ASHA American Speech-Language-Hearing Association

JSLHR

Review Article

Sharpening Our Tools: A Systematic Review to Identify Diagnostically Accurate Language Sample Measures

Michelle N. Ramos,[a] Penelope Collins,[a] and Elizabeth D. Peña[a]

[a] School of Education, University of California, Irvine

ARTICLE INFO

Article History:
Received February 24, 2022
Revision received May 28, 2022
Accepted June 17, 2022

Editor-in-Chief: Stephen M. Camarata
Editor: Emily Lund

https://doi.org/10.1044/2022_JSLHR-22-00121

ABSTRACT

Purpose: This systematic review provides a comprehensive summary of the diagnostic accuracy of English language sample analysis (LSA) measures for the identification of developmental language disorder.
Method: An electronic database search was conducted to identify English publications reporting empirical data on the diagnostic accuracy of English LSA measures for children aged 3 years or older.
Results: Twenty-eight studies were reviewed. Studies included between 18 and 676 participants ranging in age from 3;0 to 13;6 (years;months). Analyzed measures targeted multiple linguistic domains, and diagnostic accuracy ranged from less than 25% to greater than 90%. Morphosyntax measures achieved the highest accuracy, especially in combination with length measures, and at least one acceptable measure was identified for each 1-year age band up to 10 years old.
Conclusion: Several LSA measures or combinations of measures are clinically useful for the identification of developmental language disorder, although more research is needed to replicate findings using rigorous methods and to explore measures that are informative for adolescents and across diverse varieties of English.
Supplemental Material: https://doi.org/10.23641/asha.21183247

Within the field of speech-language pathology, language sample analysis (LSA) is often promoted as the gold standard for assessing language (Miller et al., 2016) and the "cornerstone of any clinical assessment battery" (Evans, 1996, p. 207) to identify language impairment, more recently termed developmental language disorder (DLD; Bishop et al., 2017). Yet, LSA is not deployed as such in typical clinical practice. Speech-language pathologists (SLPs), on the whole, do not conduct LSA regularly or adhere to consistent procedures (Fulcher-Rood et al., 2018), relying instead on standardized language tests (Fulcher-Rood et al., 2019; Selin et al., 2019), despite concerns raised around their inadequate accuracy for identifying DLD (Betz et al., 2013) and cultural and linguistic bias in test design (Castilla-Earls et al., 2020; Horton-Ikard, 2010). Among the barriers to greater adoption of LSA is a lack of clarity on the diagnostic value of a language sample, specifically which measures are the most accurate indicators of impairment and how to interpret them to determine a diagnosis of DLD. Although there is a growing body of evidence addressing these questions, it is distributed across several publications and is not readily available for easy reference by clinicians. Therefore, in this review, we seek to consolidate the existing evidence of the diagnostic accuracy of LSA into a single resource to guide the selection and interpretation of these measures in clinical practice and to inform future research and policy-directed advocacy efforts.

DLD affects approximately 7%–10% of children and is characterized by difficulties in learning and using the rules of language in the absence of intellectual, developmental, or physical disabilities that would explain the disorder (Tomblin et al., 1997). Core characteristics of DLD include particular difficulty acquiring grammatical morphology, which is often observed in children's morphosyntactic

Correspondence to Michelle N. Ramos: mnramos1@uci.edu. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

productions (Leonard, 2017; Rice & Wexler, 1996). In English, typical errors associated with DLD involve verb tense marking and agreement errors as well as difficulty producing complex sentences. Although this profile of language disabilities has been referred to as DLD in recent years, other terms are used in the literature including specific language impairment, language impairment, and primary language impairment (Bishop et al., 2017).

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

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

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

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

Prior Reviews

Previous systematic reviews of the diagnostic accuracy of LSA have focused on specific populations or sets of

measures (C. A. Dollaghan & Horner, 2011; Eisenberg et al., 2001; Eisenberg & Guo, 2016) or included LSA measures among other language assessments in their analyses (C. A. Dollaghan & Horner, 2011; Pawlowska, 2014; Shahmahmood et al., 2016). The evidence for mean length of utterance (MLU) indicates that it can provide supportive evidence of a disorder but, on its own, is not adequate for diagnosing DLD in preschool children (Eisenberg et al., 2001). Measures of morphosyntactic diversity and development (i.e., Tense Marker Total: quantifies the types of verb tense morphemes produced; Developmental Sentence Scoring [DSS] Total: rates the developmental level of forms used in eight linguistic categories) were also found inadequate for identifying impairment in this age group (Shahmahmood et al., 2016). In contrast, measures of morphosyntactic accuracy have yielded acceptable to good diagnostic accuracy for children in preschool through early elementary (Eisenberg & Guo, 2016; Shahmahmood et al., 2016). These included percent grammatical utterances (PGU), the sentence point score from the DSS, and the finite verb morphology composite (FVMC). PGU expresses grammaticality as a percentage of total utterances that are correct, whereas the sentence point score is an average of points awarded per utterance for grammaticality (i.e., 1 point for a grammatical utterance, 0 for an ungrammatical utterance). The FVMC reflects the accuracy of four clinical markers in obligatory contexts: third-person singular present *–s*, regular past tense *–ed*, and copula and auxiliary BE.

