utterance and number of different words. In many instances, these metrics can be used to develop a profile of language ability by comparing them to existing normative databases found in several software packages, including Computerized Language Analysis (CLAN; MacWhinney, 2000), Systematic Analysis of Language Transcripts (SALT; Miller & Iglesias, 2012), and Sampling Utterances and Grammatical Analysis Revised (SUGAR; Pavelko & Owens, 2023). Lastly, the inclusion of LSA may improve the diagnostic accuracy of an assessment battery (Castilla-Earls et al., 2020), indicating that it provides information above and beyond that which other assessment methods offer. For example, the pairing LSA with standardized assessment can yield an improvement in diagnostic accuracy compared to standardized testing alone (Lazewnik et al., 2019).

Considerations for Elicitation and Analysis

LSA is a ubiquitous assessment method among SLPs and has been described as "one of the most valuable resources in the language clinician's toolkit" (Ebert, 2020, p. 182). Despite being widely recognized for the value it provides, there are still questions about best practices for undertaking LSA and specific approaches that may improve its clinical utility. In a recent systematic review of 28 studies, Ramos et al. (2022) summarized the evidence of efficacy of LSA in the identification of DLD in monolinguals and found that classification accuracy ranged from poor to good (25%–90%). Much of the variability in diagnostic accuracy can be attributed to differences in the way that language samples were analyzed, given the wide variety of possible metrics to choose from. In their review, Ramos et al. identified 46 unique measures that were used to quantify LSA outcomes, across a range of linguistic domains. Although most studies focused on microstructure, several included measures of narrative macrostructure. The analysis of microstructure is concerned with the measurement of morphosyntax and semantics and includes areas such as length, accuracy, proficiency, and semantic productivity. Common microstructure metrics include mean length of utterance and number of different words. The analysis of macrostructure, on the other hand, is concerned with the examination of language samples collected through narrative tasks to identify the presence of specific elements not captured in an analysis of microstructure. Measures of macrostructure may include an examination of story grammar elements (e.g., Fichman et al., 2017), mental state terms (e.g., Altman et al., 2016), or causal relations (e.g., Kupersmitt & Armon-Lotem, 2019).

In addition to the measures used to analyze language samples, the method of elicitation can also take a variety of forms, such as narrative retell (e.g., Lazewnik et al., 2019), play (e.g., De Anda et al., 2023), conversation (e.g., Hewitt et al., 2005), and picture description (e.g., Restrepo, 1998). Narrative retell is one of the most common types of tasks and is frequently recommended

when conducting LSA with bilinguals (Castilla-Earls et al., 2020; Rojas & Iglesias, 2009; Squires et al., 2014). In contrast to other methods of elicitation, narrative retell tasks offer some advantages because of their relatively structured administration, which may lead to more consistent results across children. In addition, existing LSA databases frequently include narrative samples, such as Child Language Data Exchange System (MacWhinney, 2000) and SALT (Miller & Iglesias, 2012), which facilitates ease of comparison to age-matched peers on a similar task. Despite the popularity of narrative tasks, other types of elicitation methods also demonstrate value (Ebert, 2020). When collecting language samples from young children, for example, a play-based approach may be a more straightforward means of elicitation. For older children, on the other hand, it may be more useful to use elicitation methods that do not rely on picture books (Ebert & Pham, 2017). Expository and persuasive tasks, for example, can yield more complex language than other types of tasks and may be more appropriate for older children (Pezold et al., 2020). In their systematic review, Ramos et al. (2022) identified a range of elicitation methods and found that narrative tasks, including both tell and retell, were the most commonly used methods but that no single elicitation task demonstrated superior diagnostic accuracy. LSA exhibits many appealing characteristics, including its ecological validity, flexibility, and the potential to improve diagnostic accuracy in the identification of DLD in bilinguals.

Purpose

Despite the range of potential benefits that it offers, LSA has seen somewhat limited adoption as a clinical tool relative to other assessment methods. Many school-based SLPs report that they do not use LSA at all, or only do so in a limited capacity (Pavelko et al., 2016). Among SLPs working with bilingual children, 28%–40% report using LSA as part of their assessment batteries, depending on the language being examined (Arias & Friberg, 2017). Instead, many school-based clinicians rely on standardized assessment to make disability determinations (Fulcher-Rood et al., 2018). Limited widespread adoption may be partly attributable to a lack of information regarding best practices for implementation as well as how to derive meaningful diagnostic information from LSA (Ramos et al., 2022). Although these barriers affect the use of LSA for all clients, they present a particular challenge when working with bilinguals, given that bilingual assessment relies heavily on the use of nonstandardized measures. Despite being a commonly recommended assessment method intended to reduce bias, there are several questions regarding how LSA is best used with bilingual children. Given the paucity of existing assessment tools for this population, a better understanding of the ability of LSA to differentiate between DLD and typical language (TL) among bilinguals is needed. Although previous systematic reviews have examined LSA for speakers of English (Ramos et al., 2022) or alternative assessment measures (Dollaghan & Horner, 2011; Orellana et al., 2019; Ortiz, 2021; Schwob et al., 2021), there are no systematic reviews specifically examining the use of LSA in bilingual children. Thus, the aim of this study is to provide a systematic review of the literature on LSA for the purposes of differentiating children with DLD from those with TL. This study seeks to answer the following questions:

1. What is the range of LSA methods that have been examined in studies of DLD in bilingual children?
2. Which LSA methods best differentiate bilingual children with TL from those with DLD?

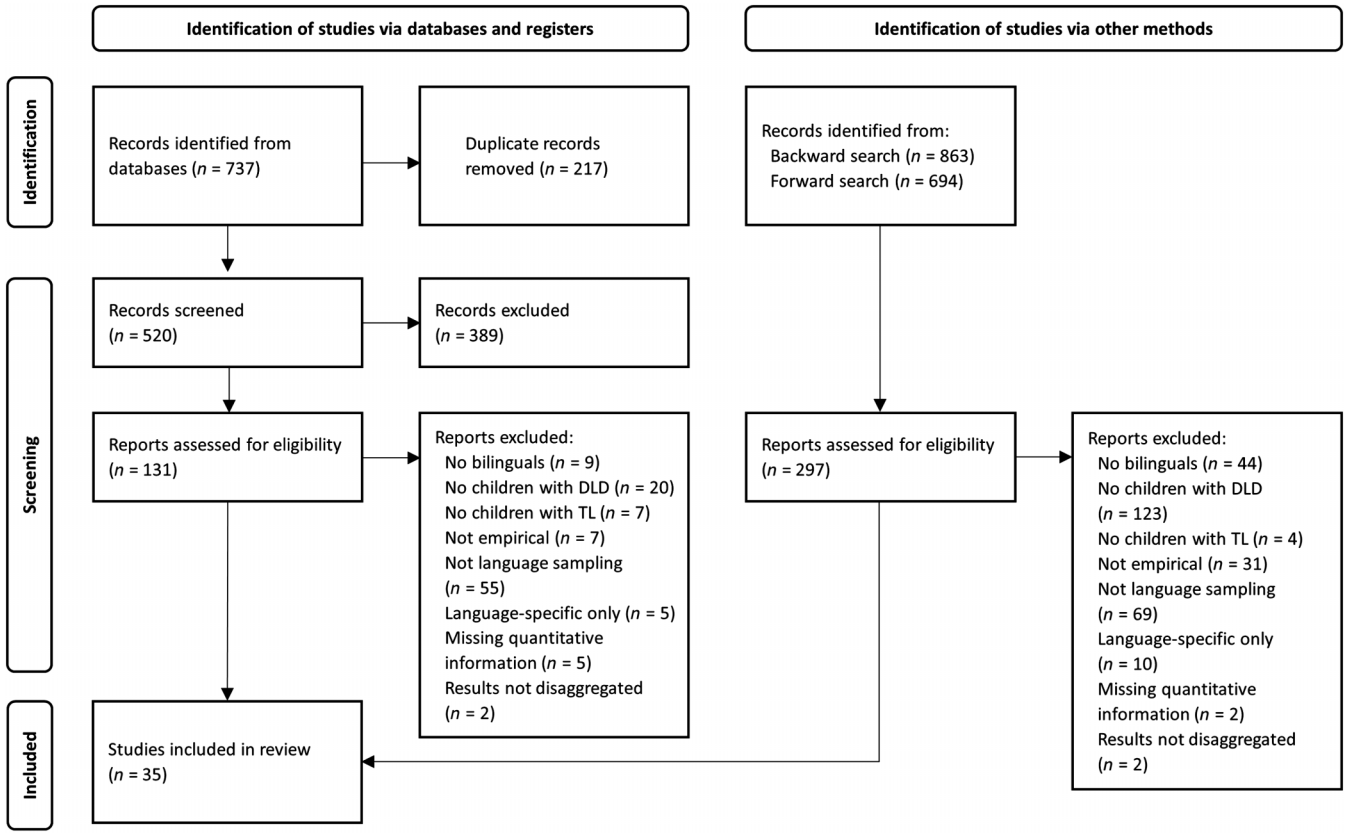
Method

Using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (Page, McKenzie, et al., 2021), the following sections describe the procedures used to identify relevant studies, extract data, evaluate study quality, estimate publication bias, and undertake quantitative analysis. Much of the methodology used in the present study was guided by the *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* (Deeks et al., 2023).

Search Procedure

To identify relevant studies, we conducted a search of several electronic databases, followed by backward and forward searches, as shown in Figure 1. Potential articles were identified based on a predefined set of eligibility criteria, as described in the following section. The search of electronic databases included ERIC, EBSCO Academic Search Ultimate, Medline, ProQuest Dissertations and Theses Global, and PsycINFO. We searched for peer-reviewed articles and dissertations published through April 2023 using a set of key words intended to target the population of interest (i.e., bilinguals), the disorder in question (i.e., DLD), and the method of measurement (i.e., LSA). We adopted a search strategy similar to that used by Ramos et al. (2022) because of the study's similar focus. The database search included the following key words: (bilingual* OR multilingual* OR "dual language learner*") AND ("developmental language disorder" OR DLD OR "language impair*" OR "language disorder*" OR DLD OR SLI OR PLI OR LI) AND ("language sampl*" OR elicitation OR collection OR narrative OR retell OR "index of productive syntax" OR "developmental sentence scor*" OR "mean length of utterance" OR productivity OR "type token ratio" OR "number of different word*" OR "subordination index" OR "argument

Figure 1. Study search. DLD = developmental language disorder; TL = typical language.



structure" OR "lexical measure*" OR grammatical* OR "grammar measure*" OR "syntax measure*" OR "syntactic measure*") AND (classif* OR identif* OR predict* OR discrim* OR differentiate* OR distinguish* OR diagnos* OR "diagnostic accuracy" OR sensitivity OR specificity OR "likelihood ratio*" OR AUC).

We identified potentially relevant studies from the initial database search by reviewing their titles and abstracts. After removing duplicates, we reviewed the full text of these articles to determine whether they met the eligibility criteria. Using the set of studies identified in the database search, we then conducted backward and forward searches to find additional eligible studies. The backward search consisted of searching the references of studies identified in the database search, and the forward search was completed by reviewing all articles that cited these studies. We completed both backward and forward searches using the online tool "Dimensions."

Eligibility Criteria

Prior to beginning our search, we developed a set of eligibility criteria based on the population of interest, method of measurement, and reported outcomes. Inclusionary and exclusionary criteria are described below.

