ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

The Reliability of Short Conversational Language Sample Measures in Children With and Without Developmental Language Disorder

Amy Wilder[a] and **Sean M. Redmond**[a]

[a] Department of Communication Sciences and Disorders, The University of Utah, Salt Lake City

ARTICLE INFO

Article History:
Received November 25, 2021
Revision received January 20, 2022
Accepted January 21, 2022

Editor-in-Chief: Stephen M. Camarata
Editor: Megan York Roberts

https://doi.org/10.1044/2022_JSLHR-21-00628

ABSTRACT

Purpose: Language sample analysis (LSA) represents an ecologically valid method for diagnosing, identifying goals, and measuring progress in children with developmental language disorder (DLD). LSA is, however, time consuming. The purpose of this study was to determine the length of sample needed to obtain reliable LSA measures for children in kindergarten and first grade with typical language (TL) and DLD using automated analyses from the Systematic Analysis of Language Transcripts software.
Method: Play-based conversational language samples collected on kindergarten to first-grade children with TL (n = 21) and DLD (n = 21) from a community-based sample were analyzed. Eight LSA measures were calculated from 1-, 3-, 5-, 7-, and 10-min sample cuts and compared to 20-min samples for reliability.
Results: Reliability estimates were similar for the TL and DLD groups except for errors and omissions, which showed overall higher levels of reliability in the DLD group and reached acceptable levels at 3 min. Percent grammatical utterances were reliable at 7 min in the DLD group and not reliable in shorter samples in the TL group. The subordination index was reliable at 10 min for both groups. Number of different words reached acceptable reliability at the 3-min length for the DLD group and at the 10-min length for the TL group. Utterances and words per minute were reliable at 3 min and mean length of utterance at 7 min in both groups.
Conclusions: Speech-language pathologists can obtain reliable LSA measures from shorter, 7-min conversational language samples from kindergarten to first-grade children with DLD. Shorter language samples may encourage increased use of LSA.
Supplemental Material: https://doi.org/10.23641/asha.19529287

Developmental language disorder (DLD), a communication disorder characterized by persistent difficulty using and understanding language that cannot be explained by hearing loss, intellectual disability, or another medical condition, affects 7%–12% of the school-age population (Bishop et al., 2016; Leonard, 2014; Norbury et al., 2016; Tomblin et al., 1997). School-age children with DLD are at risk for increased social, emotional, and behavioral problems as well as decreased academic achievement and quality of life (Dubois et al., 2020; Eadie et al., 2018; Langbecker et al., 2020). Adequate early identification and intervention for students with DLD may alleviate adverse outcomes. To effectively provide these services, speech-language pathologists (SLPs) need psychometrically robust assessment tools that measure functional language skills to address the clinical needs of children with DLD. Language sample analysis (LSA) represents a well-established, ecologically valid assessment method for language disorders that addresses various assessment objectives, including diagnosis, goal identification, and progress monitoring (Finestack et al., 2020; Owens, 2016; Paul et al., 2017).

As a diagnostic tool, LSA offers several advantages over standardized omnibus language tests. Language

Correspondence to Amy Wilder: amy.wilder@utah.edu. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

samples can be used to assess difficulties using language skills for daily communicative purposes, such as having conversations, telling personal narratives, or explaining the rules of a game. Standardized language tests, which often involve pointing to pictures, repeating sentences, or completing a cloze sentence task, are, in contrast, noticeably unnatural. LSA offers an alternative to standardized tests when evaluating children who are English language learners, bilingual, from special populations, or difficult to test due to behavior issues. The majority of English standardized tests do not include normative data from bilingual children and, therefore, may not be appropriate, when used in isolation, for identifying cases of DLD in bilingual children (Ebert & Pham, 2017; Kohnert, 2010). LSA may be used as an unbiased measure and increase diagnostic accuracy when evaluating children who are English language learners, those who speak nonmainstream dialects, and those with low socioeconomic status (Ebert & Pham, 2017; Horton-Ikard, 2010; Lai & Schwanenflugel, 2016; Pieretti & Roseberry-McKibbin, 2016; Roseberry-McKibbin, 2007). Also, many language difficulties that are not evident from performances on standardized tests may be revealed through LSA (Price et al., 2010). Unlike standardized tests, which should only be repeated at designated intervals to preserve their validity, usually once per year, LSA can be repeated as often as needed to monitor progress and adjust goals. The information gained from LSA can be used to set functional goals and design interventions that are suited to individual children and are more likely to generalize to environments outside the therapy room (Owens, 2016). However, one substantial disadvantage of LSA has been that it is time consuming.

Traditional guidelines suggest language samples of 15–30 min or 50–200 utterances are required to reliably measure children's language abilities (Lee, 1974; Miller, 1981; Tyack & Gottsleben, 1974; Retherford et al., 2019). In addition to the time required to collect the sample, transcription and analysis, when done by hand, can add 1–3 hr of clinician time, depending on the length of the sample, the particular analyses used, and the clinician's proficiency with these tasks. School-based SLPs consistently list time constraints as the number one reason for not using LSA (Kemp & Klee, 1997; Pavelko et al., 2016; Westerveld & Claessen, 2014). Despite over 20 years of advances in computer-assisted LSA technology providing SLPs with significant reductions in analysis time (Heilmann et al., 2020; Pezold et al., 2020; Ratner & MacWhinney, 2016), surveys indicate the use of LSA by SLPs declined from 85% in 1994 (Kemp & Klee, 1997) to 67% in 2013 (Pavelko et al., 2016). One solution might be to leverage the potential power of shorter language samples. Many SLPs have reported using shorter samples of 1–10 min or of 50 or fewer utterances, and 50%–60% reported transcribing samples in real time (Kemp & Klee, 1997; Pavelko et al., 2016). Fortunately, there is some evidence for the reliability of LSA measures based on shorter samples compared to longer language samples. Even so, debate continues regarding sufficient sample length for individual measures. Unfortunately, much of the research has focused on language samples collected from children with typical language (TL) skills. Evidence of robust reliability for LSA measures using short language samples from children with language disorders is sorely needed. This study aims to determine the shortest sample length needed to obtain reliable LSA measures in conversational samples from kindergarten and first-grade children with DLD and TL.

Reliability

Reliability quantifies the amount of error inherent in a measurement (Streiner et al., 2015). The reliability of an instrument can be estimated in several ways. Test–retest reliability, or stability, refers to the change or lack of change in examinee responses over time or across different testing situations. Interrater reliability refers to the degree of agreement between independent assessors who score, rate, or code the instrument. Split-half reliability, which typically compares the scores from at least two sections of a standardized test (e.g., odd vs. even items on a subtest), has been adopted to determine whether measures from a shorter language sample result in roughly the same scores as those from a longer sample. Internal consistency, typically used to measure the agreement of multiple items within a scale, has also been adopted to measure the reliability of shorter samples compared to longer samples.

Relative reliability refers to how consistently individuals maintain their rank position across repeated measures, whereas absolute reliability measures the amount of variation in an individual's scores across repeated measurements. Relative reliability is often estimated with Pearson correlation, Cronbach's alpha, or intraclass correlation coefficients (ICCs). However, Pearson correlations may underestimate split-half reliability when there is an imbalance of items between each segmentation of the test. The Spearman–Brown formula corrects for this by converting the Pearson correlation to an estimate of the equivalent reliability level for the full-length test (Kaplan & Saccuzzo, 2018). Cronbach's alpha coefficients estimate internal consistency or the relatedness of items within a scale or subtest. To determine the length of language sample needed to obtain measures equivalent to a full-length reference sample, each cut of the language sample is compared to the full-length sample. This method results in measuring a series of two-item scales rather than several items within one scale. Using mathematical models and collected data, Eisinga et al. (2013) found that Spearman–Brown represented the most accurate

method for calculating reliability on a two-item scale across three measures: Spearman–Brown, Pearson, and Cronbach's alpha.