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

Purpose

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

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

Method

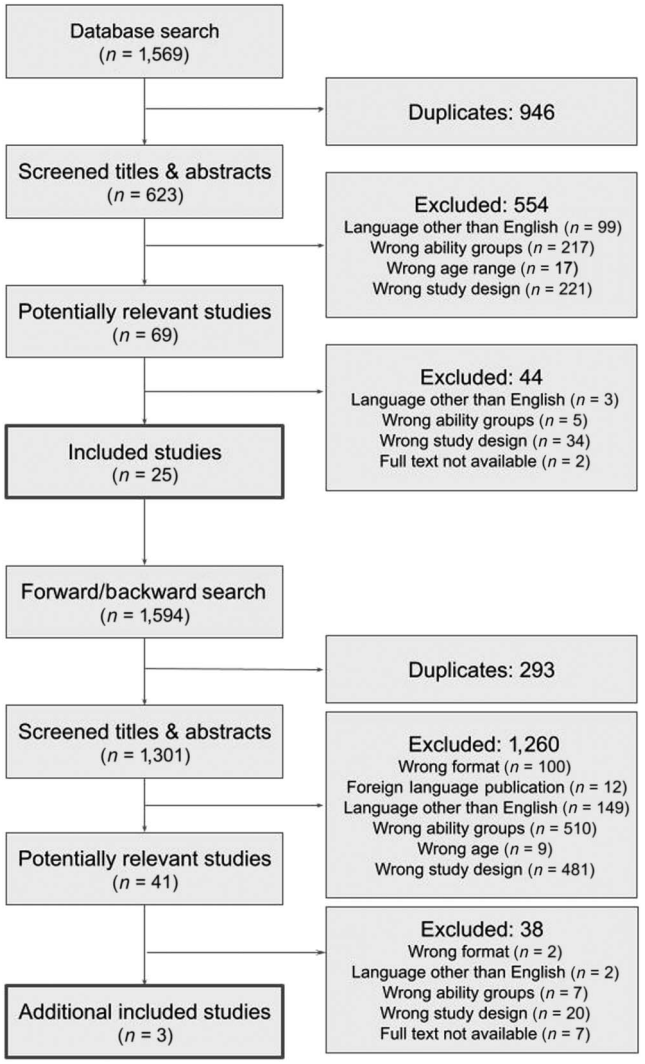
Literature Search Strategy

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

Study Selection Criteria

The search and selection process is summarized in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses chart in Figure 1. Titles and abstracts of the 623 unique results returned by the database searches were screened for relevance based on the following inclusion criteria: an empirical study published as a journal article, thesis or dissertation, conference paper, or chapter in an edited volume; language sample data were elicited in English; the participant sample included participants both with and without DLD; DLD was a primary diagnosis without comorbidities (e.g., studies with participants with

Figure 1. Search and selection process.



language impairment secondary to another diagnosis were excluded); participants were aged 3–18 years (e.g., studies that only included toddlers younger than 36 months were excluded); and the study design and analytic methods addressed diagnostic accuracy (e.g., studies that only examined the statistical significance of group differences were excluded).

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

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

Forward and backward citation chaining from the 25 remaining studies was conducted using SnowGlobe (McWeeny et al., 2021) as well as the reference and "Cited By" lists exported from Scopus for four studies that were incompatible with SnowGlobe. This process yielded 1,301 unique results that had not appeared in the electronic database search results. These studies were screened using the previously described procedures and criteria. The first author screened all studies, and the research assistant screened 20% of the studies, with 94% reliability for titles and abstracts and 100% reliability for full texts. One thousand two hundred sixty (1,260) and 38 studies were excluded during these phases, respectively.

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

Results

Twenty-eight (28) studies were ultimately included in this review (see Figure 1 for the complete selection

process) and are listed in Table 1. Language sample elicitation tasks across the corpus included play, narrative tell and retell, conversation, and expository. The size of participant samples ranged from 18 to 676 children, with an average of 159 participants. Participant age also varied significantly across studies, ranging from 2;0 to 13;6 (years;months), although 4-, 5-, and 6-year-olds were included most frequently (14, 16, and 14 studies, respectively), followed by 3- and 7-year-olds (10 studies each). Participants of 25 studies were monolingual speakers of mainstream English (ME) from the United States and Canada, one of which also included British English speakers. Three studies included speakers of African American English (AAE), two of which also included speakers of Southern White English (SWE). Two studies included bilingual speakers of English.

RQ 1: LSA Measures Examined for Diagnostic Accuracy

Because of the plethora of analyses that can be conducted from a language sample, the first RQ explored which measures have been examined for diagnostic accuracy in order to establish the scope of evidence that is available for LSA. Reviewed studies examined a wide range of language sample measures across the domains of morphology, syntax, semantics, discourse, and pragmatics. These measures are summarized in Table 2.

Table 1. Studies included for review.