Inclusionary Criteria

The inclusionary criteria are as follows:

- Studies needed to include bilingual children with DLD and TL in their samples. We considered any study that included participants under the age of 18 years. Because the goal of this study was to examine the use of LSA in bilinguals broadly, we did not focus on a specific language. Rather, the only linguistic requirement was that participants were speakers of multiple languages.
- Studies needed to examine the use of language sampling for bilingual children as their index measure.
- Because the goal of the present systematic review was to evaluate measures that could be applied to any language, studies needed to examine aspects of language sampling that were applicable to bilinguals broadly.
- Studies needed to report quantitative outcomes for language sample measures disaggregated from other outcomes. In addition, studies needed to report

quantitative metrics that could be used to derive standardized mean differences.

- Studies needed to be empirical in nature.

Exclusionary Criteria

The exclusionary criteria are as follows:

- Studies that focused exclusively on children with DLD or TL, as well as those that solely included monolinguals, were ineligible.
- Studies that used language sampling as their reference measure, and not as their index measure, were ineligible.
- Studies that examined language-specific elements, which could only be reasonably applied to a single language (e.g., language-specific morphosyntactic elements), were ineligible. To ensure broad applicability to a range of languages, we excluded measures that focused on a single lexical class.
- Studies that reported language sample metrics as part of a composite, in the absence of disaggregated metrics, were ineligible.

The first author completed all searches, and a research assistant independently reviewed 25% of studies at each stage of the search to ensure the reliability of the study selection procedure. The research assistant was first trained on the eligibility criteria through joint review of sets of 10 studies from the database search with the first author, until 90% agreement was achieved. The first author then randomly selected studies for independent review by the research assistant, resulting in 87% agreement following the database search and 94% agreement following a full-text review. Reviewers met to resolve any disagreements in study selection at each stage of the search.

Data Extraction

The first and second authors extracted relevant variables from each study by using a coding matrix in Microsoft Excel. The coding matrix included a range of relevant qualitative and quantitative variables, partially drawn from those used in previous systematic reviews of language assessment methods (Dollaghan & Horner, 2011; Orellana et al., 2019; Ortiz, 2021; Ramos et al., 2022). Variables included study purpose, major findings, country, study design, bilingual classification method, reference measure characteristics, index measure characteristics, sample sizes for DLD and TL, age range, language background, socioeconomic status, and quantitative outcomes. We contacted authors of three primary studies to request clarification on study details and received additional information from all authors.

Coders started by independently extracting data for two randomly selected articles and then reviewed results to come to consensus. This initial coding procedure served to ensure interrater reliability. Following this initial coding, the first author independently recorded data for 100% of the included studies, and the second author independently recorded data for 30% of the included studies, which were selected at random. The coders then jointly reviewed extracted data, identified any disagreements, and came to consensus. We calculated interrater reliability within the set of double-coded articles by dividing the number of individual matrix cells for which no disagreement was present by the total number of cells. Interrater reliability for data extraction was approximately 98%.

Study Quality

We evaluated study quality using indicators adapted from the Critical Appraisal of Diagnostic Evidence Scale (Dollaghan, 2007). These quality indicators represent potential sources of bias in study outcomes and are important to consider when interpreting results. Study quality indicators included sample size, gate design, representativeness of sample, validity and reliability of reference measure, uniformity of reference measure administration across groups, independent administration of reference measure, masked (i.e., blinded) administration of index measure, and reporting of reliability.

Quantitative Analysis

To compare quantitative outcomes across studies, we derived effect sizes from reported means and standard deviations separately for the DLD and TL groups. If studies reported multiple outcomes (e.g., different measures, languages), we recorded those data separately. For studies that reported language sample outcomes using a dynamic assessment approach (i.e., test–teach–retest), only results from the first phase of testing were recorded to ensure consistency in the construct being measured. As a measure of learning capacity, dynamic assessment includes measures from before and after a mediated learning experience, and outcomes from the first test phase are most comparable to the static LSA measures collected in other studies. We calculated standardized mean differences to estimate summary effects across studies using Hedges's g (Hedges, 1981). Because many studies reported multiple outcome measures, we estimated multilevel meta-regression models with random effects to account for dependent effect sizes (Van den Noortgate et al., 2015). All models were estimated using robust variance estimation to account for uncertainty in effect estimates (Moeyaert et al., 2017) and restricted maximum likelihood estimation to reduce the possibility of biased parameter estimates (Langan et al.,

2019). To quantify heterogeneity across studies, we used the I^2 and τ^2 statistics, which provide estimates of variability unattributable to random variation (Higgins & Thompson, 2002). All analyses were conducted using R (R Core Team, 2023) along with the metafor (Viechtbauer, 2010), meta (Schwarzer, 2007), and dmetar (Harrer et al., 2019) packages.

Effect sizes, as standardized mean differences, provide information about the degree to which LSA differentiates DLD from TL. An effect size of $g = 1$, for example, would indicate that children with DLD performed 1 *SD* lower on LSA measures than children with TL. To determine the statistical significance of individual LSA measures, we examined the 95% confidence intervals of each effect estimate. Confidence intervals that do not include zero indicate that the associated effect size was significantly greater than zero and, in the context of the present study, signify that the specific LSA measure was effective at differentiating DLD from TL.

To estimate effect sizes by outcome measure, we conducted a subgroup analysis according to the measures reported by each study. Using a classification scheme derived from Ramos et al. (2022), study outcomes were grouped into the following discrete categories: (a) morphosyntactic accuracy, (b) morphosyntactic length, (c) morphosyntactic proficiency, (d) semantics, (e) discourse productivity, and (f) narrative macrostructure. We first estimated individual intercept-only models for each outcome measure subgroup and then calculated standardized mean differences for each subgroup to provide descriptive information about effect sizes for different outcome measures. We displayed effect estimates for each outcome measure using forest plots. For instances in which studies reported multiple values for the same outcome measures, we derived study-level pooled values to facilitate ease of data visualization.

To identify characteristics of LSA that were associated with differences in effect size, we conducted a moderator analysis by estimating a meta-regression model that included the following variables: outcome measure, elicitation method, language, and mean age. Elicitation methods included (a) story tell, (b) story retell, (c) conversation, (d) play, (e) personal narrative, (f) picture description, or (g) multiple means of elicitation. In the moderator analysis, any elicitation method that was represented by a single study was placed into an aggregate "other" group to ensure an adequate number of observations for each. Language of elicitation was a categorical predictor with three levels, corresponding to whether elicitation was conducted in the first language (L1), in the second language (L2), or using a bilingual task. The mean age in months was included as a continuous predictor.

To evaluate the presence of publication bias, we examined the distribution of effect sizes across studies. A lack of normality in the distribution of effect sizes, such that studies with smaller effect sizes are underrepresented, is indicative of publication bias (Page, Sterne, et al., 2021). We evaluated possible overrepresentation of large effect sizes by examining the degree to which asymmetry was present in this distribution through examination of a funnel plot, in which effect sizes are plotted by their standard deviation. In addition, we conducted a variant of Egger's test that used a multilevel model with robust variance estimation to account for dependent effect sizes (Rodgers & Pustejovsky, 2021).

Results

See Figure 1 for an overview of search results and study selection. The initial database search yielded 737 studies, of which 520 were unique. A review of titles and abstracts resulted in 131 candidate studies. Following a full-text review of these studies, we identified 22 that met the inclusion criteria. The backward and forward searches yielded an additional 1,296 unique studies. After screening the titles of these studies, we identified 297 candidates. We then reviewed the full text of these studies and identified 13 that met the inclusion criteria. The final article pool comprised 35 studies.

Research Question 1: Examination of LSA in Bilinguals

Study Characteristics

See Table 1 for an overview of study characteristics. Sample sizes across studies ranged from 12 to 178 participants, with participant ages ranging from 2;0 (years; months) to 11;9. There was a diverse range of languages represented across studies, including Albanian, Amharic, Arabic, Assyrian, Bengali, Bulgarian, Catalan, Cantonese, Chinese, Danish, Dari, English, Farsi, Frisian, German, Gujarati, Hebrew, Hindi, Hungarian, Kirundi, Mandarin, Navajo, Nepali, Pashto, Polish, Portuguese, Punjabi, Romanian, Russian, Somali, Sinhala, Spanish, Suryoyo, Tamil, Tarifit-Berber, Twi, Turkish, Ukrainian, Urdu, Uyghur, and Vietnamese. The most common language pair was Spanish–English, which was reported in 14 studies. The countries in which studies were conducted also varied substantially and included Canada, Greece, Israel, Italy, Malaysia, the Netherlands, Spain, and the United States.

Classification of Measures

Studies reported a wide variety of outcome measures comprising several language domains including morphosyntax, semantics, and discourse (see the Appendix).

Table 1. Study characteristics.

Study	Sample size[a]	Age range	Languages	Elicitation method	Outcome measure
Altman et al. (2016)	31	5;4–6;6	L1: English L2: Hebrew	Story retell	Internal state terms Mean length of communication unit Mean length of three longest communication units Number of different words Story grammar Total number of communication units Total number of words Word formation Word choice
Blom et al. (2022)	20	5–6	L1: Turkish L2: Dutch	Conversation Story tell	Mean length of utterance in words Omission errors Substitution errors
Boerma & Blom (2017)	66	4;10–7;2	L1: Multiple L2: Dutch	Story tell	MAIN production score
Bonifacci et al. (2020)	55	6–7[b]	L1: Multiple L2: Italian	Story tell	BVL macrostructure Mean length of utterance Number of different words Words per minute
De Anda et al. (2023)	52	2;0 and 2;6	L1: Spanish L2: English	Play	Mean length of utterance in words Number of different words Sentence diversity
Fichman et al. (2017)	49	5;7–6;8	L1: Russian L2: Hebrew	Story retell	Story grammar Total number of clauses Total number of words
Fichman et al. (2020)	18	5;8–6;3	L1: Russian L2: Hebrew	Story tell	Communication units Mean length of utterance in words Total number of words
Fichman et al. (2021)	48	5;7–6;8	L1: Russian L2: Hebrew	Story retell	Enabling relations Internal state terms Story grammar Mean length of clause in words Number of different words Total clauses Total number of words
Fiestas (2008)	98	6;3–9;2	L1: Spanish L2: English	Story tell	Mean length of utterance in words Mean number of communication units Number of different words Story score total
Govindarajan & Paradis (2019)	87	5–6[b]	L1: Multiple L2: English	Story tell	ENNI story grammar score Mean length of communication unit Number of different words Referring expressions
Guiberson (2016)	62	2;0–2;11	L1: Spanish L2: English	Play	Mean length of utterance in words Number of different words
Guiberson (2020)	184	3;0–5;10	L1: Spanish L2: English	Story tell	Length of longest utterance Mean length of utterance in words Mean length of three longest utterances Number of different words Total number of words
Guiberson et al. (2015)	82	3;1–5;9	L1: Spanish L2: English	Story retell	Mean length of utterance in words Ungrammaticality index
Henderson et al. (2018)	90	4;0–5;11	L1: Navajo L2: English	Story retell	PEARL language complexity PEARL story grammar
Iluz-Cohen & Walters (2012)	37	4;11–7;5	L1: English L2: Hebrew	Story retell Story tell	Story grammar Total number of words

(*table continues*)