There is no conventional cutoff separating acceptable from unacceptable relative reliability coefficients. However, several researchers (Gavin & Giles, 1996; Guo & Eisenberg, 2015) have adopted $r \geq .90$ as a suitable threshold for Pearson correlation coefficients, following guidelines provided by McCauley and Swisher (1984). For alpha coefficients, Devellis (2012) suggested $\alpha = .70–.80$ as an "acceptable" reliability standard and $\alpha \geq .80$ as "very good" but further stipulated that scales intended for clinical diagnostic purposes should have higher reliabilities, preferably in the mid-90s.

One potential issue with estimates of relative reliability is that they increase as sample variability increases. That is, with greater variability across individuals, individual scores can change more without changing the individual's rank in the group. This could be problematic when estimating reliability on samples collected from children with DLD, a fairly heterogeneous group. In contrast, sample variability does not influence absolute reliability since it estimates the agreement of different scores from the same individual. Absolute reliability is often estimated with the coefficient of variation (CV) or limits of agreement (LOA) using Bland–Altman (B-A) plots (Bland & Altman, 1986). The CV is calculated by dividing the standard deviation by the group mean. CV estimates are limited by the lack of a standard interpretation of CV values. Also, acceptable CV values may vary by measure.

The B-A method involves plotting the differences between two scores against the means of both scores, calculating their LOA, and then determining the percentage of data points within the LOA (Bland & Altman, 1986, 1999). LOA may be set at a default value of ±1.96 *SD*s of the mean difference or may be alternatively set at a predetermined clinically acceptable difference. This strategy is predicated on the existence of consensus within the field on what constitutes a "clinical meaningful unit of change" for a particular measure. Alternatively, if one measure is considered the reference measure, the differences of the scores can be plotted against the scores from the reference measure. B-A plot analysis shows bias between the measures as the distance between zero and the mean differences. For example, if the mean difference is −5, the second measure, on average, measures 5 units more than the first one. Visual analysis also shows any relation between difference and magnitude (see B-A plot interpretation guide in Supplemental Material S4 for additional information). The B-A method may represent a more appropriate estimate of reliability for ratio data and may better represent clinical acceptability (Bruton et al., 2000; Pavelko et al., 2020).

Reliability of LSA Measures in Children With TL

Mean Length of Utterance

Mean length of utterance (MLU) in morphemes is a conventional measure of overall grammatical development sensitive to age-related changes in younger children (Brown, 1973; Miller & Chapman, 1981). Stable differences in MLU have been demonstrated between children with TL and DLD over a wide range of ages (3–8 years; Bedore & Leonard, 1998; Pavelko & Owens, 2019; Rice et al., 2006). MLU values for 5- to 7-year-old children with DLD often fall in the 4.0–4.5 range compared to the 5.0–5.5 range associated with children with TL (Rice et al., 2010), representing an expected average MLU difference between affected and unaffected cases of 1.0. Previous studies yielded mixed results when determining the minimum sample needed for reliable MLU calculation for children with typically developing language.

Darley and Moll (1960) found "adequate" split-half reliability ($r = .85$) for MLU in 50-utterance picture description language samples from 5-year-old children with TL ($n = 150$). In contrast, Gavin and Giles (1996) found that 100 utterances were required to reach "acceptable" levels of relative reliability ($r = .82$, $p < .0001$), whereas 175 utterances were needed for "very high" reliability ($r = .93$, $p < .001$) for MLU using two parent–child play-based language samples collected on 2- to 3-year-old young children with TL ($n = 20$) within a 3- to 14-day period. These different findings could be attributed to differences in participants' age, the sampling contexts, or sample sizes.

More recently, Heilmann et al. (2010) examined the relative and absolute reliability of MLU, calculated using Systematic Analysis of Language Transcripts (SALT; Miller et al., 2019), in both conversational and narrative samples from children with TL ($n = 231$) in a range of ages (younger group = 2;8–5;11 [years;months], older = 6;0–13;13). A repeated-measures analysis of variance (RMANOVA) found no significant differences for MLU among 1-, 3-, and 7-min transcript lengths ($p = .54$, $\eta^2 < .01$). No significant interactions were found between sample length, context, and age (ps $\geq .12$, η^2s $< .01$), suggesting a negligible amount of variance was due to the interactions between length, context, and age. When examining the relative reliability of MLU using Cronbach's alpha tests, alpha values ranged from "undesirable" to "acceptable" (αs = .56–.79) when comparing 1-min ($M_{\text{utterances}}$ = 12–16) and 3-min ($M_{\text{utterances}}$ = 36–48) cuts to 7-min samples across sampling contexts and age groups. Absolute reliability was measured using the CV. CV values for MLU ranged from 0.15 to 0.37 and showed differences of 0.04–0.14 from 1- to 7-min samples and 0.03–0.07 from 3- to 7-min samples, indicating minor changes in MLU

between 3- and 7-min samples. These results suggest that obtaining an MLU that meets the more stringent criteria of $\alpha \geq .90$ requires samples longer than 3 min or 40 utterances. However, differences in MLU between the 3- and 7-min samples do not appear to be clinically meaningful.

Using SALT analysis, Guo and Eisenberg (2015) compared MLUs from 1-, 3-, 7-, and 10-min cuts to 22-min play-based conversational samples from children with TL ($n = 60$) ages 3;0–3;11. RMANOVAs showed no significant differences for MLU among sample lengths ($ps > .29$, $\eta_p^2 < .019$). Using Pearson correlations, MLU reached "acceptable" relative reliability at the 10-min cut ($r = .93$, $p < .01$, $M_{utterances} = 90$). These findings indicate 10-min samples, or around 90 utterances, are needed for reliable MLU measures in 3-year-old children with TL. Taken together, the findings of these four studies suggest that language samples around 10 min or 60–100 utterances in length will result in estimates of MLU consistent with those from longer samples. In contrast, Pavelko et al. (2020) found MLU was reliable in 25-utterance conversational samples using a different protocol for MLU calculation than the traditional Brown (1973) conventions.

Pavelko et al. (2020) measured the relative and absolute reliability of MLU from 25- and 50-utterance samples from children with TL ($n = 220$) ages 3;2–7;10 using the Sampling Utterances and Grammatical Analysis Revised (SUGAR) method (detailed in Pavelko & Owens, 2017). Pavelko et al. employed a mixed-model analysis to estimate relative reliability using between-subjects variability and absolute reliability using within-subject variability. The results revealed no significant differences for mean MLU scores between the 25- and 50-utterance samples ($p = .64$, $d = -0.03$, 95% confidence interval [CI] [−0.22, 0.16]). The results also showed significant variation for MLU; however, the 95% CI for the effect size included zero, indicating this result was not clinically significant ($p = .0001$, $d = 0.04$, 95% CI [−0.22, 0.14]). To examine the clinical significance of these results, the authors examined B-A plots with the differences of MLUs for 25- and 50-utterance samples plotted against the means. LOA were calculated using ±1.96 *SD*s of the differences. The analyses showed that 95% of the data points fell within the LOA. At face value, this represents a desirable outcome. However, in practice, their estimated LOA (−2.21, 2.29) would allow for differences between estimates of MLU to be as high as 4.5. Even so, the authors proposed their results indicated reliable MLU scores could be obtained from 25-utterance conversational language samples in children ages 3;2–7;10 with TL skills.