Source	*N*	Age range	Elicitation task	Analyzed measure(s)
Bedore & Leonard (1998)	38	3;7–5;9 (years;months)	Play Picture description	Mean length of utterance Noun morphology composite Verb morphology composite
Castilla-Earls & Fulcher-Rood (2018)	100	4;0–6;11	Narrative retell	Grammaticality and Utterance Length Instrument
Charest et al. (2020)	377	4–9 years	Narrative tell	Moving-average type–token ratio Number of different words
Dunn et al. (1996)	242	2;6–6;11	Play	Mean length of utterance Percent structural errors
Eisenberg & Guo (2013)	34	3;0–3;11	Picture description	Percent grammatical utterances Percent sentence point Percent verb tense usage
Fletcher & Peters (1984)	29	3;4–6;11	Play Picture description Narrative retell	Unmarked verb forms Verb types
Gavin et al. (1993)	47	2;0–4;2	Conversation Play	Stage 1 major utterances Three-element noun phrases Verb phrase errors
Gladfelter & Leonard (2013)	55	4;0–5;6	Play	Finite verb morphology composite Tense and agreement productivity score Tense Marker Total
Guo & Eisenberg (2014)	36	3;0–3;11	Play	Finite verb morphology composite Tense and agreement productivity score
Guo & Schneider (2016)	129	6 & 8 years	Narrative tell	Errors per C-unit Finite verb morphology composite Percent grammatical C-units
Guo et al. (2019)	377	4–9 years	Narrative tell	Percent grammatical utterances
Guo et al. (2020)	377	4–9 years	Narrative tell	Finite verb morphology composite
Heilmann et al. (2010)	488	3;0–13;6	Conversation	10 SALT measures
Hewitt et al. (2005)	54	5;5–6;7	Conversation Narrative retell	IPSyn total Mean length of utterance Number of different words
Hoffman (2009)	48	8–10 years	Narrative tell	Proportion “restricted” utterances
Klee et al. (2017)	48	2–4 years	Play	Lexical diversity *D* Mean length of utterance
Liles et al. (1995)	114	7;6–12;6	Narrative retell	Cohesive ties Mean no. of subordinate clauses per T-unit Mean no. of words per subordinate clause Percent grammatical T-units
Moyle et al. (2011)	100	5;5–9;9	Conversation Expository	Mean length of utterance (morphemes) Noun morphology composite Verb morphology composite
Oetting & McDonald (2001)	93	4–6 years	Play	35 nonmainstream patterns
Oetting et al. (2021)	106	5 years	Play	Eight tense/agreement forms

(*table continues*)

Table 1. *(Continued).*

Source	*N*	Age range	Elicitation task	Analyzed measure(s)
Ooi & Wong (2012)	18	3;8–5;11	Play Conversation	IPSyn total Lexical diversity *D* Mean length of utterance (words)
Overton et al. (2021)	37	< 6 years	Play	DSS total IPSyn
Pavelko & Owens (2019)	306	3;0–7;11	Conversation	Clauses per sentence Mean length of utterance (SUGAR) Total words Words per sentence
Rudolph et al. (2019)	676	6;11–7;3	Play	Finite verb morphology composite
Scheffel (1997)	37	8;4–13;2	Expository (map task)	Expansions References to map Total turns Total words
Schneider et al. (2006)	377	4;0–9;11	Narrative retell	Story grammar score
Smyk (2012)	73	5;3–8;0	Narrative	Errors per T-unit Mean length of utterance Number of different words Percent maze words
Souto et al. (2014)	112	4;0–5;10	Play	DSS sentence point DSS total Finite verb morphology composite Mean tense/agreement Mean top 5 tense/agreement

Note. SALT = Systematic Analysis of Language Transcripts; IPSyn = Index of Productive Syntax; DSS = Developmental Sentence Scoring; SUGAR = Sampling Utterances and Grammatical Analysis Revised.

Morphosyntax

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

In addition to accuracy, morphosyntax was also measured in terms of *proficiency*—used here to refer to expertise with morphosyntactic production—and *length*. These measures quantified the range or diversity of forms (e.g., Tense Marker Total, tense and agreement productivity score [TAPS]), developmental sophistication (e.g., DSS, Index of Productive Syntax [IPSyn]), and complexity (e.g., clauses per sentence [CPS], DSS coordination score) of participants' morphosyntactic production. Length was most often examined at the utterance level (i.e., MLU) and generally calculated in either words or morphemes. Some unique variations included mean words per subordinate clause rather than per utterance (Liles et al., 1995), categorical rating of length (i.e., one to three words, four to seven words; Castilla-Earls & Fulcher-Rood, 2018), and the inclusion of a wider range of structures, such as derivational morphemes *–ly* and *–ful* (Pavelko & Owens, 2019).

Semantics

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

Pragmatics and Discourse

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

Table 2. Description and frequency of language sample analysis (LSA) measures analyzed in the included studies.