Table 1. *(Continued).*

Study	Sample size[a]	Age range	Languages	Elicitation method	Outcome measure
Jacobson & Schwartz (2002)	20	4;1–5;4	L1: Spanish L2: English	Play	Mean length of utterance in words
Jacobson & Walden (2013)	48	5–10[b]	L1: Spanish L2: English	Story retell	D Number of different words Omissions
Kapantzoglou et al. (2017)	40	4–5[b]	L1: Spanish L2: English	Story retell Story tell	D Grammatical errors per communication unit Mean length of utterance in words Subordination index
Kapantzoglou et al. (2021)	62	5–7	L1: Spanish L2: English	Story retell	SELPS proficiency score Total number of words
Kupersmitt & Armon-Lotem (2019)	105	5–7	L1: English/ Russian L2: Hebrew	Story tell	Ratio of causal relations
Lazewnik et al. (2019)	30	4;1–5;10	L1: Spanish L2: English	Story retell	Mean length of utterance in words
Marini et al. (2019)	22	7;0–10;6	L1: Italian L2: German	Story tell	Complete sentences Global coherence errors Lexical informativeness Local coherence errors Paragrammatic errors
McCabe & Bliss (2005)	31	8–11	L1: Spanish L2: English	Personal narrative	Actions Codas Evaluations Orientations
Ooi & Wong (2012)	61	3;8–6;9	L1: Chinese L2: English	Play Conversation	D Index of productive syntax Mean length of utterance in words
Paradis et al. (2013)	178	4;10–8;7	L1: Multiple L2: English	Story tell	ENNI story grammar score
Paradis et al. (2022)	63	5–7	L1: Multiple L2: English	Conversation Story tell	Clausal density Clausal density without sentential complement clause Complex sentences Mean length of utterance in words Simple sentences
Restrepo (1998)	62	5–7	L1: Spanish L2: English	Conversation Picture description Story retell	Mean length of terminal unit Number of errors per terminal unit
Rezzonico et al. (2015)	20	4[b]	L1: Multiple L2: English	Story retell	Information score[c] Number of different words Sentence length score
Sanz-Torrent et al. (2008)	12[d]	3;5–4;3 & 4;7–5;3	L1: Catalan L2: Spanish	Conversation	Number of utterances
Shivabasappa et al. (2018)	30	5–6 & 6–7[b]	L1: Spanish L2: English	Story retell Story tell	Core vocabulary score Occurrence score
Simon-Cereijido & Gutiérrez-Clellen (2007)	48	4[b]	L1: Spanish L2: English	Story tell	Mean length of utterance in words Ungrammaticality index Theme arguments
Smyk (2012)	73	5;3–8;0	L1: Spanish L2: English	Story retell	Mean length of utterance Number of different words Number of errors per terminal unit Percent maze words
Squires et al. (2014)	166	5–6 and 6–7[b]	L1: Spanish L2: English	Story retell	MISL macrostructure MISL microstructure

(*table continues*)

Table 1. *(Continued).*

Study	Sample size[a]	Age range	Languages	Elicitation method	Outcome measure
Tsimpli et al. (2016)	30	5;5–11;9	L1: Multiple L2: Greek	Story retell	Internal state terms Number of different words Subordination index Story structure complexity
Verhoeven et al. (2011)	24	7–9[b]	L1: Multiple L2: Dutch	Story tell	Mean length of utterance in words Story length Ungrammatical utterances

Note. L1 = first language; L2 = second language; MAIN = Multilingual Assessment Instrument for Narratives (Gagarina et al., 2019); BVL = Batteria Valutazione Linguaggio 4–12 (Marini et al., 2015); PEARL = Predictive Early Assessment of Reading and Language (Peterson & Spencer, 2014); D = lexical diversity measure; SELPS = Spanish–English Language Proficiency Scale (Smyk et al., 2013); ENNI = Edmonton Narrative Norms Instrument (Schneider et al., 2005); MISL = Monitoring Indicators of Scholarly Language (Gillam et al., 2016).

[a]Sample sizes include bilingual participants only. [b]Age range estimated from reported mean and standard deviation. [c]From Renfrew Bus Story Test (Cowley & Glasgow, 1994). [d]Excludes MLUw control group (MLUw = mean length of utterance in words).

Across all studies, there were 58 unique outcome measures. Each measure was categorized by its linguistic domain, following the aforementioned classification scheme described by Ramos et al. (2022). The following sections describe how measures were grouped into each of these categories.

Morphosyntax. Measures of morphosyntax represented approximately half of all metrics, comprising 32 unique measures across 25 studies. Measures of "morphosyntactic accuracy" constituted nine unique measures across as many studies. Accuracy measures were those that examined the overall grammaticality of utterances (e.g., percentage of ungrammatical utterances, number of errors per communication unit) or the presence of specific error patterns (e.g., omission and substitution errors). Measures of "morphosyntactic length" were those that examined the length of utterances, consisting of 10 unique measures across 19 studies. The most common length metric was mean length of utterance in words (MLUw), which was reported in 15 studies. Additional length measures included mean length of clause, mean length of communication unit, mean length of longest communication unit, mean length of terminal utterance, mean length of utterance in morphemes, and mean length of five longest utterances in words. In addition, Guiberson (2020) reported two novel metrics: mean length of three longest utterances and length of longest utterance. Lastly, eight studies examined "morphosyntactic proficiency," consisting of 12 unique measures. Proficiency measures were those that quantified the presence of specific morphosyntactic elements (e.g., simple and complex sentences) or overall complexity (e.g., subordination index, clausal density). Three studies reported proficiency measures from standardized tests, including the Monitoring Indicators of Scholarly Language (MISL; Gillam et al., 2012), Predictive Early Assessment of Reading and Language (PEARL; Peterson & Spencer, 2014), and Spanish–English Language Proficiency Scale (SELPS; Smyk et al., 2013).

Semantics. Metrics used to examine semantic ability comprised five unique measures across 15 studies. Semantic measures primarily consisted of lexical diversity metrics, the most common of which was number of different words, which was reported in 12 studies. Three studies reported D, a lexical diversity measure from the software program CLAN (Jacobson & Walden, 2013; Kapantzoglou et al., 2017; Ooi & Wong, 2012). One study examined the production of a set of predefined, high-frequency words with two metrics: core vocabulary and occurrence scores (Shivabasappa et al., 2018).

Discourse. At the discourse level, there were two main categories of measures across studies: discourse productivity and narrative macrostructure. Several productivity metrics quantified the number of specific linguistic elements across the entire language sample, comprising four unique measures across nine studies. These measures included the number of C-units, clauses, words, or utterances. Two studies examined fluency by measuring words per minute (Bonifacci et al., 2020) and percent maze words (Smyk, 2012). In addition, 16 studies examined narrative macrostructure, consisting of 19 unique measures. Macrostructure metrics included a range of measures, such as enabling relations, global coherence, and internal state terms. In addition, several studies examined specific aspects of story grammar such as goals, attempts, and outcomes. Seven reported scores derived from standardized tests, including the Batteria Valutazione Linguaggio 4–12 (Marini et al., 2015), the Edmonton Narrative Norms Instrument (ENNI; Schneider et al., 2005), the MISL (Gillam et al., 2012), the Multilingual Assessment Instrument for Narratives (MAIN; Gagarina et al., 2019), the PEARL (Peterson & Spencer, 2014), and the Renfrew Bus Story Test information score (Cowley & Glasgow, 1994).

Composite. In addition to measures used in isolation, a single study reported outcomes for composite metrics. Henderson et al. (2018) examined the ability of a measure that included both narrative macrostructure and

microstructure to distinguish between children with DLD and TL.

Elicitation Methods

Elicitation methods varied across studies and were classified as one of the following types: (a) story tell, (b) story retell, (c) conversation, (d) play, (e) personal narrative, and (f) picture description. Each of these tasks was used as the sole means of language sample elicitation in at least one study, except for picture description. Seven studies reported using multiple methods of elicitation. Narrative tasks, including "story tell" and "story retell," were the most common type of task across studies. Story tell tasks, which were used in 16 studies, required children to independently generate a story, often with a wordless picture book. Story retell tasks, which were used in 16 studies, required children to produce oral narratives following a presentation of the story by the examiner. Like with story tell tasks, wordless picture books were a common choice for stimuli in story retell tasks. "Conversation" was the method of elicitation in five studies (Blom et al., 2022; Ooi & Wong, 2012; Paradis et al., 2022; Restrepo, 1998; Sanz-Torrent et al., 2008). In conversational tasks, children were engaged in open-ended, semi-structured conversation on topics of relevance and interest. "Play" was the method of elicitation in four studies, three of which focused on children between 2 and 3 years (De Anda et al., 2023; Guiberson, 2016; Ooi & Wong, 2012). One study used elicitation of "personal narratives" in which children were prompted to produce factual descriptions of past events through a conversational map procedure (McCabe & Bliss, 2005). Lastly, one study (Restrepo, 1998) used "picture description" as its means of elicitation, alongside other tasks.

There was notable variability with respect to the language of elicitation. Twelve studies elicited samples in each of the participants' languages, six studies focused specifically on L1, and 12 studies focused on L2. Several studies that used L2-only tasks included samples with a diverse range of languages (e.g., Boerma & Blom, 2017; Bonifacci et al., 2020; Govindarajan & Paradis, 2019), limiting the ability to feasibly administer tasks in the L1. Six studies used bilingual tasks where participants were not restricted to a single language. A single study included an explicit code-switching condition, in which participants were encouraged to freely alternate between their languages (Iluz-Cohen & Walters, 2012).

Research Question 2: Language Sampling Methods That Differentiate DLD and TL in Bilinguals