Besides MLU, the most researched LSA measure, examinations of the reliability of other LSA measures have yielded mixed findings. Unlike MLU, these measures do not have readily available developmental frameworks to extrapolate a clinically meaningful unit of growth or change.

Words per Minute

Words per minute (WPM) is a measure of verbal productivity and fluency shown to increase with age (Miller, 1981). Two of the previously reviewed studies examined the reliability of WPM in shorter language samples with fairly similar results. Heilmann et al. (2010) found no significant differences for WPM in 1- and 3-min samples compared to 7-min samples ($p = .67$, η^2s $< .01$), with no significant interaction effects between length, age group, and sampling context ($ps \geq .28$, η^2s $< .01$). Cronbach's alpha values showed "respectable" relative reliability for WPM in the 1-min samples for both age groups in both contexts (αs $\geq .80$) and "good" relative reliability in the 3-min samples for the older age group in both contexts (αs $\geq .92$). CV values showed modest differences ranging from .02 to .16 between the 1- and 7-min samples and from .01 to .03 between the 3- and 7-min samples, indicating good absolute reliability.

In contrast, Guo and Eisenberg (2015) found WPM was significantly larger in 3-min samples compared to 22-min samples ($p = .001$, η^2s $= .168$). However, there were no significant differences for WPM in the 1-, 7-, and 10-min samples compared to the 22-min samples ($ps \geq .07$, η^2s $< .06$). Pearson correlations showed WPM reached "adequate" reliability at the 7-min length ($r = .92$, $p < .01$). These studies suggest a reliable WPM measure can be obtained from a sample of around 7 min or 60 utterances in younger children and from a sample of 3 min or 30–40 utterances in older children.

Number of Different Words

Number of different words (NDW) is a measure of lexical diversity that has been shown to differentiate children with DLD from children with TL (Hewitt et al., 2005; Watkins et al., 1995). Gavin and Giles (1996) calculated test–retest reliability coefficients for NDW using play-based language samples from preschool children with TL. They found that 150 utterances were required to reach "acceptable" reliability ($r = .83$, $p < .001$) and 175 utterances were required for "very high" reliability ($r = .93$, $p < .001$). In contrast, Guo and Eisenberg (2015) found "acceptable" reliability ($r = .92$, $p < .01$) for 10-min samples ($M_{utterances} = 91$) when compared with 22-min play-based samples from preschool children with TL. Heilmann et al. (2010) also found NDW to be reliable in shorter samples. In their sample, NDW reached "very good" relative reliability (αs $\geq .81$) and absolute reliability (CV differences = 0.05–0.13) in 1-min samples in younger and older age groups in both narrative and conversational contexts. Because reliability was calculated using different metrics, it is hard to synthesize results across these studies.

Sentence Complexity

Measures of sentence complexity have been shown to detect language growth throughout the school-age years

(Nippold et al., 2007). Studies have reported that children with DLD, as a group, use simpler sentence structures than children with TL (Bishop & Donlan, 2005; Marinellie, 2004; Scott & Windsor, 2000). However, little is known about the reliability of sentence complexity measures from short language samples. Two studies examined the reliability of sentence complexity measures with conflicting results. Darley and Moll (1960) found that a structural complexity measure scoring utterances based on the number of phrases and clauses did not show adequate reliability (r = .69) in 50-utterance picture description samples from typically developing 5-year-old children. Pavelko et al. (2020) examined the relative and absolute reliability of clauses per sentence (CPS) in 25- and 50-utterance conversational samples from children ages 3;2–7;10. They found significant differences between CPS mean scores (p = .012, d = −0.2) with significant variation (p = .0001, d = 0.19). Further analysis using a B-A scatter plot indicated that the variation for CPS was not clinically significant as 96% of scores fell within the lines of agreement (−0.20, 0.24; M difference = 0.019), suggesting differences between the 25- and 50-utterance samples were not clinically meaningful. Differences in measures and analyses used across studies make direct comparisons difficult. Therefore, whether sentence complexity measures can be reliably obtained from short language samples remains an open question.

Additional Measures

An increased number of mazes, including false starts, revisions, disfluency repetitions, and filler words (e.g., *um*, *uh*, *er*), may indicate difficulty with sentence formulation or word-finding problems (Miller, 1996). Some studies have found significant group differences for the percentage of maze words (PMW) used by children with TL and DLD (MacLachlan & Chapman, 1988; Thordardottir & Weismer, 2002). Heilmann et al. (2010) found relative reliability levels for the PMW ranged from "unacceptable" to "acceptable" (αs = .39–.79) in 1- and 3-min conversational and narrative samples from children with TL. Differences in CV values ranged from 12% to 40% between 1- and 7-min samples and from 4% to 16% between 3- and 7-min samples, indicating variability across sample lengths. However, if children with DLD produce relatively more maze words, this measure may be reliable in short samples from children with DLD.

Eisenberg and Guo (2015) calculated split-half reliability for the percent grammatical utterances (PGU) in picture description language samples from 3-year-old children with TL. Comparing sample cuts from two 7-picture sets ($M_{utterances}$ = 32, 30) to the entire 15-picture samples ($M_{utterances}$ = 67) showed no significant differences for PGU (ps > .45, η_p^2 < .02). Pearson correlations showed "acceptable" levels of relative reliability (rs > .95, ps < .01), suggesting 30-utterance samples are adequate for measuring PGU in this age group for children with TL. Heilmann et al. (2010) also examined the reliability of omissions and errors in short language samples. Cronbach's alpha values indicated this measure was unreliable in 1- and 3-min conversational or narrative samples from children with TL (αs = .51–.69). Differences in CV values ranged from 0.64 to 1.45, representing significant variability, considering the mean number of errors and omissions (EAO) ranged from 0.7 to 1.6. Again, this measure may be more reliable in children with DLD than those with TL as they produce grammatical EAO at higher rates.

In summary, some LSA measures, including MLU, WPM, and NDW, have demonstrated acceptable levels of relative reliability when based on shorter language samples. Multiple independent investigations of relative reliability suggest language samples around 7 min or 50–60 utterances in length are probably sufficient. Estimates of absolute reliability, including the CV and B-A plots provided by some investigators, are more challenging to interpret without agreed-upon cutoff values. Other LSA measures, including sentence complexity, PMW, and EAO, may require longer samples when used with children with TL. Curiously, more studies have investigated the reliability of LSA measures in children with TL than in children with language disorders. Evidence for the reliability of short language samples is needed in this population because SLPs typically collect language samples on children with language disorders. A few studies have examined the reliability of LSA measures from short language samples in children with DLD with results similar to those from children with TL.

Reliability of LSA Measures in Children With Language Disorders

MLU

Cole et al. (1989) examined reliability for MLU using play-based samples from children with language delay ages 4;4–6;8 (n = 10). MLU showed good split-half reliability (ρ = .95) and good test–retest reliability (ρ = .92) from two 100-utterance samples collected 2 weeks apart. Tilstra and McMaster (2007) found MLU in words was unreliable (rs = .01–.63) when comparing three short picture-elicited narrative samples ($M_{utterances}$ = 12) from children ages 5–9 years (n = 45), including children with TL; those receiving special services for reading, speech, or language; and English language learners. Casby (2011) compared MLU for 10-, 20-, and 50-utterance cuts with 100- to 150-utterance conversational samples from children with DLD (n = 10) ages 3;0–11;8. Pearson correlations ranged from .52 to .94 depending on the cuts' utterance length and location (e.g., first 10, middle 10, last 10, or 10 random utterances). Selecting 50 random utterances showed good reliability (r = .94). These results suggest

that, similar to children with TL, samples of at least 50 utterances are required for reliable MLU measures in children with DLD.