LSA measure	Frequency	Description
Morphosyntax: accuracy		
DSS sentence point[a]	1/28 (4%)	Total points awarded to grammatical sentences (1 point if no errors)
Errors per T-unit[b]	1/28 (4%)	Number of grammatical errors divided by total T-units
(Finite) Verb morphology composite[a,c,d,e,f,g,h,i,j]	9/28 (32%)	Percentage of correct productions in obligatory contexts of regular past tense, third-person singular present, copula BE, and auxiliary BE. Modifications also included auxiliary DO[a] or irregular past tense.[f]
Nonmainstream patterns[k]	1/28 (4%)	Total occurrences of 35 grammatical surface features that are possible in Southern African American English and/or Southern White English
Noun morphology composite[c,d]	2/28 (7%)	Percentage of correct productions in obligatory contexts of possessive *–s*, plurals, and articles
Omitted bound morphemes (SALT)[l]	1/28 (4%)	Number of obligatory morphemes that were omitted
Omitted words (SALT)[l]	1/28 (4%)	Number of obligatory words that were omitted
Percent grammatical T-units[m]/utterances[n]	2/28 (7%)	Number of grammatical utterances divided by total utterances
Percent structural errors[o]	1/28 (4%)	Percentage of utterances that contain a morphological or syntactic error (e.g., word order, omitted morpheme, omitted word, telegraphic speech)
Percent verb tense usage[e]	1/28 (4%)	Percentage of correct production in obligatory contexts of tense marking including copula/auxiliary BE, auxiliary DO, bound tense markers, and irregular past or third-person verb forms
Tense and agreement forms[p]	1/28 (4%)	Percentage of occurrence in possible contexts of mainstream overt, nonmainstream overt, and zero forms of eight targets (past tense regular, past tense irregular, verbal *–s* habitual, verbal *–s* nonhabitual, four auxiliary BE forms)
Unmarked verb forms[q]	1/28 (4%)	Number of lexical verbs produced without premodification or inflection
Verb phrase errors[r]	1/28 (4%)	Number of errors occurring within verb phrases
Morphosyntax + semantics: accuracy		
Errors per C-unit[h]	1/28 (4%)	Number of grammatical errors[s] divided by total C-units
Percent grammatical utterances[e]/C-units[h,t]	3/28 (11%)	Percentage of utterances not containing any coded errors[s]
Percent sentence point[e]	1/28 (4%)	Percentage of utterances awarded a point for containing no errors[s] (excluding C-units with a missing subject or missing main verb)
Proportion "restricted" utterances[u]	1/28 (4%)	Percentage of utterances with a complete clause (i.e., subject and predicate) and one or more syntactic or semantic errors
Utterance errors (SALT)[l]	1/28 (4%)	Number of utterances that contained a syntactic error, or three or more word-level omissions/errors, or that did not make sense
Word errors (SALT)[l]	1/28 (4%)	Number of incorrect productions of lexical items
Morphosyntax: proficiency		
Clauses per sentence[y]	1/28 (4%)	Number of clauses in the sample divided by total sentences
DSS total[a,v]	2/28 (7%)	Total points across all utterances divided by total utterances (sentence point plus 1–8 points awarded for each form produced within eight categories: main verb, indefinite pronouns/noun modifiers, personal pronouns, secondary verbs, negatives, conjunctions, interrogative reversals, and *wh*-questions)
IPSyn total[v,w,x]	3/28 (11%)	Total ratings across four categories: noun phrases, verb phrases, question and negation, and sentence structure. Each structure is rated for frequency in the sample: 0 = *never*, 1 = *once*, and 2 = *twice or more*.
Mean subordinate clauses per T-unit[m]	1/28 (4%)	Number of subordinate clauses divided by total T-units
Mean tense/agreement[a]	1/28 (4%)	Sum of DSS main verb category scores for each utterance divided by total utterances that earned at least a score of 1 for this category
Mean top 5 tense/agreement[a]	1/28 (4%)	Average of the five highest scores in the DSS main verb category
(Tense and agreement) Productivity score[f,g]	2/28 (7%)	Number of different uses (i.e., with different subjects, different lexical verbs inflected, different morphemes within the category) up to five of morphemes in five categories (copula, auxiliary BE, auxiliary DO, third-person singular, regular past), with 0–25 possible points
Tense Marker Total[f]	1/28 (4%)	Number of forms occurring at least once in samples from a set of 15 targets (cop/aux/3PS/regpast/DO), with 0–15 possible points
Morphosyntax: length		
GLi: length[n]	1/28 (4%)	Length of each utterance is rated as one of three intervals ($\leq$ 3 words, 4–7 words, or $\geq$ 8 words), and a weighted average is calculated.
Mean length of utterance (MLU)[b,c,d,l,o,w,x,y,z]	9/28 (32%)	Number of free and inflectional morphemes[c,d,l,w,z] or words[x,y] divided by total utterances/sentences
MLU: SUGAR[y]	1/28 (4%)	Number of morphemes divided by total utterances (18 derivational morphemes, and each word in a proper name = 1 morpheme; all contractions: *hafta*, *wanna*, and *gotta* = 2; *gonna* = 3)
Mean words per subordinate clause[m]	1/28 (4%)	Number of words within subordinate clauses divided by total subordinate clauses
Stage 1 major utterances[r]	1/28 (4%)	One-word utterances produced as commands, questions, or statements
Three-element noun phrases[r]	1/28 (4%)	Noun phrases with three words (i.e., determiners, modifiers, prepositions)

(*table continues*)