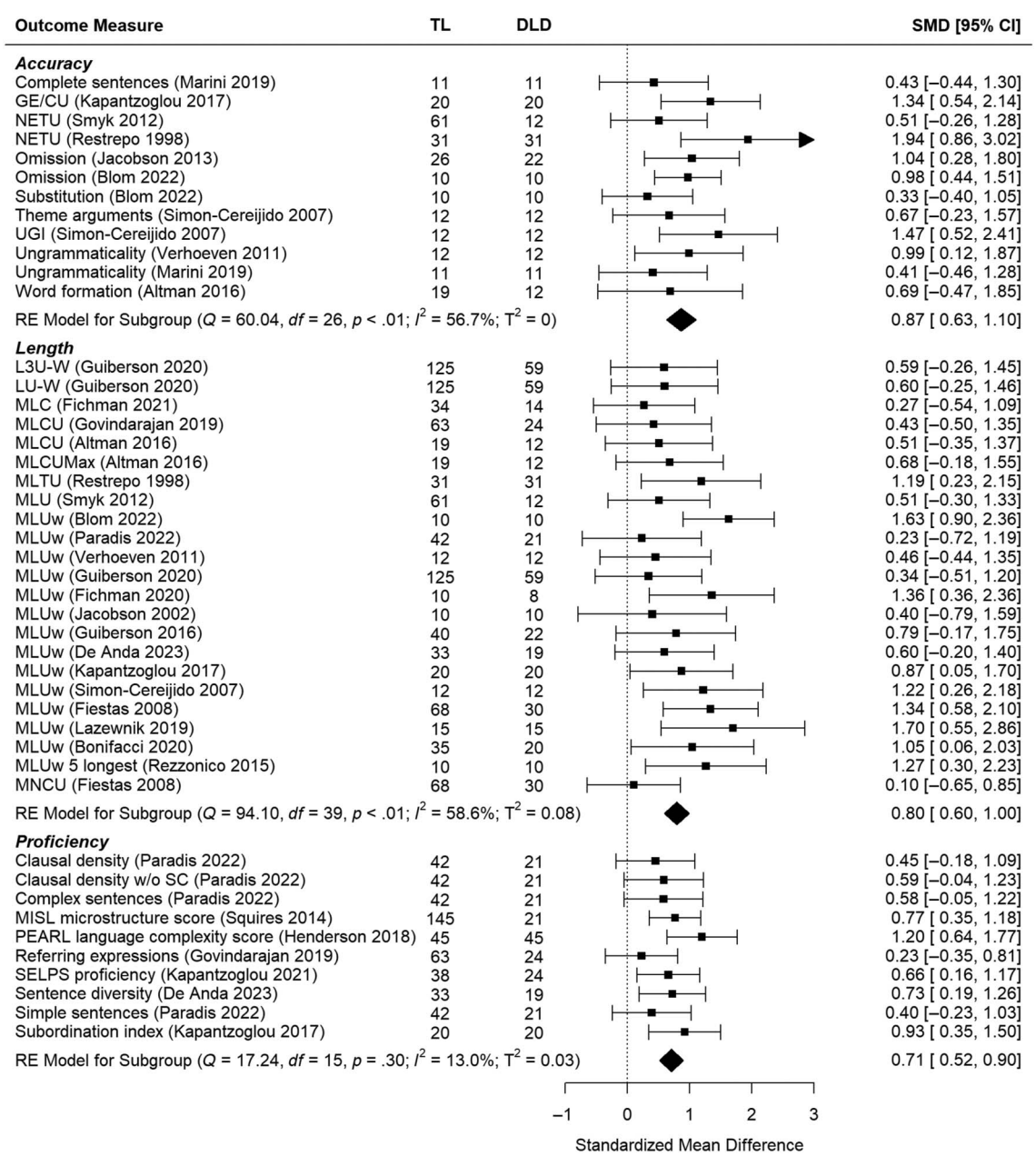
When LSA is used in practice, it is expected that children with DLD will demonstrate measurably different levels of performance than their peers with TL. Across studies, the pooled estimate representing the ability of LSA measures to distinguish children with DLD from those with TL was significant (g = 0.78, 95%CI [0.66, 0.89]). On average, children with DLD performed 0.78 *SD* lower on LSA measures than those with TL. There was a moderate amount of heterogeneity (I^2 = 50.98%, τ^2= 0.09), suggesting that differences in reported effect sizes across studies were not completely attributable to random variation (Higgins & Thompson, 2002). This is not an unexpected outcome, given the diversity of measures included. We identified two effect sizes from as many studies (Shivabasappa et al., 2018; Tsimpli et al., 2016) as outliers and removed them from the analysis following an examination of influence diagnostics (Viechtbauer & Cheung, 2010). We also excluded the composite measure reported by Henderson et al. (2018), because the data set already included the constituent measures from which they were composed. Lastly, we omitted results from Guiberson et al. (2015), two measures from Kapantzoglou et al. (2017), and two measures from Ooi and Wong (2012) because of the absence of metrics needed to calculate standardized mean differences (i.e., means and standard deviations). The final analysis included 204 individual effect sizes across studies, representing 58 unique measures. To determine the degree to which different methods of LSA were associated with differences in task performance, we examined two variables that are commonly controlled by examiners: (a) outcome measure and (b) elicitation method.

Outcome Measures

Figures 2 and 3 provide a summary of the subgroup analysis, showing the effects sizes for each measure along with the pooled effect estimates for each category. The pooled effect sizes for all outcome measure categories were significantly greater than zero, providing evidence of their ability to distinguish between the language abilities of children with DLD from those with TL. Across studies, effect sizes ranged from g = 0.10 to g = 1.95. Although many measures exhibited effect sizes significantly greater than zero, others exhibited no detectable effects.

The category with the largest pooled effect size was morphosyntactic accuracy (g = 0.87), with individual effect sizes ranging from g = 0.33 to g = 1.94. There was a moderate amount of between-studies heterogeneity (I^2 = 56.7%, τ^2 = 0), indicating that the pooled estimate may not be representative of all accuracy measures. Among accuracy metrics, several exhibited effect sizes that were significantly greater than zero: grammatical errors per communication unit, number of errors per terminal unit, omission errors, and ungrammaticality. The effect sizes for grammatical errors per communication unit (Kapantzoglou et al., 2017) and number of errors per

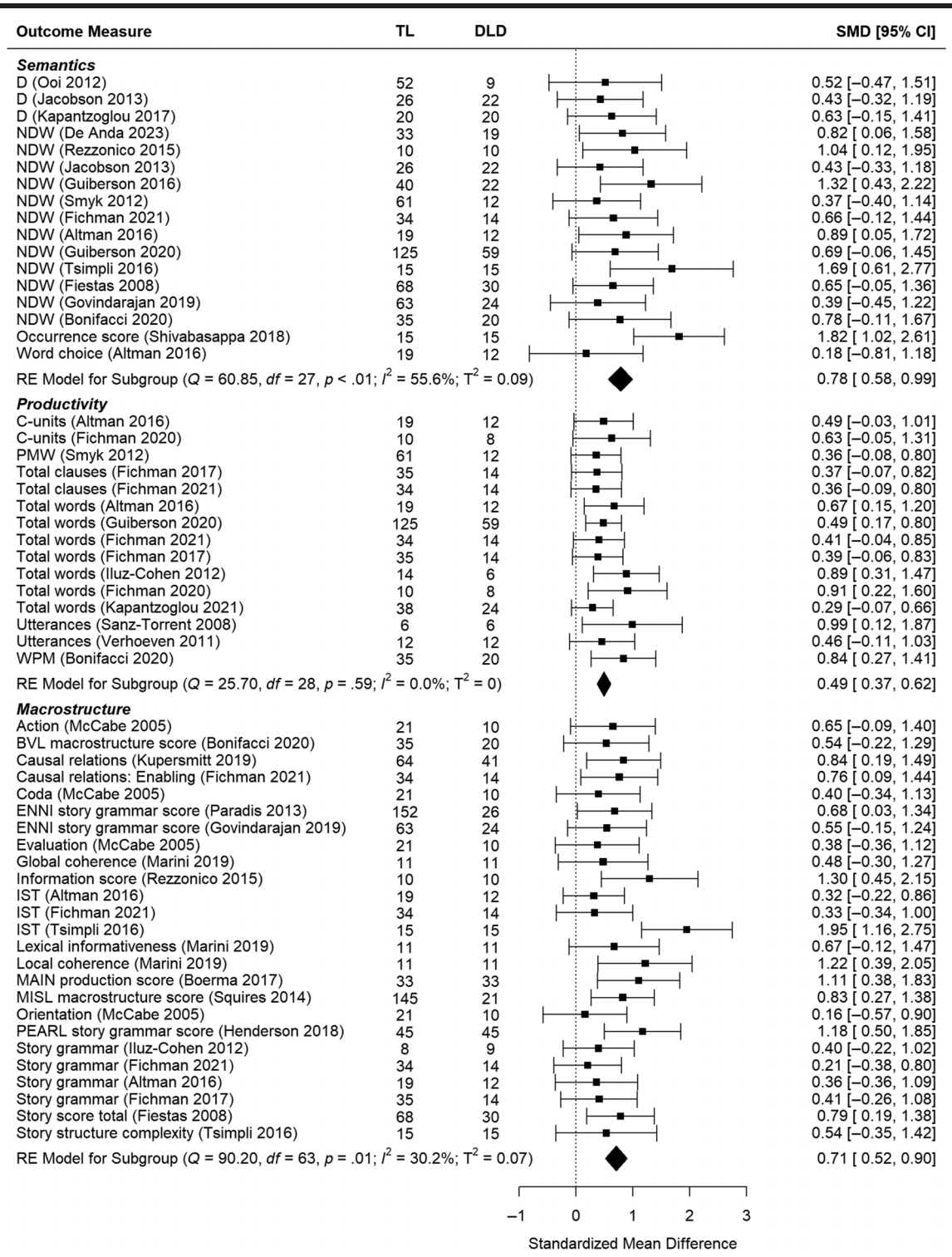
Figure 2. Effect sizes by type of outcome measure: morphosyntax. CI = confidence interval; DLD = developmental language disorder; GE/CU = grammatical errors per C-unit; NETU = number of errors per T-unit; UGI = ungrammaticality index; L3U-W = mean length of three longest utterances; LU-W = length of longest utterance; MLC = mean length of clause; MLCU = mean length of C-unit; MLCUMax = mean length of three longest C-units; MLTU = mean length of T-unit; MLU = mean length of utterance; MLUw = mean length of utterance in words; MNCU = mean number of clauses per utterance; MISL = Monitoring Indicators of Scholarly Language (Gillam et al., 2016); PEARL = Predictive Early Assessment of Reading and Language (Peterson & Spencer, 2014); SC = sentential complement clause; SELPS = Spanish–English Language Proficiency Scale (Smyk et al., 2013); SMD = standardized mean difference; TL = typical language.



terminal unit (Restrepo, 1998) were $g = 1.34$ and $g = 1.94$, respectively. The effect sizes for omission errors (Blom & Boerma, 2016; Jacobson & Walden, 2013) ranged from $g = 0.98$ to $g = 1.04$. Lastly, two measures of ungrammatical utterances (Simon-Cereijido & Gutiérrez-Clellen, 2007; Verhoeven et al., 2011) ranged from $g = 0.99$ to $g = 1.47$.

Across studies, the pooled effect size for measures of morphosyntactic length was $g = 0.80$, with individual effect sizes ranging from $g = 0.10$ to $g = 1.70$ and a moderate amount of heterogeneity ($I^2 = 58.6\%$, $\tau^2 = 0.08$). The effect sizes for two measures were significantly greater than zero: MLUw and mean length of terminal unit.

Figure 3. Effect sizes by type of outcome measure: semantics, discourse productivity, and narrative macrostructure. CI = confidence interval; DLD = developmental language disorder; D = lexical diversity measure; NDW = number of different words; WPM = words per minute; BVL = Batteria Valutazione Linguaggio 4–12 (Marini et al., 2015); ENNI = Edmonton Narrative Norms Instrument (Schneider et al., 2005); IST = internal state terms; MAIN = Multilingual Assessment Instrument for Narratives (Gagarina et al., 2019); MISL = Monitoring Indicators of Scholarly Language (Gillam et al., 2016); PEARL = Predictive Early Assessment of Reading and Language (Peterson & Spencer, 2014); SMD = standardized mean difference; TL = typical language.



MLUw was the most frequently reported measure, with effect sizes ranging from $g = 0.23$ to $g = 1.70$. In addition, several studies reported variants of MLUw, including length of longest utterance (Guiberson, 2020), mean length of three longest utterances (Guiberson, 2020), and mean length of five longest utterances (Rezzonico et al., 2015). Of these variants, only mean length of five longest utterances exhibited an effect size that was significantly greater than zero ($g = 1.27$). The effect size for mean length of terminal unit (Restrepo, 1998) was $g = 1.19$.

The pooled effect size for measures of morphosyntactic proficiency was $g = 0.71$, with individual effect sizes ranging from $g = 0.23$ to $g = 1.20$. The low amount of heterogeneity ($I^2 = 13\%$, $\tau^2 = 0.03$) suggests that the pooled effect size provides a reliable estimate of the true effect size magnitude. Five measures exhibited effect sizes that were significantly greater than zero: MISL microstructure, PEARL language complexity, SELPS proficiency, subordination index, clausal density, and sentence diversity. Of these measures, three were derived from standardized tests. Squires et al. (2014) used the MISL microstructure score to evaluate the complexity of elements in children's narratives ($g = 0.77$), while Henderson et al. (2018) used the PEARL language complexity score ($g = 1.20$) and Kapantzoglou et al. (2017) used the SELPS proficiency score ($g = 0.66$). Subordination index, a measure of the ratio of clauses to terminal units (Kapantzoglou et al., 2017), exhibited an effect size of $g = 0.93$. Paradis et al. (2022) similarly examined two measures of clausal density but reported outcomes corresponding to smaller effect sizes, ranging from $g = 0.45$ to $g = 0.59$. Lastly, the effect size for sentence diversity (De Anda et al., 2023), which was used to quantify unique subject–verb combinations, was $g = 0.73$.