Additional Measures

Tilstra and McMaster (2007) also examined the reliability of WPM, clauses per C-unit, and total grammatical errors among three short picture-elicited narrative samples ($M_{utterances}$ = 12). WPM showed acceptable reliability in the first- and third-grade groups but not in the kindergarten group. Conversely, grammatical errors showed good reliability in the kindergarten group but not in the first- and third-grade group. Clauses per C-unit showed poor reliability in all three age groups. Guo et al. (2021) examined the split-half reliability of a clausal density measure in narrative samples of children ages 4–9 years with TL and with DLD. Reliability was calculated for the two groups combined using samples that ranged in length from 33 to 174 utterances with means ranging from 58 to 81 by age group. Pearson correlations by age group ranged from .54 to .86, $p < .001$, which the authors interpreted as showing "appropriate" reliability. However, only the 5-year-old group met the criteria of $r \geq .80$. Guo et al. (2019) examined the split-half reliability of PGU in narrative samples from children with TL (n = 300) and DLD (n = 77) ages 4–9 years. The samples varied in length from 9 to 148 utterances; mean utterances by age group ranged from 52 to 80. They found "acceptable" reliability for PGU in the 4-, 5-, 7-, and 8-year-old groups (rs $\geq .82$, ps $< .001$), whereas reliability was nearly at acceptable levels in the 6-year-old group ($r = .79$, $p < .001$) and was below acceptable levels in the 9-year-old group. ($r = .63$, $p < .001$).

In summary, limited information suggests that MLU and WPM show similar levels of reliability in children with DLD compared to children with TL, indicating these measures may be reliably obtained from shorter language samples for children with DLD. Reliable measures of clausal density appear to require samples greater than 12 utterances for children with DLD but may be reliable in 70-utterance samples for certain age groups. Finally, PGU may be obtained reliably in children with DLD using samples around 70 utterances in length. However, further study is needed to determine the minimum sample length required for reliable measures for PGU. Further examination into the length of sample needed to assess additional LSA measures reliably in children with DLD is needed.

The Current Study

The purpose of this observational, case–control study was to determine the shortest conversational language sample length, as measured in minutes, needed for reliable LSA measures based on automated SALT analyses for children in kindergarten and first grade with TL skills and DLD. To this end, we addressed the following research questions.

1. Are there significant main effects for sample length? Specifically, are LSA measures based on differing sample lengths (1, 3, 7, and 10 min) significantly different from a 20-min standard reference when examined with a series of RMANOVAs?
2. Do the effects of length across LSA measures vary as a function of group status (TL, DLD)? In other words, are there significant Length × Group interactions?
3. Do LSA measures from shorter samples demonstrate adequate levels of relative reliability as determined with Spearman–Brown correlation coefficients ≥ .90?
4. Do LSA measures from shorter samples demonstrate adequate levels of absolute reliability as determined with CVs and B-A plots showing ≥ .95% of data points within LOA of ±1.96 *SD*s of mean differences?
5. Does the sample length needed for adequate levels of reliability ($\rho \geq .90$, LOA ≥ 95%) for LSA measures vary as a function of group status (TL, DLD)?
6. Are there significant differences, measured with Fisher Z transformation, in the reliability levels of short-sample LSA measures collected on children with DLD and children with TL?

Based on prior studies, we predicted that the number of utterances, WPM, NDW, and MLU from shorter language samples would show similar levels of reliability across children with TL and children with DLD. We also predicted that lengths needed to establish reliability would vary across LSA measures. We also predicted that EAO and PGU would be more reliable measures in samples collected from children with DLD than samples from children with TL, as children with DLD produce grammatical EAO more frequently.

Method

Approval was obtained from the University of Utah's institutional review board for the following procedures. Reporting follows the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement (Von Elm et al., 2007). See Supplemental Material S8 for STROBE checklist.

Participants

Language samples used for this study were collected as part of a previous investigation of school-based

language screening protocols (Redmond et al., 2019). Some of these samples were used in an earlier study of children's productions of infinitive clauses (Wilder & Redmond, 2021). In the language screening study, students in kindergarten through third grade (n = 254) enrolled in regular education and those receiving school-based services for speech disorders, language disorders, reading disabilities, learning disabilities, or behavioral disorders were recruited. Children with multilingual status, with clinically low levels of nonverbal IQ (< 70), or who failed a hearing screening were excluded. Participants were also required to pass the Test of Early Grammatical Impairment phonological probe. This probe determines whether a child can accurately produce a small set of phonemes, /s, z, t, d/, accurately when they appear in word-final position. Sufficient accuracy with these particular phonemes in word-final contexts ensures that observed morphological omissions are due to children's grammatical limitations rather than their phonological limitations. The percentage of intelligible words in our language samples ranged from 96% to 100% (M = 98.5%) for the TL group and from 91% to 100% (M = 96.4%) for the DLD group.

In the language screening study, play-based conversational language samples were collected from all kindergarten and first-grade participants who came in for lab-based confirmatory assessments (n = 119). From that subgroup, we identified 24 participants who met our criteria for DLD. These criteria included a standard score (M = 100, SD = 15) of ≤ 85 on the Clinical Evaluation of Language Fundamentals–Fourth Edition (CELF-4; Semel et al., 2003) Core Language Index and a standard score of ≥ 70 on the Naglieri Nonverbal Ability Test (NNAT; Naglieri, 1997). A cutoff score of 85 on the CELF-4 Core Language Index shows 100% sensitivity and 82% specificity for identifying cases of DLD (Nitido & Plante, 2020) and is commonly used by both researchers and clinicians (Betz et al., 2013; Caesar & Kohler, 2009; Finestack & Satterlund, 2018). Because we had access to a community-based sample of children, we elected to take advantage of the opportunity and included unidentified cases of language disorder to enhance our external validity. Three participants, one who did not complete our nonverbal assessment and two whose language samples were shorter than 20 min, were excluded from the current study, resulting in a final sample of n = 21 participants in the DLD group (M_{age} = 6;2, range: 5;5–7;7). Of the 21 children identified as having DLD, 11 were receiving speech pathology services at the time of the study, and four had previously been diagnosed as having a speech/language impairment but were not receiving services at the time of the study, according to parental report.

We matched participants with DLD to participants with TL by age within 3 months (mean age = 6;2, range: 5;6–7;5). Criteria for assignment to the TL comparison group included a standard score between 70 and 120 on the NNAT and a CELF-4 Core Language Index score between 86 and 115. Upper limits for NNAT and CELF scores were imposed on the TL group to ensure that group comparisons would not be distorted by the overrepresentation of children with significantly above-average/gifted verbal or nonverbal skills. There was no significant group difference for NNAT scores; both group means were close to expected average levels of performance. As expected, there were significant differences between groups in their CELF-4 core language scores, with the TL group mean aligning with an "average" level of performance and the DLD group mean consistent with "moderate-to-severe" levels of impairment. Additional exclusionary criteria for both the DLD and TL comparison groups included a history of hearing loss, neurological damage, a genetic syndrome, or a diagnosis of autism spectrum disorder. Other participant characteristics, including sex, race, ethnicity, and maternal education, are presented in Table 1.