Table 2. *(Continued).*

LSA measure	Frequency	Description
Semantics		
Lexical diversity *D*[x,z]	2/28 (7%)	Repeated ratio of number of different words to total words calculated using CLAN software's *vocd* program
Moving-average type–token ratio[aa]	1/28 (4%)	Average of type–token ratios (ratio of different word types to total word tokens) calculated for successive 100-word cuts of the transcript
Number of different words[b,l,w,aa]	4/28 (14%)	Number of different word roots produced in the sample. Alternative calculations used the first 200 words of the sample,[aa] the first 41 utterances,[aa] and 50 utterances.[w]
Verb types[q]	1/28 (4%)	Number of different/unique verbs produced in the sample
Pragmatics/discourse		
Between-utterance pauses (SALT)[l]	1/28 (4%)	Total seconds of pausing between two utterances (no speech for ≥ 2 s)
Complete cohesive ties[m]	1/28 (4%)	Total intersentential cohesive ties (conjunctive, reference, lexical, ellipsis) that were complete (i.e., information referred to by the cohesive marker is easily found and defined without ambiguity)
Expansions[ab]	1/28 (4%)	Whether child's response to examiner's question about a nonexistent map feature expanded on features of the map/discovered
Mean turn length[l]	1/28 (4%)	Total main body words divided by total conversational turns
Percent maze words[bl]	2/28 (7%)	Percentage of words that were reduplications, revisions, filled pauses, or false starts
References[ab]	1/28 (4%)	Total number of map features mentioned by the child
Story grammar[ac]	1/28 (4%)	Total points awarded for inclusion of story grammar elements based on a story-specific rubric (character[s], setting, initiating event, etc.)
Total number of words[y,ab]	2/28 (7%)	Total words produced in the sample (including unintelligible words[a])
Total turns[ab]	1/28 (4%)	Total number of conversational turns in the sample

Note. Frequency indicates the number of studies and the percentage of the total included studies that analyzed the measure. DSS = Developmental Sentence Scoring; SALT = Systematic Analysis of Language Transcripts; IPSyn = Index of Productive Syntax; cop = copula; aux = auxiliary; 3PS = third-person singular; regpast = regular past; GLi = Grammaticality and Utterance Length Instrument; SUGAR = Sampling Utterances and Grammatical Analysis Revised.

[a]Souto et al., 2014. [b]Smyk, 2012. [c]Bedore & Leonard, 1998. [d]Moyle et al., 2011. [e]Eisenberg & Guo, 2013. [f]Gladfelter & Leonard, 2013. [g]Guo & Eisenberg, 2014. [h]Guo & Schneider, 2016. [i]Guo et al., 2020. [j]Rudolph et al., 2019. [k]Oetting & McDonald, 2001. [l]Heilmann et al., 2010. [m]Liles et al., 1995. [n]Castilla-Earls & Fulcher-Rood, 2018. [o]Dunn et al., 1996. [p]Oetting et al., 2021. [q]Fletcher & Peters, 1984. [r]Gavin et al., 1993. [s]Coded errors included missing verb, missing obligatory argument/constituent, pronoun substitution, tense marking, grammatical morphemes (articles, plural *–s*, obligatory present participle *–ing*, prepositions), and lexical/other. [t]Guo et al., 2019. [u]Hoffman, 2009. [v]Overton et al., 2021. [w]Hewitt et al., 2005. [x]Ooi & Wong, 2012. [y]Pavelko & Owens, 2019. [z]Klee et al., 2017. [aa]Charest et al., 2020. [ab]Scheffel, 1997. [ac]Schneider et al., 2006.

et al., 1995). Discourse fluency was measured using the proportion of maze words to total words (Heilmann et al., 2010; Smyk, 2012), between-utterance pause length, and words per minute (Heilmann et al., 2010).

RQ 2: Diagnostic Accuracy of LSA Measures

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

Morphosyntax Measures: Accuracy

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

and 4- to 10-year-olds' narratives (Guo et al., 2019; Guo & Schneider, 2016; Hoffman, 2009), with 83%–100% sensitivity and 82%–96% specificity overall and reaching good accuracy for 9-year-olds (90% sensitivity/specificity). Similarly, errors per C-unit yielded 91% sensitivity and 82% specificity for 6-year-olds' narrative samples as well as 94% sensitivity and 80% specificity for 8-year-olds (Guo & Schneider, 2016).

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

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

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

Morphosyntax Measures: Proficiency

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

Length Measures

Many studies have examined MLU, both independently and combined with other measures. Alone, its accuracy varies significantly. Bedore and Leonard (1998) found MLU to nearly reach good accuracy with children ages 3;7–5;9 (95% sensitivity/89% specificity), but this was not replicated with the validation sample (100% sensitivity/68% specificity). The replication findings are

consistent with other studies, which found at least one of the accuracy metrics to be inadequate (i.e., 67% sensitivity for children ages 5;5–6;7 in Hewitt et al., 2005; 72% sensitivity for children ages 5;5–9;9 in Moyle et al., 2011). Modifications to the way MLU is typically calculated, as with the Sampling Utterances and Grammatical Analysis Revised (SUGAR) protocol, resulted in better accuracy (86% sensitivity/86% specificity) with 3- to 7-year-olds' conversational samples (Pavelko & Owens, 2019).

Semantics, Pragmatics, and Discourse Measures

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

Composite Measures

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

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

Two composites were explored for bilingual speakers of English. For Malaysian Cantonese–English speakers ages 3;8–5;11, a composite of MLU in words, a Malaysian English adaptation of IPSyn total, and lexical diversity *D* fell short of acceptable levels (78% sensitivity and specificity; Ooi & Wong, 2012). For Spanish–English bilingual children ages 5;3–8;0, a composite of MLU in words, errors per T-unit, number of different words, and percent maze words yielded 83% overall diagnostic accuracy, but since the disaggregated metrics were not reported, findings should be cautiously interpreted as suggestive but not conclusive (Smyk, 2012).