The pooled effect size estimate for measures of semantics was $g = 0.78$, with individual effect sizes ranging from $g = 0.18$ to $g = 1.82$ and a moderate amount of heterogeneity ($I^2 = 55.6\%$, $\tau^2 = 0.09$). The effect sizes for two measures were significantly greater than zero: number of different words and occurrence score. Of these measures, number of different words was the most frequently reported metric, with effect sizes ranging from $g = 0.37$ to $g = 1.69$ across 12 studies. The effect size for occurrence score (Shivabasappa et al., 2018) was $g = 1.82$.

Discourse productivity measures exhibited the lowest pooled effect size ($g = 0.49$), ranging from $g = 0.29$ to $g = 0.99$, but also the lowest amount of heterogeneity ($I^2 = 0\%$, $\tau^2 = 0$). The effect sizes for three productivity measures were significantly greater than zero: total number of words, number of utterances, and words per minute. The effect sizes for total number of words, which ranged from $g = 0.29$ to $g = 0.91$, were significantly greater than zero in four studies (Altman et al., 2016; Fichman et al., 2020, 2021; Guiberson, 2020) and nonsignificant in one study (Kapantzoglou et al., 2021). The effect sizes for number of utterances ranged from $g = 0.46$ to $g = 0.99$. Sanz-Torrent et al. (2008) reported a significant effect size, but Verhoeven (2011) did not. Lastly, the effect size for words per minute (Bonifacci et al., 2020) was $g = 0.84$.

Narrative macrostructure made up the single largest group of measures and exhibited a significant effect size ($g = 0.71$), with individual effect sizes ranging from $g = 0.16$ to $g = 1.95$. Although this was the most diverse category of measures, it exhibited lower heterogeneity than some other categories ($I^2 = 30.2\%$, $\tau^2 = 0.07$). Within this group, several measures derived from standardized tests exhibited effect sizes significantly greater than zero, including the Renfrew Bus Story Test information score ($g = 1.30$) in Rezzonico et al. (2015), ENNI story grammar score ($g = 0.68$) in Paradis et al. (2013), MAIN production score ($g = 1.11$) in Boerma and Blom (2017), MISL macrostructure score ($g = 0.83$) in Squires et al. (2014), and PEARL story grammar score ($g = 1.18$) in Henderson et al. (2018). In addition, several other measures demonstrated effect sizes significantly greater than zero: causal relations, internal state terms, local coherence (Marini et al., 2019), and internal response (Iluz-Cohen & Walters, 2012). Causal relations (Fichman et al., 2021; Kupersmitt & Armon-Lotem, 2019), which refer to how individuals describe connections between events in a narrative, focusing on specific types of relations (i.e., enabling, physical, motivational, and psychological), had effect sizes ranging from $g = 0.76$ to $g = 0.84$. Internal state terms (also referred to as "mental state terms") focus on a specific set of lexical items that describe the psychological or emotional state of characters in a narrative. Of the three studies that included these measures, only Tsimpli et al. (2016) reported a significant effect size ($g = 1.95$). Local coherence is a measure of the relatedness of utterances within a discourse ($g = 1.22$). Lastly, Fiestas (2008) reported a significant effect size ($g = 0.79$) for a composite narrative score, which included components, ideas and language, and episode structure.

Two studies examined differences in code-switching for children with DLD and TL. Kapantzoglou et al. (2021) found that children with DLD and TL code-switched at similar rates, suggesting the limited utility of code-switching as a clinical maker of DLD. Conversely, Iluz-Cohen and Walters (2012), in an examination of descriptive data, reported higher proportions of code-switches for children with DLD. The authors also reported differences in length for story retell tasks in which bilingual children were encouraged to code-switch, compared to tasks in which a single language was used.

Moderator Analysis

To determine the degree to which variation in overall effect size could be explained by differences in LSA procedures or participant characteristics, we estimated a meta-regression model with the following moderators: outcome measure, elicitation method, language of task, and mean age of participants. The moderator representing outcome measures comprised six groups, using the previously described means of classification (i.e., morphosyntactic accuracy, morphosyntactic length, morphosyntactic proficiency, semantics, discourse productivity, and narrative macrostructure) with narrative macrostructure as the reference group. Elicitation methods included story retell, story tell, multiple, or other and used story retell as the reference group. Language of task included L1, L2, or bilingual task. A test of moderator heterogeneity (see Table 2) was nonsignificant, indicating that variation in effect sizes across studies could not be fully accounted for by the moderators $Qm(11) = 7.78$, $p = .73$. The covariates for type of outcome measure, elicitation method, language of task, and mean age were all nonsignificant, indicating that differences in these variables were not associated with meaningful variation in effect size. Results of the meta-regression suggest that the ability of LSA to distinguish between DLD and TL is not associated with differences in any of the included moderator variables.

Quality of Evidence

We adapted Dollaghan (2007) to assess the methodological quality of each study. With respect to the samples used for each study, we examined three characteristics: sample size, gate design, and representativeness. Twenty-four (69%) used samples of 30 participants or greater. The size and diversity of samples is an important consideration for generalizability, and studies with larger samples are more precise. The second key characteristic was the distinction between one- and two-gate designs. A single study utilized a one-gate design (Squires et al., 2014) and was the only study to include a sample that was representative of the prevalence of DLD in the general population. One-gate designs use the same criteria for entry for all participants, regardless of clinical profile, comprising a broad, representative sample, which presumably includes both children with and without DLD (Dollaghan & Horner, 2011). Two-gate designs, in which children with DLD are preselected, are susceptible to spectrum bias, as they may not include the full range of ability levels present in the population. Nonetheless, the disadvantage of one-gate designs is that they may be unfeasible to implement in many cases, given the large number of participants required to adequately sample a sufficient number of children with DLD.

Table 2. Analysis of moderators of language sample analysis effect size.

Coefficient	β	95% CI	*p*
Outcome measure[a]			
Accuracy	0.12	[–0.12, 0.35]	.34
Length	0.07	[–0.1, 0.25]	.42
Productivity	–0.04	[–0.24, 0.15]	.68
Proficiency	0.01	[–0.21, 0.24]	.92
Semantics	0.07	[–0.13, 0.26]	.5
Elicitation method[b]			
Story retell	0.08	[–0.16, 0.32]	.49
Multiple	0.3	[–0.08, 0.68]	.12
Other	–0.05	[–0.45, 0.35]	.8
Language[c]			
L1 + L2	–0.07	[–0.4, 0.26]	.66
L2	–0.04	[–0.18, 0.09]	.52
Age	0	[–0.01, 0]	.27

Note. CI = confidence interval; L1 = first language; L2 = second language.

[a]Reference group = narrative macrostructure. [b]Reference group = story tell. [c]Reference group = L1.

The reference measure used to determine the clinical status of participants is a key consideration, because of the potential effect on study outcomes. A study that uses a flawed reference measure may incorrectly classify participants as having DLD, leading to inaccurate conclusions about the validity of the index measure being evaluated. We considered several dimensions of reference measure testing, including validity and reliability, uniformity in administration across groups, and independent testing. With respect to validity and reliability, because there is no universally agreed-upon gold standard for DLD identification in bilingual children, we considered two characteristics when evaluating reference measures that are considered to be best practice: multiple converging sources of information and assessment in both languages (Bedore & Peña, 2008; Castilla-Earls et al., 2020; Ebert, 2020). Twenty-seven of the included studies (77%) used a reference measure that adhered to these criteria. Regarding reference measure administration, 26 studies (74%) gave the same measure to participants with DLD and TL. Many studies that used different reference measures for each group relied on the presence of a preexisting diagnosis as a means of DLD classification. A lack of uniformity in measuring language ability may result in unidentified cases of DLD in the TL group, or potential misdiagnoses in the DLD group as seen in previous studies (Hamann & Abed Ibrahim, 2017; Tuller et al., 2018). Lastly, we examined whether studies used independent testing for their reference measure. Independent testing is preferred because it helps to ensure the objectivity of clinical determinations. Only 10 studies (29%)

reported using independent testing in the identification of DLD.

Two aspects of the index measure were considered in the evaluation study quality: blinded testing (which we refer to as masked/masking) and reliability. Keeping examiners masked to the clinical status of participants is preferred, because it ensures uniform administration of index measure. Masking was considered to be present if studies reported its use during language sample elicitation or transcription. Only seven studies (20%) reported using masking in their index measure. The absence of masking is consistent with the results of previous systematic reviews (Orellana et al., 2019; Ortiz, 2021; Ramos et al., 2022), highlighting an important consideration for future studies. Lastly, with respect to reliability, 33 studies (94%) reported measuring reliability in their index measure, which was commonly measured through the use of multiple individuals transcribing and scoring a language sample.

Publication Bias

We evaluated the presence of possible publication bias by examining the distribution of effect sizes across studies. Figure 4 shows a funnel plot of the standardized mean differences from each study on the *x*-axis by their standard error on the *y*-axis. Visual inspection of the funnel plot reveals some asymmetry; studies with larger effect sizes appear overrepresented, as demonstrated by a clustering of studies outside the funnel on the right side of the plot. Additionally, the test for asymmetry was significant ($\beta = 2.69$, SE = 0.76, $p = .004$), indicating the absence of normality in the distribution of effect sizes across studies. Studies with larger effect sizes were overrepresented in the sample, while those with smaller effects were more limited in number.

Discussion

The goal of this systematic review was to examine the literature on the use of LSA to assess language performance of bilingual children with and without DLD. The pooled standardized mean difference for LSA was significant, providing evidence of its effectiveness in differentiating between DLD and TL. However, the presence of possible publication bias indicates the need for caution in drawing definitive conclusions about the overall magnitude of the effect. Across studies, there was remarkable variability in the type of outcome measures used to analyze language samples, encompassing several linguistic domains. An examination of effect sizes for individual measures revealed considerable variation in the ability of different outcome measures to distinguish DLD from TL, and only some measures exhibited effect sizes that were significantly greater than zero. Despite this variability, differences in the type of outcome measure used were not significantly associated with effect size, as demonstrated in the moderator analysis. Other moderators, including elicitation method, language of task, and mean age, were similarly nonsignificant, indicating that variation in effect size is related to other factors. The following section provides a qualitative overview of the results of the present systematic review in the context of the extant LSA literature.