Table 1. Participant characteristics means (standard deviations) and ranges.

Group	Sex	Age (years;months)	Mom's ed[a]	Race/ethnicity	NNAT	CELF
TL	11 male	6;2	4.05*	20 White	104.86	104.71***
n = 21	10 female	(0;7)	(0.74)	1 African American	(9.15)	(8.22)
		5;6–7;4	3–5	2 Hispanic	85–118	88–115
DLD	14 male	6;2	3.43*	17 White	97.57	66.19***
n = 21	7 female	(0;8)	(0.93)	1 African American	(14.48)	(14.26)
		5;5–7;7	2–5	2 Asian, 1 "other"	73–122	40–84
				1 Hispanic		

Note. ed = education; NNAT = Naglieri Nonverbal Ability Test; CELF = Clinical Evaluation of Language Fundamentals; TL = typical language; DLD = developmental language disorder.

[a]1 = *some high school*, 2 = *high school diploma/GED*, 3 = *some college/technical training*, 4 = *four-year college degree*, and 5 = *some grad school/advanced degree*.

*p < .05. ***p < .001.

Language Sampling Procedure

Graduate student clinicians elicited 30-min play-based conversational samples in a laboratory setting using a standard kit that included a house, a barn, little people, farm animals, and furniture. Prior to data collection, examiners were trained to proficiency on evidence-based elicitation procedures, including speaking with short sentences, pausing, and limiting their use of yes/no questions. Examiners were also trained to use open-ended prompts to elicit children's utterances (see Hadley, 1998). The play sessions were audio- and video-recorded for later transcription.

Transcription Procedure and Accuracy

Graduate student research assistants who were naive to the participants' group status transcribed and coded the conversational samples following standard conventions from the SALT (Miller et al., 2019). Transcribers were required to complete SALT training courses (SALT Software, 2021) and reach at least 85% agreement for morpheme, word, utterance boundary, and SALT codes with a standardized practice sample before transcribing and coding data for this study. A second research assistant checked transcription and coding for each conversational sample. Any disagreements between the original transcriber/coder and the checker were resolved through consensus. Twenty percent of the vetted samples from the TL and DLD groups (five samples each) were randomly selected and independently transcribed and coded again to estimate interrater reliability. ICCs based on consistency for average measures were calculated using a two-way random-effects model for each of the eight LSA measures. ICC values between .75 and .9 indicate good reliability, and values ≥ .90 indicate excellent reliability (Koo & Li, 2016). The following ICCs for the TL group indicated excellent levels of interrater reliability: utterances per minute (UPM) = .998, WPM = .997, NDW = .996, MLU = .960, PMW = .975, EAO = .987, PGU = .965, and subordination index (SI) = .949. Interrater reliability for the DLD group ranged from good to excellent, indicated by the following ICCs: UPM = .998, WPM = .984, NDW = .983, MLU = .994, PMW = .804, EAO = .979, PGU = .942, and SI = .963.

LSA Procedure

The first 5 min of the conversational samples were excluded from analysis to control for any potential warm-up effects (see Miller, 1981) and due to the noise of dumping out and setting up toys resulting in increased unintelligible utterances. The samples were divided into 1-, 3-, 7-, 10-, and 20-min cuts starting at the 6-min time code. These time points were selected following Guo and Eisenberg (2015) and Heilmann et al. (2010). Eight language sample measures were generated using SALT software (Miller & Iglesias, 2012), providing 40 measures on each sample. See Table 2 for a description of these measures.

Statistical Analyses

The data were analyzed using IBM SPSS Statistics Version 26 software (IBM Corp., 2019). We completed a series of RMANOVAs with the language sample measures as dependent variables, DLD or TL status as a between-group variable, and sample length as a within-group repeated measure. Absolute value measures (total utterances, NDW, and EAO) were divided by the number of minutes in the segment for RMANOVAs. Next, Spearman–Brown correlation coefficients were calculated to measure the relative reliability of each measure at the 1-, 3-, 7-, and 10-min lengths compared to the 20-min samples. Additionally, Pearson correlations and Cronbach's alpha tests

Table 2. Descriptions of selected language sample measures calculated with Systematic Analysis of Language Transcripts (SALT).

Language sample measure	Description
Number of total utterances	All utterances, including partial or abandoned utterances and those with unintelligible words.
Words per minute	The number of words, including those in partial utterances and excluding maze words, divided by the number of minutes in the sample length.
Number of different words	The number of different words, including those in partial utterances and excluding maze words.
Mean length of utterance (MLU) in morphemes	MLU calculated by SALT using the analysis set (complete and intelligible utterances).
Percentage of maze words	The number of maze words, including false starts, revisions, disfluency repetitions, and filler words, divided by the number of total words.
Errors and omissions	The combined number of grammatical errors and omissions.
Percent grammatical utterances	The number of grammatical utterances, those with no grammatical errors or omissions, divided by the number of total utterances.
Subordination index	The ratio of the total number of clauses (main and subordinate) to the number of C-units. Subordinate clauses included adverbial, relative, and complement clauses excluding infinitive clauses. Independent coordinate clauses were transcribed as a separate C-unit.

(Cronbach, 1951) were calculated as another measure of relative reliability to allow for comparison to previous studies. Absolute value measures were divided by the number of minutes when calculating Pearson correlations. Alpha values for the absolute value measures were based on standardized items due to differences in scaling across sample lengths, whereas alpha values for the ratio measures (WPM, MLU, PMW, PGU, and SI) were not adjusted.

The CV was calculated as an estimate of absolute reliability for each LSA measure across time cuts and by group. B-A plots for MLU and SI measures were generated using MedCalc (2021) for Windows (Version 20.014) as an additional estimate of absolute reliability and for comparison with Pavelko et al. (2020). The individual differences between LSA measures for each time cut and the 20-min length were plotted on the *y*-axis, and the individual LSA measures for the 20-min length were plotted on the *x*-axis as a reference standard (Bland & Altman, 1986, 1999; Krouwer, 2008). Separate plots were generated for each LSA measure by time cut (1, 3, 7, and 10 min) by group (TL, DLD). Following Pavelko et al. (2020), LOA were calculated using ±1.96 *SD*s of the differences. B-A plots were also interpreted using clinically significant LOA set at ±1.0 for MLU, following Pavelko et al. (2020). We selected an MLU difference of < 1.0 based on the logic that each increase of 1.0 on the MLU scale generally aligns with a grammatical stage under Brown's (1973) influential model. Thus, we would consider estimates of MLU that are discrepant by more than 1.0 to represent a clinically significant level of disagreement because they are associated with different stage assignments (Miller & Chapman, 1981; Paul et al., 2017). An MLU value of 1.0 also corresponds to the magnitude of observed group differences provided by previous comparisons of children with DLD and TD in the age range considered here. Finally, Fisher *Z* transformations were applied to the obtained Pearson correlations across our measures to identify potentially significant group differences in our reliability estimates.

Results

Significant Main Effects for Length Were Observed

Means, standard deviations, and ranges for the eight language sample measures are presented in Table 3, divided by group (TL, DLD) and sample length (1, 3, 7, 10, and 20 min). RMANOVAs were calculated to examine differences between the shorter samples and the 20-min samples. Mauchly's test indicated the assumption of sphericity was violated for the main effects of sample length and the Length × Group interaction effects for all eight language sample measures. Therefore, Greenhouse–Geisser estimates of sphericity were used. A series of RMANOVAs revealed significant differences across sample lengths for NDW per minute ($p < .001$, $\eta_p^2 = .627$). Main effects for length were nonsignificant for the seven other language sample measures (ps = .20–.98, η_p^2s = .003–.040). See Table 4 for a summary of the results.