Best Diagnostic Accuracy

Examining diagnostic accuracy by age, there are multiple options for monolingual speakers of ME with at least acceptable accuracy for each year interval between the ages of 3 and 10 years and at least one measure or model with good accuracy for each year interval except 6 years old (see Table 3). For 3-year-olds, studies found that MLU combined with a verb composite score (referred to as the FVMC in later studies; Bedore & Leonard, 1998) or age combined with three LARSP categories (Gavin et al., 1993)

Table 3. Language sample analysis measures with the best diagnostic accuracy by age.

Measure	Elicitation task	Materials	Sensitivity	Specificity	Overall	Cutoff
Mainstream English speakers						
3-year-olds						
Age + Stage 1 utterances + VP errors + 3-element NP[a]	Conversation/play	Toys	91%	92%	—	Yes
FVMC + MLU[b]	Play/picture description	Toys, picture sequences	95%	95%	—	No
4-year-olds						
FVMC modified (4;0–4;6 [years;months])[c]	Play	Toys	100%	100%	—	Yes
FVMC[d]	Play	Toys	93%	94%	—	Yes
FVMC[e]	Narrative tell	Picture sequences (ENNI)	92%	94%	94%	Yes
DSS sentence point[d]	Play	Toys	93%	94%	—	Yes
5-year-olds						
FVMC modified (5;0–5;6)[c]	Play	Toys	92%	93%	—	Yes
FVMC[d]	Play	Toys	91%	93%	—	Yes
FVMC[e]	Narrative tell	Picture sequences (ENNI)	100%	90%	92%	Yes
DSS sentence point[d]	Play	Toys	100%	100%	—	Yes
6-year-olds						
MLU (SUGAR) + clauses/sentence[f]	Conversation	Personal topics (SUGAR protocol)	97%	82%	—	Yes
Errors per C-unit[g]	Narrative tell	Picture sequences (ENNI)	91%	82%	85%	Yes
Unmarked verb forms + verb types[h]	Conversation/narrative	Toys, board game, picture sequence, wordless picture book	89%	90%	—	Yes
FVMC[e,g]	Narrative tell	Picture sequences (ENNI)	82%	90%	89%	Yes
Percent grammatical C-units[g,i]	Narrative tell	Picture sequences (ENNI)	82%	90%	89%	Yes
10 SALT measures[j]	Conversation	Personal topics (SALT protocol)	80%	85%	—	No
Grammaticality and Utterance Length Instrument[k]	Narrative retell	Wordless picture book	83%	92%	—	No
7-year-olds						
Cohesive ties + grammaticality + subordinate clauses/T-unit + words/subordinate clause[l]	Narrative retell	Movie	—	—	98%	No
Percent grammatical utterances[i]	Narrative tell	Picture sequences (ENNI)	92%	88%	89%	Yes
FVMC[e]	Narrative tell	Picture sequences (ENNI)	85%	86%	86%	Yes
MLU (SUGAR) + clauses per sentence[f]	Conversation	Personal topics (SUGAR protocol)	97%	82%	—	Yes
8-year-olds						
Cohesive ties + grammaticality + subordinate clauses/T-unit + words/subordinate clause[l]	Narrative retell	Movie	—	—	98%	No
Errors per C-unit[g]	Narrative tell	Picture sequences (ENNI)	94%	80%	84%	Yes
Percent "restricted" utterances[m]	Narrative tell	Wordless picture book	83%	88%	—	Yes
Percent grammatical utterances[i]	Narrative tell	Picture sequences (ENNI)	88%	84%	85%	Yes

(*table continues*)

Table 3. *(Continued).*

Measure	Elicitation task	Materials	Sensitivity	Specificity	Overall	Cutoff
9-year-olds						
Percent grammatical utterances[i]	Narrative tell	Picture sequences (ENNI)	90%	90%	90%	Yes
Cohesive ties + grammaticality + subordinate clauses/T-unit + words/subordinate clause[l]	Narrative retell	Movie	—	—	98%	No
10-year-olds						
Cohesive ties + grammaticality + subordinate clauses/T-unit + words/subordinate clause[l]	Narrative retell	Movie	—	—	98%	No
Percent "restricted" utterances[m]	Narrative tell	Wordless picture book	83%	88%	—	Yes
African American English and Southern White English speakers (4- to 6-year-olds)						
35 nonmainstream patterns[n]	Play	Toys (gas station, picnic/park, baby dolls, food, Legos, beads), picture scenes	87%	94%	90%	No
Southern White English speakers (4- to 6-year-olds)						
Irregular past + auxiliary DO + irregular third + infinitive TO + *don't* agreement[n]	Play	Toys (gas station, picnic/park, baby dolls, food, Legos, beads), picture scenes	87%	95%	—	No
Past tense (strategic scoring)[o]	Play	Toys (gas station set, picnic/park set, baby doll set), action pictures (visiting doctor's office, fishing, grocery shopping, washing a car)	89%	89%	89%	Yes

Note. Em dashes indicate data not reported. VP = verb phrase; NP = noun phrase; FVMC = finite verb morphology composite; MLU = mean length of utterance; ENNI = Edmonton Narrative Norms Instrument; DSS = Developmental Sentence Scoring; SUGAR = Sampling Utterances and Grammatical Analysis Revised; SALT = Systematic Analysis of Language Transcripts.