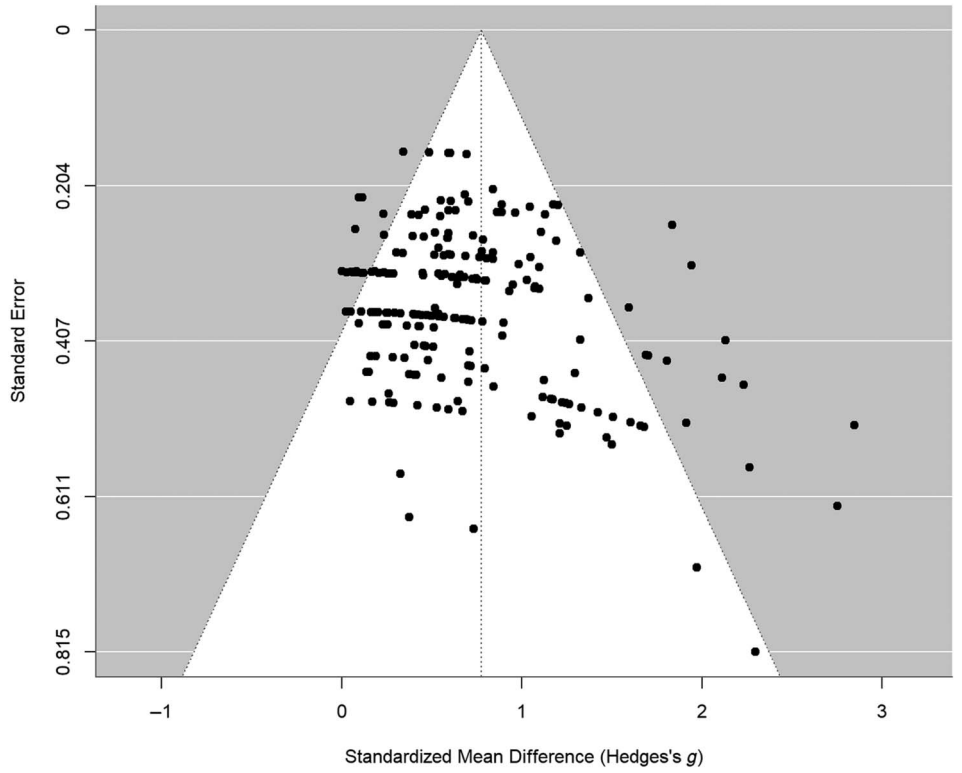
Outcome Measures

One of the main ways in which studies varied was with respect to the outcome measures used to quantify LSA results. Included measures characterized a variety of domains of language ability including morphosyntax, semantics, and discourse. Much of the variation in outcome measures was attributed to how language samples were elicited across studies. In many cases, the elicitation task dictated the type of outcome measure that could be used. Narrative macrostructure measures, for example, would only sensibly be used to quantify results of a language sample elicited with a narrative task. Other measures, such as MLUw, may be less sensitive to the elicitation method and therefore more broadly applicable.

The ability of outcome measures to distinguish between DLD and TL varied substantially, as exemplified by the range of effect sizes. Although no single measure demonstrated superior evidence of efficacy, several exhibited a stronger ability to differentiate DLD from TL than others, as exemplified by effect sizes significantly greater than zero. Measures of morphosyntactic accuracy exhibited the largest overall effect size. Within this category, grammatical errors per communication unit, number of errors per T-unit, omission errors, and ungrammaticality indices exhibited the best performance. Measures of morphosyntactic length with the largest effect sizes included mean length of terminal unit, mean length of utterance in morphemes, and MLUw. In the domain of semantics, the largest effect sizes were observed for number of different words and occurrence score. For discourse productivity, total number of words, number of utterances, and words per minute exhibited the largest effect sizes. Lastly, in the domain of narrative macrostructure, the largest effect sizes were observed for measures of causal relations, internal state terms, lexical informativeness, local coherence, and several measures derived from standardized assessments.

Among the range of metrics included across studies, two stood out for their high frequency of use: MLUw and number of different words. Although the effects for these measures were not significant in all studies in which they

Figure 4. Funnel plot of publication bias.



were reported, their frequent inclusion allows for a more precise estimate of their ability to differentiate between DLD and TL in bilinguals. In contrast with these high-frequency metrics, most other measures were included in only one or two studies. This disparity in frequency of use makes it difficult to draw direct comparisons between measures, as greater precision of effect estimates can be obtained from those measures that are represented in multiple studies. Variants of MLUw reported by Guiberson (2020), for example, may be useful in measuring language ability, but their inclusion in a single study indicates the need for further investigation.

The single most commonly reported measure across studies was MLUw. MLUw is commonly used to make cross-linguistic comparisons because it is argued to be less sensitive to variation in inflectional morphology, compared to other measures (Gutiérrez-Clellen et al., 2000). Despite its popularity, MLUw is not without limitations. Although it may accomplish the goal of making cross-linguistic comparisons for some language pairs, it may be less effective with others. MLUw, for example, may not be an ideal metric for polysynthetic languages, in which morphologically dense words are permissible, potentially reducing the number of words needed in any given utterance (Rozendaal & Baker, 2008). Like other metrics, MLUw was also not universally effective at differentiating DLD from TL in the present systematic review, as several studies reported data indicating a null effect. These considerations highlight the fact that several factors may affect the precision of any given metric. Although the moderator analysis did not find significant associations among the included moderators, other unobserved variables are likely to play a role in moderating the measurement ability of LSA.

Elicitation Methods

Results from the moderator analysis did not reveal evidence of an association between elicitation method and effect size. Although several different elicitation methods were represented across studies, no single method was significantly more effective in differentiating DLD from TL in bilingual children. In spite of this, it is important to acknowledge the limited number of studies that utilized elicitation methods other than story tell/retell. Approximately 90% of studies used narrative tasks as their sole means of elicitation or as part of their elicitation methods. Although other elicitation methods were present, they comprised a small minority of tasks. A greater variety of elicitation methods is required to draw accurate comparisons. A single study (Kapantzoglou et al., 2017) reported

outcomes for both story tell and retell tasks, but no other studies directly compared elicitation methods within the same sample. Considering the growing use of alternative methods of elicitation, such as home language sampling collected using video conference software (Manning et al., 2020), a greater focus on the relationship between elicitation method and task performance is warranted.

In addition to the elicitation task, language of elicitation is a critical consideration. Language of elicitation was not a significant moderator, suggesting that it did not impact the ability of LSA to measure differences in language ability for children with DLD and TL across studies. This is reflected in several studies that directly compared LSA outcomes for L1 and L2, reporting similar levels of performance across languages (Fichman et al., 2017; Marini et al., 2019; McCabe & Bliss, 2005). Despite this result, cross-linguistic LSA can provide clinically relevant information not captured by effect size metrics. Language difficulties are likely to manifest differently in each language, as exemplified by Altman et al. (2016), who reported that children with DLD produced distinct error patterns in Hebrew compared to English. Some outcome measures may also be sensitive to the language of the task, as demonstrated in Shivabasappa et al. (2018), where authors reported differences in core vocabulary usage in L1 compared to L2. While these types of differences may not always be reflected in quantitative outcomes, they are certainly relevant for forming a clinical impression and treatment planning.

Another relevant consideration for language of elicitation is the use of code-switching. Most studies elicited language samples in each language that participants spoke. This is a common method of elicitation and is widely considered to be best practice (Bedore & Peña, 2008; Castilla-Earls et al., 2020; Ebert, 2020). A single study included a code-switching condition by modeling a narrative in which the examiner attempted to elicit code-switches by alternating between languages (Iluz-Cohen & Walters, 2012). Although this is an uncommon approach, it does highlight the potential utility of language samples elicited with the explicit goal of encouraging children to use both languages, particularly in light of previous studies that have identified differences in measured language ability when code-switching is included in the analysis (Hiebert & Rojas, 2021; Kekejian, 2022). In LSA, code-switched utterances are generally excluded due to the complexity they introduce in analysis (Ebert, 2020). Although code-switching may not be a reliable indicator of DLD (Gutiérrez-Clellen et al., 2009; Kapantzoglou et al., 2021), its consideration in language sampling may be useful. Results from previous studies highlight the variability in the frequency of code-switching in language samples restricted to one language (Gutiérrez-Clellen et al., 2009; Halpin & Melzi, 2021; Kapantzoglou et al., 2021; Raichlin et al., 2019), but providing a linguistic context in which children have the opportunity to use their full linguistic repertoire may yield unique and clinically useful information (Gross et al., 2022).

Limitations

Although we made efforts to comprehensively identify all relevant studies, it is possible that some studies were missed. Regarding the range of languages represented, the focus on bilinguals broadly, and not on a single language background, may limit the ability to generalize conclusions to a specific language. In addition to linguistic diversity, there was also a wide variety of outcome measures. Although it is useful to examine results for a range of different LSA methods, this renders interpreting outcomes for low-frequency measures challenging. For measures included in only one study, it is difficult to draw conclusions about their true effects, in contrast with those measures whose use was reported more frequently (e.g., MLUw). In terms of the extant literature available for synthesis, there was likely a bias toward studies and corresponding research groups that were able to publish research in English.

With respect to quantitative analyses, conclusions derived from pooled effect estimates may not be universally representative of every measure's ability to differentiate between DLD and TL. Some measures may be more effective than others, a factor that may be obscured by summary metrics. The analysis of moderators should similarly be interpreted with an understanding of its limitations. The lack of significance in the moderators does not indicate that age or language of task is not important, but rather that they did not contribute to differences in effect size among the included studies. Age, for example, may influence the type of task chosen, as young children may benefit from elicitation methods that include play-based activities. Lastly, while this study focused on LSA, it did not examine diagnostic accuracy. Results from this study provide information about the ability of LSA to differentiate between DLD and TL, but not its precision in identifying children with DLD.

Future Research

There are several areas in which future studies on LSA in bilinguals can build upon extant research. The first is the age in which LSA has been investigated. Across studies, children from ages 2;0 to 11;9 were represented. Although adolescents were included in some studies, teenagers were not. The use of LSA in older children would improve our understanding of which approaches are

appropriate for a wider range of ages. Ramos et al. (2022) identified similar limitations among LSA measures for English-speaking children, as some measures may be less sensitive to language difficulties at older ages. Regarding elicitation methods, narrative tasks made up the vast majority across studies. Further investigation into other means of elicitation would provide more insight into potential differences in outcome measures for certain types of tasks. The use of code-switching in LSA is another consideration for future studies, given how infrequently it has been considered in previous LSA research. As a linguistic phenomenon common to many bilinguals, a much better understanding of the role that code-switching plays in LSA is needed.

Lastly, with respect to study design, future studies should strive to include reference standards that adequately account for cross-linguistic ability. Although this may not always be possible, particularly in cases where children from diverse language backgrounds are included, the validity of the reference standard is a key consideration. The lack of a universally accepted reference standard for the identification of DLD in bilinguals presents a major barrier to those studies that wish to establish the presence of a communication disorder. It is precisely for this reason that the field needs a robust set of methods to ensure an accurate classification for all participants. With respect to measure administration, future studies should report the use of masking and independent testing. Studies with larger samples that are representative of the prevalence of DLD in the general population are also warranted.

Clinical Implications

Results provide insight into the degree to which different LSA measures and elicitation methods can differentiate between children with DLD and those with TL. Although examiners should be conscious of the specific measures that will best capture language ability for the specific languages being measured, the applicability of LSA to bilinguals broadly is evident from the diverse language backgrounds represented across studies. Clinicians can use these results to guide their decisions about which measures may be best in specific contexts. Combinations of measures are likely to provide a greater amount of detail than any single measure. Using measures of narrative macrostructure with morphosyntactic measures can provide rich information for the purposes of forming a clinical impression or for treatment planning. Elicitation in each language a client speaks is best practice, and clinicians can use LSA to make cross-linguistic comparisons, which are a valuable part of assessment. When selecting elicitation tasks, clinicians should know of the benefits of using structured tasks, such as story tell and retell, but additional elicitation methods, such as play and conversation, are also beneficial and may be a more appropriate choice for some clients.

Data Availability Statement

The data used in this study are available from the corresponding author on reasonable request.