Length × Group Interactions Were Nonsignificant

Length × Group interactions were nonsignificant for all eight LSA measures, indicating length effects for LSA measures did not vary as a function of group status (ps = .06–.75, η_p^2s = .005–.071). Main effects for group were significant (ps ≤ .02) for six of the eight LSA measures (WPM, NDW, MLU, EAO, PGU, and SI) with medium-to-large effect sizes (η_p^2 = .13–.40). See Table 4 for a summary of the results.

Estimates of Relative Reliability Varied Across Measures and Groups

Spearman–Brown coefficients were calculated to further examine the reliability of these measures in shorter language samples relative to the 20-min benchmark. In the TL group, ρ coefficients ranged from a low of −.36 for MLU in a 1-min sample to a high of .97 for number of total utterances (NTU) in a 7-min sample. The DLD group showed a range of ρ coefficients from a low of .21 for SI in a 1-min sample to a high of .98 for NDW and WPM in a 10-min sample. Table 5 provides a summary of the results divided by group and sample length. Following McCauley and Swisher (1984), we considered ρ coefficients ≥ .90 as acceptable levels of reliability.

Cronbach's alpha coefficients were similar to Spearman–Brown coefficients and ranged from .55 to .97 in the TL group and from .14 to .98 in the DLD group (see Supplemental Material S1 for results). Sample lengths needed to reach adequate levels of reliability were identical to those with ρ coefficients except for WPM in the DLD group, which increased from 3 to 7 min, and NDW in the TL group, which decreased from 10 to 7 min. The results for the TL group replicated results from Heilmann et al. (2010) for WPM, NDW, and errors and showed slightly lower alpha values for MLU and PMW.

Pearson coefficients showed overall lower levels of reliability compared to Spearman–Brown coefficients, ranging from −.22 to .94 in the TL group and .18 to .97 in the DLD group, resulting in longer sample lengths to reach acceptable levels of reliability for nearly all measures in both groups (see Supplemental Material S2). These results replicated the work of Guo and Eisenberg (2015) for WPM and NDW and showed a lower correlation for MLU by comparison for the 3-min length. Our results for MLU, WPM, and NDW replicated the results from Guo and Eisenberg for the 7- and 10-min segments.

Table 3. Means (standard deviations) and ranges of language sample measures by group and length.

Measure	Group	1 min		3 min		7 min		10 min		20 min	
		M (*SD*)	Range	*M* (*SD*)	Range	*M* (*SD*)	Range	*M* (*SD*)	Range	*M* (*SD*)	Range
Frequency-based measures											
Number of total utterances (NTU)	TL	12.24 (4.93)	3–25	35.05 (12.02)	12–68	81.43 (28.23)	38–163	115.14 (38.32)	50–222	230.1 (73.73)	102–440
	DLD	9.38 (4.88)	2–17	30.38 (14.12)	8–53	72.14 (28.83)	25–110	105.33 (38.22)	46–156	212.81 (66.50)	88–303
Number of different words (NDW)**	TL	29.62 (10.62)	3–47	65.57 (19.10)	34–110	116.52 (29.53)	69–195	145.76 (32.99)	86–232	232.95 (45.86)	142–353
	DLD	20.95 (12.42)	4–48	48.05 (23.88)	9–84	84.38 (34.32)	24–139	111.38 (37.48)	48–174	170.76 (49.67)	72–250
Errors and omissions (EAO)***	TL	0.43 (0.75)	0–3	1.43 (1.40)	0–6	3.38 (2.20)	0–10	4.67 (2.90)	1–13	10.67 (4.53)	2–21
	DLD	1.24 (1.67)	0–7	3.86 (3.21)	0–12	10.05 (6.95)	0–24	14.57 (9.59)	0–33	30.43 (19.71)	2–81
Ratio-based measures											
NTU per minute	TL	12.24 (4.93)	3–25	11.68 (4.01)	4–23	11.63 (4.03)	5–23	11.51 (3.84)	5–22	11.50 (3.69)	5–22
	DLD	9.38 (4.88)	2–17	10.13 (4.71)	3–18	10.31 (4.12)	4–16	10.53 (3.82)	5–16	10.64 (3.32)	4–15
Words per minute*	TL	45.62 (21.62)	3–104	45.16 (18.94)	19–99	43.69 (17.66)	18–99	43.57 (16.83)	17–96	45.32 (16.56)	17–95
	DLD	31.29 (23.60)	4–83	31.64 (20.56)	4–72	30.61 (18.44)	5–76	32.09 (17.42)	8–76	32.80 (15.82)	7–71
NDW per minute**	TL	29.62 (10.62)	3–47	21.86 (6.37)	11–37	16.64 (4.20)	10–28	14.58 (3.30)	9–23	11.65 (2.29)	7–18
	DLD	20.95 (12.42)	4–48	16.02 (7.96)	3–28	12.05 (4.90)	3–20	11.14 (3.75)	5–17	8.54 (2.48)	4–13
MLU**	TL	4.76 (1.19)	2.5–6.8	5.05 (0.90)	3.7–6.5	5.01 (0.84)	3.9–7.7	5.00 (0.85)	4.0–7.7	5.28 (0.96)	4.0–8.2
	DLD	4.28 (1.25)	1.9–6.3	4.18 (1.30)	1.8–7.1	4.01 (0.91)	2.2–6.3	4.20 (0.87)	2.8–6.8	4.28 (0.88)	2.8–6.5
Percentage of maze words	TL	1.81 (2.89)	0–9	2.07 (1.81)	0–6	2.91 (1.60)	0–6	3.12 (1.53)	1.1–7.3	3.42 (1.74)	1–6.6
	DLD	3.35 (5.71)	0–20	3.16 (4.41)	0–20	3.52 (2.730	0–11.5	3.37 (1.67)	0.5–6.3	3.87 (2.18)	0.3–9
EAO per minute***	TL	0.43 (0.75)	0–3	0.48 (0.47)	0–2	0.50 (0.32)	0–1	0.50 (0.30)	0–1	0.56 (0.23)	0–1
	DLD	1.24 (1.67)	0–7	1.29 (1.07)	0–4	1.44 (0.99)	0–3	1.46 (0.96)	0–3	1.52 (0.99)	0–4
Percent grammatical utterances***	TL	97.12 (4.71)	88–100	95.26 (4.50)	82–100	95.45 (2.91)	90–100	95.71 (2.30)	91–99	94.93 (2.76)	88–99
	DLD	82.25 (23.90)	0–100	84.51 (11.39)	63–100	85.18 (7.94)	73–100	85.67 (7.30)	75–100	85.04 (7.92)	69–98
Subordination index***	TL	1.16 (0.16)	0.89–1.44	1.13 (0.12)	1–1.45	1.10 (0.06)	1–1.21	1.10 (0.05)	1.02–1.24	1.11 (0.06)	1.02–1.23
	DLD	0.95 (0.25)	0–1.33	0.97 (0.2)	0.33–1.25	0.97 (0.11)	0.77–1.13	0.98 (0.11)	0.73–1.13	0.99 (0.09)	0.8–1.11

Note. Significant between-group differences are indicated by asterisks. TL = typical language; DLD = developmental language disorder; MLU = mean length of utterance.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 4. RMANOVA results by group (TL, DLD) and sample length.