[a]Gavin et al., 1993. [b]Bedore & Leonard, 1998. [c]Gladfelter & Leonard, 2013. [d]Souto et al., 2014. [e]Guo et al., 2020. [f]Pavelko & Owens, 2019. [g]Guo & Schneider, 2016. [h]Fletcher & Peters, 1984. [i]Guo et al., 2019. [j]Heilmann et al., 2010. [k]Castilla-Earls & Fulcher-Rood, 2018. [l]Liles et al., 1995. [m]Hoffman, 2009. [n]Oetting & McDonald, 2001. [o]Oetting et al., 2021.

can achieve at least 90% sensitivity and specificity using conversation or play-based language samples, although more modest levels were found for the LARSP model in the validation study (91% sensitivity/80% specificity). For 4- and 5-year-olds, both the traditional and modified FVMCs yield good accuracy with play (Gladfelter & Leonard, 2013; Souto et al., 2014) and narrative (Guo et al., 2020) samples, as did the DSS sentence point with play samples (Souto et al., 2014). Although none of the measures examined with 6-year-olds reached 90% sensitivity and specificity, several reached the 80% threshold of acceptable level: FVMC, percent grammatical C-units (PGCU), errors per C-unit, MLU + CPS, GLi, unmarked verbs + verb types, and a combination of 10 SALT measures.

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

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

Quality of Evidence

The 15 publications reporting the measures in the previous section were examined for design features that indicate the quality or strength of the evidence using a checklist published by C. A. Dollaghan (2004). A one-gate design that recruits all participants from the same population is more likely to result in a participant sample that represents a continuum of ability or severity than a two-gate design that recruits from different sources (e.g., TD from a local school and DLD from a clinic). Selection of a valid and accurate gold standard reference measure and blinded administration of the measure to all participants by independent examiners ensure that group assignment reflects accurate and objective classification of impairment status. Positive and negative likelihood ratios (LR+ and LR−) are diagnostic accuracy metrics that are less vulnerable to small participant samples, although there is less consensus on a recommended threshold (Klee et al., 2017). Intermediate values of above 4.0 for LR+ and below 0.4 for LR− are suggested as a minimum to be considered conclusive, with values of 10.0 for LR+ and 0.2 for LR− indicating high likelihood of accurate classification in the corresponding ranges (C. A. Dollaghan, 2004). In addition, confidence intervals indicate how precise the diagnostic accuracy is likely to be across different groups. Clinical feasibility for LSA measures can include whether the cutoff value or regression equation for the measure(s) was reported, the number of measures that must be calculated, the length of the LSA transcript required, and access to required materials.

All studies used a two-gate design or did not report this information clearly. Twelve studies used a clinical criterion (i.e., a previous diagnosis by an SLP and/or current enrollment in language therapy) as the gold standard reference measure either alone, in addition to a standardized test or a parent report measure, or confirmed by such a measure (see Table S2 in Supplemental Material S1). A clinical criterion is widely regarded as an appropriate gold standard (C. Dollaghan & Campbell, 1998). Three studies used standardized tests as the reference measure, namely, the Test of Language Development–Primary: Second Edition or Preschool Language Scale–Third Edition (Bedore & Leonard, 1998), the Structured Photographic Expressive Language Test–Third Edition (Castilla-Earls & Fulcher-Rood, 2018), and the Norm-Referenced Diagnostic Evaluation of Language Variation (DELV-NR; Oetting et al., 2021). Of the tests used in these studies, only the Test for Examining Expressive Morphology and the DELV-NR have evidence of at least acceptable diagnostic accuracy with the age group and cutoff scores used (Eisenberg & Guo, 2013; Nitido & Plante, 2020; Spaulding et al., 2006). None of the studies clearly reported whether administration of measures was blinded.

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

Eight studies reported the cut score(s) or the regression equation used in determining the diagnostic accuracy of the measure (see Table 3). The cutoff for individual measures from two additional studies could be derived based on the midpoint between group means (Gladfelter & Leonard, 2013; Souto et al., 2014). Of these, analyses required calculation of only one to two LSA measures, except for the LARSP model that required three measures and child age. The analysis samples ranged from 33 to over 375 utterances long. Procedures that elicit 50–100 utterances will be more feasible in clinical practice (Heilmann, 2010; Pavelko et al., 2016) than those requiring more time-intensive elicitation or significantly longer samples (e.g., 1-hr protocol in Fletcher & Peters, 1984). The elicitation methods and materials are clearly described and publicly available for fidelity of implementation in a clinical setting for all but two studies.

Discussion

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

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

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

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

Limitations

One limitation of the current study is that, although the search terms allowed for the inclusion of a wide age range, the actual age range represented in the reviewed studies is fairly narrow. A substantial number of LSA measures have been tested with children between the ages of 3 and 6 years, but very few studies, which

examined a limited selection of measures or composites, were available for children past the age of 9 years and none for children older than 13 years. Considering that the accuracy of measures varies by age even among young children, as we see with FVMC, we cannot assume that "good" measures will still be useful if they have not been tested on children of that age. This mirrors the larger trend in speech-language pathology research, and so calls to expand research on adolescent language also apply in this case.