References

Altman, C., Armon-Lotem, S., Fichman, S., & Walters, J. (2016). Macrostructure, microstructure, and mental state terms in the narratives of English–Hebrew bilingual preschool children with and without specific language impairment. *Applied Psycholinguistics, 37*(1), 165–193. https://doi.org/10.1017/S0142716415000466

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

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

Blom, E., & Boerma, T. (2016). Why do children with language impairment have difficulties with narrative macrostructure? *Research in Developmental Disabilities, 55,* 301–311. https://doi.org/10.1016/j.ridd.2016.05.001

Blom, E., Boerma, T., Karaca, F., de Jong, J., & Küntay, A. C. (2022). Grammatical development in both languages of bilingual Turkish–Dutch children with and without developmental language disorder. *Frontiers in Communication, 7,* Article 1059427. https://doi.org/10.3389/fcomm.2022.1059427

Boerma, T., & Blom, E. (2017). Assessment of bilingual children: What if testing both languages is not possible? *Journal of Communication Disorders, 66,* 65–76. https://doi.org/10.1016/j.jcomdis.2017.04.001

Bonifacci, P., Atti, E., Casamenti, M., Piani, B., Porrelli, M., & Mari, R. (2020). Which measures better discriminate language minority bilingual children with and without developmental language disorder? A study testing a combined protocol of first and second language assessment. *Journal of Speech, Language, and Hearing Research, 63*(6), 1898–1915. https://doi.org/10.1044/2020_JSLHR-19-00100

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

Cowley, J., & Glasgow, C. (1994). *Renfrew Bus Story Test*. AGS.

De Anda, S., Cycyk, L. M., Durán, L., Biancarosa, G., & McIntyre, L. L. (2023). Sentence diversity in Spanish–English bilingual toddlers. *American Journal of Speech-Language Pathology, 32*(2), 576–591. https://doi.org/10.1044/2022_AJSLP-22-00149

Deeks, J., Bossuyt, P., Leeflang, M., & Takwoingi, Y. (2023). *Cochrane handbook for systematic reviews of diagnostic test accuracy*. Wiley.https://doi.org/10.1002/9781119756194

De Lamo White, C., & Jin, L. (2011). Evaluation of speech and language assessment approaches with bilingual children. *International Journal of Language & Communication Disorders, 46*(6), 613–627. https://doi.org/10.1111/j.1460-6984.2011.00049.x

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

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

Ebert, K. D. (2020). Language sample analysis with bilingual children: Translating research to practice. *Topics in Language Disorders, 40*(2), 182–201. https://doi.org/10.1097/TLD.0000000000000209

Ebert, K. D., & Pham, G. (2017). Synthesizing information from language samples and standardized tests in school-age bilingual assessment. *Language, Speech, and Hearing Services in Schools, 48*(1), 42–55. https://doi.org/10.1044/2016_LSHSS-16-0007

Fichman, S., Altman, C., Voloskovich, A., Armon-Lotem, S., & Walters, J. (2017). Story grammar elements and causal relations in the narratives of Russian–Hebrew bilingual children with SLI and typical language development. *Journal of Communication Disorders, 69,* 72–93. https://doi.org/10.1016/j.jcomdis.2017.08.001

Fichman, S., Armon-Lotem, S., Walters, J., & Altman, C. (2021). Story grammar elements and mental state terms in the expression of enabling relations in narratives of bilingual preschool children. *Discourse Processes, 58*(10), 925–942. https://doi.org/10.1080/0163853x.2021.1972391

Fichman, S., Walters, J., Melamed, R., & Altman, C. (2020). Reference to characters in narratives of Russian–Hebrew bilingual and Russian and Hebrew monolingual children with developmental language disorder and typical language development. *First Language, 42*(2), 263–291. https://doi.org/10.1177/0142723720962938

Fiestas, C. E. (2008). *The dynamic assessment of narratives: A bilingual study* [Doctoral dissertation, The University of Texas at Austin]. ProQuest Dissertations and Theses Global. https://www.proquest.com/dissertations-theses/dynamic-assessment-narratives-bilingual-study/docview/304474223/se-2?accountid=14696

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

Gagarina, N. V., Klop, D., Kunnari, S., Tantele, K., Välimaa, T., Balčiūnienė, I., Bohnacker, U., & Walters, J. (2019). MAIN: Multilingual Assessment Instrument for Narratives. *ZAS Papers in Linguistics, 56,* 155–155. https://doi.org/10.21248/zaspil.56.2019.414

Gillam, S. L., Gillam, R. B., Fargo, J. D., Olszewski, A., & Segura, H. (2016). Monitoring Indicators of Scholarly Language: A progress-monitoring instrument for measuring narrative discourse skills. *Communication Disorders Quarterly, 38*(2), 96–106. https://doi.org/10.1177/1525740116651442

Gillam, S. L., Gillam, R. B., & Reece, K. (2012). Language outcomes of contextualized and decontextualized language intervention: Results of an early efficacy study. *Language, Speech, and Hearing Services in Schools, 43*(3), 276–291. https://doi.org/10.1044/0161-1461(2011/11-0022)

Govindarajan, K., & Paradis, J. (2019). Narrative abilities of bilingual children with and without developmental language disorder (SLI): Differentiation and the role of age and input factors. *Journal of Communication Disorders, 77,* 1–16. https://doi.org/10.1016/j.jcomdis.2018.10.001

Gross, M. C., López González, A. C., Girardin, M. G., & Almeida, A. M. (2022). Code-switching by Spanish–English bilingual children in a code-switching conversation sample: Roles of language proficiency, interlocutor behavior, and parent-reported code-switching experience. *Language, 7*(4), Article 246. https://doi.org/10.3390/languages7040246

Guiberson, M. (2016). Telehealth measures screening for developmental language disorders in Spanish-speaking toddlers. *Telemedicine and e-Health, 22*(9), 739–745. https://doi.org/10.1089/tmj.2015.0247

Guiberson, M. (2020). Alternatives to traditional language sample measures with emergent bilingual preschoolers. *Topics in Language Disorders, 40*(2), E1–E6. https://doi.org/10.1097/TLD.0000000000000208

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

Gutiérrez-Clellen, V. F., Restrepo, M. A., Bedore, L., Peña, E. D., & Anderson, R. (2000). Language sample analysis in Spanish-speaking children: Methodological considerations. *Language, Speech, and Hearing Services in Schools, 31*(1), 88–98. https://doi.org/10.1044/0161-1461.3101.88

Gutiérrez-Clellen, V. F., Simon-Cereijido, G., & Erickson Leone, A. (2009). Code-switching in bilingual children with specific language impairment. *International Journal of Bilingualism, 13*(1), 91–109. https://doi.org/10.1177/1367006909103530

Halpin, E., & Melzi, G. (2021). Code-switching in the narratives of dual-language Latino preschoolers. *International Journal of Bilingual Education and Bilingualism, 24*(9), 1271–1287. https://doi.org/10.1080/13670050.2018.1553928

Hamann, C., & Abed Ibrahim, L. (2017). Methods for identifying specific language impairment in bilingual populations in Germany. *Frontiers in Communication, 2,* Article 16. https://doi.org/10.3389/fcomm.2017.00016

Harrer, M., Cuijpers, P., Furukawa, T., & Ebert, D. D. (2019). *dmetar: Companion R package for the guide "Doing Meta-Analysis in R."* R Foundation for Statistical Computing.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107–128. https://doi.org/10.3102/10769986006002107

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

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

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,* Article 106110. https://doi.org/10.1016/j.jcomdis.2021.106110

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

Iluz-Cohen, P., & Walters, J. (2012). Telling stories in two languages: Narratives of bilingual preschool children with typical and impaired language. *Bilingualism: Language and Cognition, 15*(1), 58–74. https://doi.org/10.1017/S1366728911000538

Jacobson, P. F., & Schwartz, R. G. (2002). Morphology in incipient bilingual Spanish-speaking preschool children with specific

language impairment. *Applied Psycholinguistics, 23*(1), 23–41. https://doi.org/10.1017/S0142716402000024

Jacobson, P. F., & Walden, P. R. (2013). Lexical diversity and omission errors as predictors of language ability in the narratives of sequential Spanish–English bilinguals: A cross-language comparison. *American Journal of Speech-Language Pathology, 22*(3), 554–565. https://doi.org/10.1044/1058-0360(2013/11-0055)

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

Kekejian, C. R. (2022). *Translanguaging in the context of speech-language pathology: An exploratory study of linguistic ability using narrative retells with Armenian–English bilingual children* [Doctoral dissertation, The University of Utah]. https://www.proquest.com/docview/2741073763/abstract/D4B5EC94F5584B63PQ/1

Kupersmitt, J. R., & Armon-Lotem, S. (2019). The linguistic expression of causal relations in picture-based narratives: A comparative study of bilingual and monolingual children with TLD and DLD. *First Language, 39*(3), 319–343. https://doi.org/10.1177/0142723719831927

Langan, D., Higgins, J. P. T., Jackson, D., Bowden, J., Veroniki, A. A., Kontopantelis, E., Viechtbauer, W., & Simmonds, M. (2019). A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. *Research Synthesis Methods, 10*(1), 83–98. https://doi.org/10.1002/jrsm.1316

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

MacWhinney, B. (2000). The CHILDES Project: Tools for analyzing talk: Volume I. Transcription format and programs, Volume II. The database. *Computational Linguistics, 26*(4), 657. https://doi.org/10.1162/coli.2000.26.4.657

MacWhinney, B., & Wagner, J. (2010). Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository. *Gesprachsforschung: Online-Zeitschrift Zur Verbalen Interaktion, 11,* 154–173.

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

Marini, A., Marotta, L., Bulgheroni, S., & Fabbro, F. (2015). *Batteria per la Valutazione del Linguaggio in Bambini dai 4 ai 12 anni* [Battery for the assessment of language in children aged 4 to 12]. Giunti O.S. Organizzazioni Speciali.

Marini, A., Sperindè, P., Ruta, I., Savegnago, C., & Avanzini, F. (2019). Linguistic skills in bilingual children with developmental language disorders: A pilot study. *Frontiers in Psychology, 10,* Article 493. https://doi.org/10.3389/fpsyg.2019.00493

McCabe, A., & Bliss, L. S. (2005). Narratives from Spanish-speaking children with impaired and typical language development. *Imagination, Cognition and Personality, 24*(4), 331–346. https://doi.org/10.2190/cjq8-8c9g-05lg-0c2m

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

Miller, J. F., & Iglesias, A. (2012). *Systematic Analysis of Language Transcripts* (Research Version) [Computer software]. SALT Software.

Moeyaert, M., Ugille, M., Beretvas, S. N., Ferron, J., Bunuan, R., & Van den Noortgate, W. (2017). Methods for dealing with multiple outcomes in meta-analysis: A comparison between averaging effect sizes, robust variance estimation and multilevel meta-analysis. *International Journal of Social Research Methodology, 20*(6), 559–572. https://doi.org/10.1080/13645579.2016.1252189

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

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

Ortiz, J. A. (2021). Using nonword repetition to identify language impairment in bilingual children: A meta-analysis of diagnostic accuracy. *American Journal of Speech-Language Pathology, 30*(5), 2275–2295. https://doi.org/10.1044/2021_AJSLP-20-00237

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *The BMJ, 372,* Article n71. https://doi.org/10.1136/bmj.n71

Page, M. J., Sterne, J. A. C., Higgins, J. P. T., & Egger, M. (2021). Investigating and dealing with publication bias and other reporting biases in meta-analyses of health research: A review. *Research Synthesis Methods, 12*(2), 248–259. https://doi.org/10.1002/jrsm.1468

Paradis, J., Duncan, T. S., Thomlinson, S., & Rusk, B. (2022). Does the use of complex sentences differentiate between bilinguals with and without DLD? Evidence from conversation and narrative tasks. *Frontiers in Education, 6,* Article 804088. https://doi.org/10.3389/feduc.2021.804088

Paradis, J., Emmerzael, K., & Duncan, T. S. (2010). Assessment of English language learners: Using parent report on first language development. *Journal of Communication Disorders, 43*(6), 474–497. https://doi.org/10.1016/j.jcomdis.2010.01.002

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

Pavelko, S. L., & Owens, R. E. (2023). A sweet tutorial to the SUGAR method of language sampling. *Perspectives of the ASHA Special Interest Groups, 8*(1), 32–49. https://doi.org/10.1044/2022_PERSP-22-00134

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

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

Peterson, D. B., & Spencer, T. D. (2014). *Predictive Early Assessment of Reading and Learning*. Language Dynamics Group.