Measure		*df*	*f*	*p*	η_p^2
Number of total utterances per minute	Length	1.72	0.20	.79	.005
	Group	1	1.55	.22	.037
	Length × Group	1.72	3.05	.06	.071
Words per minute	Length	1.48	0.38	.62	.009
	Group*	1	5.69	.022	.125
	Length × Group	1.48	0.21	.74	.005
Number of different words per minute	Length***	1.24	77.18	< .001	.659
	Group**	1	9.38	.004	.190
	Length × Group	1.24	2.76	.10	.064
MLU	Length	1.62	1.05	.34	.026
	Group**	1	11.15	.002	.218
	Length × Group	1.62	1.06	.34	.026
Percentage of maze words	Length	1.87	2.05	.14	.049
	Group	1	1.23	.27	.030
	Length × Group	1.87	0.68	.50	.017
Errors and omissions per minute	Length	1.61	0.92	.38	.023
	Group***	1	15.63	< .001	.281
	Length × Group	1.61	0.22	.75	.006
Percent grammatical utterances	Length	1.35	0.11	.98	.003
	Group***	1	26.92	< .001	.402
	Length × Group	1.35	0.81	.41	.020
Subordination index	Length	1.51	0.28	.70	.007
	Group***	1	21.32	< .001	.348
	Length × Group	1.51	1.47	.24	.035

Note. RMANOVA = repeated-measures analysis of variance; TL = typical language; DLD = developmental language disorder; MLU = mean length of utterance.

*$p < .05$. **$p < .01$. ***$p < .001$.

Estimates of Absolute Reliability Varied Across Measures

CVs showed greater variability for the DLD group than the TL group across LSA measures and time cuts.

Table 5. Spearman-Brown coefficients between 1-, 3-, 7-, and 10-min samples and 20-min samples by group (TL, DLD).

Measure	Group	1 min	3 min	7 min	10 min
Number of total utterances	TL	.88	**.95**	**.97**	**.97**
	DLD	.80	**.91**	**.95**	**.97**
Words per minute	TL	.79	**.90**	**.96**	**.96**
	DLD	.75	**.91**	**.95**	**.98**
Number of different words	TL	.71	.88	**.95**	**.96**
	DLD	.80	**.95**	**.97**	**.98**
MLU	TL	–.36	.64	**.93**	**.96**
	DLD	.55	.75	**.91**	**.96**
Percentage of maze words	TL	.51	.19	.85	.87
	DLD	.63	.43	.55	**.90**
Errors and omissions	TL	.57	.50	.63	.70
	DLD	.73	**.91**	**.92**	**.95**
Percent grammatical utterances	TL	.66	.46	.67	.81
	DLD	.53	.74	**.90**	**.94**
Subordination index	TL	.58	.30	.73	**.92**
	DLD	.21	.56	.88	**.96**

Note. Values in bold indicate $\rho \geq .90$. TL = typical language; DLD = developmental language disorder; MLU = mean length of utterance.

Differences in CV values between shorter lengths and the 20-min length for the TL group ranged from 0.01 to 0.08 for NTU, 0.02 to 0.1 for WPM, 0.03 to 0.16 for NDW, 0.0 to 0.07 for MLU, 0.02 to 1.09 for PMW, 0.19 to 1.31 for EAO, 0 to 2 for PGU, and 0.0 to 0.09 for SI. Differences for the DLD group ranged from 0.05 to 0.21 for NTU, 0.06 to 0.27 for WPM, 0.05 to 0.3 for NDW, 0.0 to 0.1 for MLU, 0.06 to 1.15 for PMW, 0.01 to 0.7 for EAO, 0 to 20 for PGU, and 0.02 to 0.17 for SI. Compared to Heilmann et al. (2010), CV values for the TL group in the 1-, 3-, and 7-min time cuts were slightly greater for NTU, WPM, PMW, and EAO; similar for NDW; and lower for MLU. See Supplemental Material S3 for a summary of the results by group and sample length.

B-A plots showed acceptable levels of reliability for nearly all measures at all sample lengths when ±1.96 *SDs* of the mean differences were used to set LOA. This includes every measure at the 1-min time cut for both groups. Lack of differentiation across lengths, groups, and measures calls into question the utility of this procedure for estimating absolute reliability in LSA. It seems overly generous. In contrast, using LOA of ±1.00 for MLU provided differentiation. At the 7-min threshold, the DLD group reached acceptable levels of reliability for MLU, and the TL group reached nearly acceptable reliability (90% within LOA). See Supplemental Materials S4–S5 for a summary of the results by group and sample length.

Group Differences Were Observed in Sample Length Needed for Adequate Reliability

Using a cutoff of $\rho \geq .90$, the TL and DLD groups required 3-min samples to reach acceptable reliability for NTU and WPM, 7-min samples for MLU, and 10-min samples for SI. The TL group required longer sample lengths to reach adequate reliability than the DLD group for NDW, PMW, EAO, and PGU. The DLD group reached acceptable reliability levels for NDW at 3 min, whereas the TL group required 10 min. The DLD group was reliable for EAO at 3 min, PGU at 7 min, and PMW at 10 min, whereas the TL group did not reach adequate reliability levels for these measures.

Significant Group Differences in Reliability Estimates Were Observed on EAO

Potential significant differences between groups in estimates of relative reliability were examined. Fisher Z converts Pearson correlation coefficients to standardized z scores, which can be used to test for significant differences between two correlation coefficients. Fisher Z values were calculated for each measure at each time cut to test for significant group differences in reliability. Significant group differences for Pearson correlations were found for EAO at the 3-min length ($z = -2.8$ $p = .005$, two-tailed), 7-min length ($z = -2.46$ $p = .014$, two-tailed), and 10-min length ($z = -2.6$ $p = .009$, two-tailed). These results indicated that errors within language samples represented a more stable feature of language samples collected on 5- to 7-year-old children with DLD than those collected on their peers with TL skills. All other group differences in reliability coefficients were nonsignificant.

Discussion

Sufficient levels of reliability are needed before clinicians can place much stock in any measures designed to help them diagnose, identify goals, or monitor progress in children with DLD. LSA is relatively unique among available clinical measures because it addresses all these objectives while also providing ecologically valid estimates of children's language skills. Enthusiasm for LSA among clinicians has historically been curbed by the perceived amount of time associated with collecting, transcribing, and analyzing samples. It also appears to be waning. One solution is to collect shorter language samples than what has been conventionally suggested. The use of shorter samples represents an appropriate strategy if shorter samples are as reliable and valid as longer samples. In this study, we examined the effect of sample length, in minutes, on the relative and absolute estimates of reliability of eight widely used LSA measures from children with and without DLD.

Reliability for Children With TL

Estimates of relative reliability for our participants with TL replicated previous findings for some measures and showed differences for others. Alpha values for the TL group at the 3-min length for WPM and NDW replicated the results of Heilmann et al. (2010), suggesting that genre effects (play-based vs. interview or narrative) might exert little influence on the stability of these measures for children in the age range examined here. Pearson correlation coefficients for WPM and NDW at the 3-min length replicated the results of Guo and Eisenberg (2015), whereas correlations for MLU were slightly lower in 3-min samples and similar in 7- and 10-min samples. This replication suggests samples from kindergarten to first-grade children yield similar reliability to those from preschool-age children for these measures.