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

The current study limited the scope of the review to only English language sample data. This allowed for a more comprehensive synthesis of the patterns of findings within a single language. However, because clinical markers of DLD are language specific (Leonard, 2017), the diagnostic accuracy level and corresponding cutoff scores or equations found cannot be generalized to other languages, even if the measure can be readily applied (e.g., grammaticality). To facilitate best practices of assessing bilingual students in both of their languages using diagnostic LSA, future research should examine the diagnostic accuracy of LSA measures in languages other than English and compare the accuracy of measures cross-linguistically.

Implications for Future Research

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

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

Clinical Application

Despite the limitations and gaps that remain to be addressed, SLPs can apply the findings of this review in current practice by incorporating the LSA measures identified as having evidence of clinical utility, albeit preliminary, into their assessments with similar clients. Clinicians can refer to Table 3 to identify the measure(s) that would provide the most accurate diagnostic classification for the client's age. For measures with an available cutoff score, Supplemental Material S2 includes a summary of procedures and interpretation guidelines. A software-specific tutorial on how to automatically generate each measure is beyond the scope of this study; however, they appear to be generally compatible with the functionality of popular programs using either embedded commands (e.g., MLU; see Pezold et al., 2020, Supplemental Material S2, p. 1) or custom codes (see Pezold et al., 2020, Supplemental Material S1, Section 2, pp. 5–6). Future tutorials should explore the options for coding transcripts and computing the measures highlighted in this review across different software programs to enable clinicians to take full advantage of computer-assisted LSA using the most efficient procedures.

Conclusions

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

References

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

*Bedore, L. M., & Leonard, L. B. (1998). Specific language impairment and grammatical morphology: A discriminant function analysis. *Journal of Speech, Language, and Hearing Research, 41*(5), 1185–1192. https://doi.org/10.1044/jslhr.4105.1185

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

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

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

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

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

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

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

Crystal, D., Fletcher, P., & Garman, M. (1976). *The grammatical analysis of language disability: A procedure for assessment and remediation*. Edward Arnold.

Dollaghan, C., & Campbell, T. F. (1998). Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research, 41*(5), 1136–1146. https://doi.org/10.1044/jslhr.4105.1136

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

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)

*Dunn, M., Flax, J., Sliwinski, M., & Aram, D. (1996). The use of spontaneous language measures as criteria for identifying children with specific language impairment: An attempt to reconcile clinical and research incongruence. *Journal of Speech and Hearing Research, 39*(3), 643–654. https://doi.org/10.1044/jshr.3903.643

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

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

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

Evans, J. (1996). Plotting the complexities of language sample analysis: Linear and non-linear dynamical models of assessment. In K. Cole, P. Dale, & D. Thal (Eds.), *Assessment of communication and language* (pp. 207–256). Brookes.

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

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

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

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

*Gladfelter, A., & Leonard, L. B. (2013). Alternative tense and agreement morpheme measures for assessing grammatical deficits during the preschool period. *Journal of Speech,*

Language, and Hearing Research, 56(2), 542–552. https://doi.org/10.1044/1092-4388(2012/12-0100)

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

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

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

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

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

Heilmann, J. J. (2010). Myths and realities of language sample analysis. *SIG 1 Perspectives on Language Learning and Education, 17*(1), 4–8. https://doi.org/10.1044/lle17.1.4

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

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

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

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

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

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

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

Leonard, L. B. (2017). *Children with specific language impairment* (2nd ed.). MIT Press.

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

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

McWeeny, S., Choe, J., & Norton, E. (2021). *SnowGlobe: An iterative search tool for systematic reviews and meta-analyses*. https://doi.org/10.17605/OSF.IO/U25RN

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. (2008). *Systematic Analysis of Language Transcripts (SALT), English & Spanish* (Version 9) [Computer software]. University of Wisconsin–Madison Waisman Center, Language Analysis Laboratory.

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

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

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

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

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

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

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

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

***Pavelko, S. L., & Owens, R. E., Jr.** (2019). Diagnostic accuracy of the Sampling Utterances and Grammatical Analysis Revised (SUGAR) measures for identifying children with language

impairment. *Language, Speech, and Hearing Services in Schools, 50*(2), 211–223. https://doi.org/10.1044/2018_LSHSS-18-0050

Pavelko, S. L., Owens, R. E., Jr., 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

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

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

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

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

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

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

Rabin, B. A., & Brownson, R. C. (2017). Terminology for dissemination and implementation research. In R. C. Brownson, G. A. Colditz, & E. K. Proctor (Eds.), *Dissemination and implementation research in health: Translating science to practice* (pp. 19–45). Oxford University Press. https://doi.org/10.1093/oso/9780190683214.003.0002

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

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

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

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

***Schneider, P., Hayward, D., & Dubé, R. V.** (2006). Storytelling from pictures using the Edmonton Narrative Norms Instrument. *Journal of Speech-Language Pathology and Audiology, 30*(4), 224–238.

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

Shahmahmood, T. M., Jalaie, S., Soleymani, Z., Haresabadi, F., & Nemati, P. (2016). A systematic review on diagnostic procedures for specific language impairment: The sensitivity and specificity issues. *Journal of Research in Medical Sciences, 21*(1), 67. https://doi.org/10.4103/1735-1995.189648

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

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

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

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

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

Tomblin, J. B., Records, N. L., Buckwalter, P., Zhang, X., Smith, E., & O'Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research, 40*(6), 1245–1260. https://doi.org/10.1044/jslhr.4006.1245