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

Raichlin, R., Walters, J., & Altman, C. (2019). Some wheres and whys in bilingual codeswitching: Directionality, motivation and locus of codeswitching in Russian–Hebrew bilingual children. *International Journal of Bilingualism, 23*(2), 629–650. https://doi.org/10.1177/1367006918763135

Ramos, M. N., Collins, P., & Peña, E. D. (2022). Sharpening our tools: A systematic review to identify diagnostically accurate language sample measures. *Journal of Speech, Language, and Hearing Research, 65*(10), 3890–3907. https://doi.org/10.1044/2022_JSLHR-22-00121

R Core Team. (2023). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

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

Rezzonico, S., Chen, X., Cleave, P. L., Greenberg, J., Hipfner-Boucher, K., Johnson, C. J., Milburn, T., Pelletier, J., Weitzman, E., & Girolametto, L. (2015). Oral narratives in monolingual and bilingual preschoolers with SLI. *International Journal of Language & Communication Disorders, 50*(6), 830–841. https://doi.org/10.1111/1460-6984.12179

Rodgers, M. A., & Pustejovsky, J. E. (2021). Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychological Methods, 26*(2), 141–160. https://doi.org/10.1037/met0000300

Rojas, R., & Iglesias, A. (2009). Making a case for language sampling: Assessment and intervention with (Spanish–English) second language learners. *The ASHA Leader, 14*(3), 10–13. https://doi.org/10.1044/leader.FTR1.14032009.10

Rozendaal, M. I., & Baker, A. E. (2008). A cross-linguistic investigation of the acquisition of the pragmatics of indefinite and definite reference in two-year-olds. *Journal of Child Language, 35*(4), 773–807. https://doi.org/10.1017/S0305000908008702

Sanz-Torrent, M., Serrat, E., Andreu, L., & Serra, M. (2008). Verb morphology in Catalan and Spanish in children with Specific Language Impairment: A developmental study. *Clinical Linguistics & Phonetics, 22*(6), 459–474. https://doi.org/10.1080/02699200801892959

Schneider, P., Dubé, R. V., & Hayward, D. (2005). *The Edmonton Narrative Norms Instrument*. University of Alberta Faculty of Rehabilitation Medicine. https://doi.org/10.1037/t75173-000

Schwarzer, G. (2007). meta: An R package for meta-analysis. *R News, 7*(3), 40–45.

Schwob, S., Eddé, L., Jacquin, L., Leboulanger, M., Picard, M., Oliveira, P. R., & Skoruppa, K. (2021). Using nonword repetition to identify developmental language disorder in monolingual and bilingual children: A systematic review and meta-analysis. *Journal of Speech, Language, and Hearing Research, 64*(9), 3578–3593. https://doi.org/10.1044/2021_jslhr-20-00552

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

Simon-Cereijido, G., & Gutiérrez-Clellen, V. F. (2007). Spontaneous language markers of Spanish language impairment. *Applied Psycholinguistics, 28*(2), 317–339. https://doi.org/10.1017/S0142716407070166

Smyk, E. (2012). *Second language proficiency in sequential bilingual children with and without primary language impairment* [Doctoral dissertation, Arizona State University]. ProQuest Dissertations & Theses. https://www.proquest.com/docview/1038158549/abstract/C87C15C5F1C94E42PQ/1

Smyk, E., Restrepo, M. A., Gorin, J. S., & Gray, S. (2013). Development and validation of the Spanish–English Language Proficiency Scale (SELPS). *Language, Speech, and Hearing Services in Schools, 44*(3), 252–265. https://doi.org/10.1044/0161-1461(2013/12-0074)

Squires, K. E., Lugo-Neris, M. J., Peña, E. D., Bedore, L. M., Bohman, T. M., & Gillam, R. B. (2014). Story retelling by bilingual children with language impairments and typically developing controls. *International Journal of Language & Communication Disorders, 49*(1), 60–74. https://doi.org/10.1111/1460-6984.12044

Tsimpli, I. M., Peristeri, E., & Andreou, M. (2016). Narrative production in monolingual and bilingual children with specific language impairment. *Applied Psycholinguistics, 37*(1), 195–216. https://doi.org/10.1017/S0142716415000478

Tuller, L., Hamann, C., Chilla, S., Ferré, S., Morin, E., Prevost, P., dos Santos, C., Abed Ibrahim, L., & Zebib, R. (2018). Identifying language impairment in bilingual children in France and in Germany. *International Journal of Language & Communication Disorders, 53*(4), 888–904. https://doi.org/10.1111/1460-6984.12397

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2015). Meta-analysis of multiple outcomes: A multilevel approach. *Behavior Research Methods, 47*(4), 1274–1294. https://doi.org/10.3758/s13428-014-0527-2

Verhoeven, L., Steenge, J., van Weerdenburg, M., & van Balkom, H. (2011). Assessment of second language proficiency in bilingual children with specific language impairment: A clinical perspective. *Research in Developmental Disabilities: A Multidisciplinary Journal, 32*(5), 1798–1807. https://doi.org/10.1016/j.ridd.2011.03.010

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods, 1*(2), 112–125. https://doi.org/10.1002/jrsm.11

Appendix (p. 1 of 2)

Description of Outcome Measures Across Studies

Category	Studies	Outcome measure description
Morphosyntax: Accuracy		
Complete sentences	1	Number of complete sentences in sample
Grammatical errors per C-unit	1	Number of grammatical errors divided by number of C-units
Number of errors per T-unit	2	Number of grammatical errors divided by number of T-units
Omission	2	Number of morpheme omissions in obligatory contexts
Substitution	1	Number of morpheme substitution errors
Theme arguments	1	Percentage of correct uses of theme arguments in obligatory contexts
Ungrammaticality[a]	4	Percentage of utterances with grammatical errors
Word formation	1	Percentage of morphological errors
Morphosyntax: Length		
Mean length of three longest utterances	1	Number of words in three longest utterances divided by three
Length of longest utterance	1	Number of words in longest utterance
Mean length of clause	1	Number of words per clause divided by the number of clauses
Mean length of C-unit	2	Number of words per C-unit divided by the number of C-units[b]
Mean length of three longest C-units	1	Number of words in three longest C-units divided by three[b]
Mean length of T-unit	1	Number of words per utterance divided by the total number of T-units
Mean length of utterance in morphemes	1	Number of morphemes per utterance divided by number of utterances
Mean length of utterance in words	15	Number of words per utterance divided by number of utterances
Mean length of five longest utterances (words)	1	Number of words of the five longest utterances divided by five
Mean number of C-units	1	Number of C-units per utterance divided by number of utterances
Morphosyntax: Proficiency		
Clausal density	1	Number of clauses divided by the total number of sentences
Clausal density without sentential complement	1	Number of clauses divided by the total number of sentences, excluding sentential complement clauses
Complex sentences	1	Number of sentences comprised of two or more clauses
ENNI referring expressions	1	ENNI score for how a child introduces a referent[c]
Index of productive syntax	1	Rating of noun phrase, verb phrase, question/negation, and phrase structure
MISL microstructure score	1	Score from MISL Microstructure subscale[d]
PEARL language complexity score	1	Score from the PEARL Language Complexity subscale[e]
SELPS proficiency	1	Proficiency score from the SELPS[f]
Sentence diversity	1	Number of unique subject–verb combinations
Simple sentences	1	Number of sentences comprised of one clause
Subordination	1	Number of subordinate clauses divided by number of C-units
Subordination index	1	Number of subordinate clauses divided by number of T-units
Semantics		
Core vocabulary score	1	Number of core vocabulary words produced, from a predefined set of 30 words
D/VocD	3	Measure of lexical diversity calculated using CLAN software[g]
Number of different words	12	Number of different words, generally in the first 100 words of a sample
Occurrence score	1	Number of times each core vocabulary word produced, from a predefined set of 30 words
Word choice	1	Percentage of words produced that are contextually inappropriate
Discourse: Productivity		
C-units	2	Total number of C-units
Percent maze words	1	Number of maze words, such as false starts, repetitions, and reformulations, divided by total number of words

(*table continues*)

Appendix (p. 2 of 2)

Description of Outcome Measures Across Studies

Category	Studies	Outcome measure description
Total number of clauses	2	Total number of clauses
Total number of words	7	Total number of words
Utterances	2	Total number of utterances
Words per minute	1	Number of words produced per minute
Discourse: Macrostructure		
Actions	1	Number of completed actions described by the speaker in a personal narrative
BVL macrostructure score	1	Macrostructure score from the BVL[h]
Causal relations	2	Presence of relations between story grammar elements including enabling, physical, motivational, and psychological
Codas	1	Number of summary statements that finish a personal narrative
ENNI story grammar score	2	Macrostructure score from the ENNI[c]
Evaluations	1	Number of utterances in a personal narrative in which the subjective experience of speaker is expressed
Global coherence	1	The number of production errors that repeat previous introduced topics, do not provide additional information, deviate from the flow of discourse, or include incongruent ideas divided by the total number of utterances
Information score	1	Information score from the Renfrew Bus Story Test[i]
Internal/mental state terms	3	Number of internal state terms divided by number of content words or number of clauses
Lexical informativeness	1	The number of lexical information units divided by the number of words
Local coherence	1	The number of utterances that were conceptually different than the previous one, including topic shifts and missing references, divided by the total number of utterances
MAIN production score	1	Production score from the MAIN[j]
MISL macrostructure score	1	Macrostructure score from the MISL[d]
Orientations	1	Statements that provide information about the setting in personal narrative
PEARL story grammar score	1	Story grammar score from the PEARL[e]
Story score total	1	Composite score of story production including story components, story ideas and language, and episode structure
Story grammar elements/story structure complexity	5	Number of story grammar elements present including initiating events, goals, attempts, and outcomes

[a]Measures reported as ungrammaticality index (Guiberson, 2016; Simon-Cereijido & Gutiérrez-Clellen, 2007), percentage of grammatical errors (Verhoeven et al., 2011), and paragrammatic errors (Marini et al., 2019). [b]Altman et al. (2016) included bound morphemes in Hebrew for concepts that would be expressed as function words in English. [c]Edmonton Narrative Norms Instrument (Schneider et al., 2005). [d]Monitoring Indicators of Scholarly Language (Gillam et al., 2016). [e]Predictive Early Assessment of Reading and Language (Peterson & Spencer, 2014). [f]Spanish–English Language Proficiency Scale (Smyk et al., 2013). [g]Computerized Language Analysis (MacWhinney & Wagner, 2010). [h]Batteria Valutazione Linguaggio 4–12 (Marini et al., 2015). [i]From Renfrew Bus Story Test information score (Cowley & Glasgow, 1994). [j]Multilingual Assessment Instrument for Narratives (Gagarina et al., 2019).