Obtained estimates of absolute reliability using CVs and B-A plots were more challenging to interpret. Their use has been limited in previous language sample studies. In many cases, what constitutes a clinically meaningful difference for a given LSA measure is unclear. In addition, conventions for suitable cutoff values for CV and LOA for B-A plots for LSA measures have yet to emerge. Compared to Heilmann (2010), CVs in our TL group across 1-, 3-, and 7-min samples were higher for UPM, WPM, and PMW, suggesting play-based samples may result in more variability for these measures compared to interview and narrative samples. CV for NDW, MLU, and EAO was similar to those in Heilmann's study, suggesting similar variability across sampling contexts for these measures in children with TL.

B-A plots with LOA calculated using ± 1.96 *SD*s of mean differences resulted in LOA that were too large to be helpful for determining reliability in 1- and 3-min samples, similar to the findings of Pavelko et al. (2020) in a 25-utterance sample. For example, our analyses for MLU in 1- and 3-min samples, along with the 25-utterance MLU from Pavelko et al. (2020), showed that 95% of the data points fell within the LOA. At face value, this represents a desirable outcome. However, in practice, our estimated LOA (−1.6, 2.1 for a 3-min MLU in the TL group) would allow for differences between estimates of MLU to be as high as 3.7. For example, it would consider a 3-min MLU of 3.0 consistent with a 20-min MLU of 6.7. Allowing discrepancies of that magnitude essentially capsizes any prospect of using the MLU measure to determine the presence or absence of a language disorder or to attribute observed improvements to intervention efforts. However, in the 7- and 10-min samples for MLU, the ± 1.96 *SD*s of LOA were similar to our selected LOA of ± 1.00 for MLU and showed reliability levels similar to estimates calculated for relative reliability. These results suggest that using LOA of ± 1.00 for MLU may be more clinically useful for determining benchmarks for absolute reliability.

Additionally, it is unclear whether results from SALT and SUGAR-based MLU calculations can be meaningfully compared. Briefly, MLU_{sugar} includes several derivational bound morphemes (e.g., *–ful*, *–er*, *–ly*); counts contracted infinitives such as *wanna*, *hafta*, and *gotta* as two morphemes and *gonna* as three morphemes; and includes one-word, incomplete, and abandoned utterances as well as utterances with up to two unintelligible words. This alternative version of MLU does not yet have an established unit of growth anchored to developmental achievements like traditional MLU calculations. In traditional MLU calculations, an increase of around 1.0 tracks with substantive qualitative changes (i.e., stages) reflecting children's emerging proficiencies with morphosyntax, phrase elaborations, and complex syntax (Guo et al., 2018).

Reliability for Children With DLD

This study contributes new information regarding the reliability of LSA measures in children with DLD. A series of RMANOVAs found no Length × Group interactions, suggesting the stability of LSA scores in different sample lengths did not vary as a function of group status. Had significant interaction effects been observed, the implication would be that determining the appropriate sample length would depend on whether children were or were not affected by DLD—limiting their value for diagnostic decisions when affected status is unknown. RMANOVAs revealed significant group differences for six of the eight LSA measures examined: WPM, NDW, MLU, EAO, PGU, and SI. These group differences indicated that these six measures might be more useful than NTU or PMW when using LSA for diagnostic purposes since they captured robust significant differences between children with DLD and those with TL. The large effect size for group differences in PGU ($\eta_p^2 = .402$) indicated that this measure might be particularly advantageous for identifying potential cases of DLD.

Estimates for relative reliability indicated that children with TL and DLD ages 5–7 years produce words, morphemes, and complex sentences at an equally stable rate, whereas children with DLD produce different words and errors at a more stable rate than children with TL. This increased stability is advantageous for using short language samples to diagnose, identify goals, and monitor progress in children with either known or suspected DLD. Our results for MLU replicated findings from Cole et al. (1989) showing Pearson correlations ≥ .9 at 100 utterances and contrast those of Casby (2011) who found Pearson correlations ≥ .9 at 50 utterances. Differences in these findings may be explained by differences in age ranges and participant sample sizes. Our results for PGU replicated findings from Guo et al. (2019), indicating similar reliability estimates for play-based conversational and narrative samples in children with DLD ages 5–7 years.

Estimates of absolute reliability, using CV, showed more variability in individual scores across time cuts for children with DLD than for children with TL. These results also suggest that the higher estimates of relative reliability in the DLD group for EAO and PGU may be partly inflated due to increased variability within the DLD group compared to the TL group. Overall, B-A plots for the DLD group showed larger mean differences and LOA ranges than the TL group, suggesting more variability among the DLD group. B-A plots using LOA of ±1.00 for MLU may be more clinically useful than ±1.96 *SD*s of mean differences for determining absolute reliability. However, an allowable range of ±1.00 for MLU may still be too wide for capturing clinically significant increases. Perhaps, setting LOA to ±0.5 would be more appropriate. We examined the consequences of setting the LOA at ±0.5 and found that 19 of the 21 children in the DLD group had MLU differences within ±0.05, suggesting a 10-min sample may be adequate for measuring progress with MLU for most children with DLD ages 5–7 years (see Supplemental Material S6).

Overall, our results suggest that, for children with DLD in kindergarten and first grade, a sample of around 7 min or 70 utterances in length appears to be comparable to a 20-min or 170-utterance sample for most measures. For MLU, PWM, and SI, samples of 10 min or 100 utterances may be needed to obtain accurate, stable measures.

Limitations

Several limitations need to be considered when evaluating the results of our study. Our study sample reflected the community from which it was drawn and consisted of 5- to 7-year-old monolingual English speakers who were predominately White and non-Hispanic. Our results may not generalize to children of different ages or from communities with different demographic characteristics. Language samples were divided by time to allow direct comparisons with previous studies and to align more closely with how clinicians plan for LSA in their sessions. However, utterance-based divisions may produce more precise reliability estimates as our time-based cuts resulted in an extensive range of utterances for both groups. We compared cuts from within the 20-min samples to the full 20-min samples. This method, consistent with Guo and Eisenberg (2015), allowed us to determine whether reducing the length of a sample would result in approximately the same values for the selected LSA measures. However, this method also resulted in some data overlap, which may have inflated our reliability estimates. Other studies (Heilmann et al., 2010; Pavelko et al., 2020) have examined the reliability of nonoverlapping sections of language samples. To examine the impact of our decision on our obtained estimates, we calculated a follow-up split-half reliability using the first and last 10 min of our samples.

Spearman–Brown correlations ranged from −.17 to .91 for the TL group, with lower correlations for EAO and PGU. Correlations for the DLD group ranged from .47 to .92, with lower values for PMW. (See Supplemental Material S7 for a summary across LSA measures.) The split-half reliability estimates for the DLD group, with the exception of PMW, were comparable to Spearman–Brown reliability estimates for the Clinical Evaluation of Language Fundamentals–Fifth Edition (CELF-5) subtests, which range from .60 to .95 (Wiig et al., 2013). Therefore, our results show that, even when using short samples, LSA has reliability estimates comparable to the CELF-5, making LSA a psychometrically robust option for routine evaluation of children with suspected DLD.

We examined language samples in a play-based context; consequently, our results may not apply to other sampling contexts, including interview, narrative, expository, and persuasive language samples. Even so, the potential influence of genre on our estimates for WPM and NDW may be modest because our results were generally consistent with previous estimates based on other collection approaches. However, other measures, including the SI, may be affected by genre as narrative and expository contexts elicit more complex sentences (Nippold et al., 2005, 2014, 2015), which may lead to increased stability in shorter samples. Additional research is needed to further quantify the effects of genre on estimates of LSA reliability. Finally, the value of our absolute reliability estimates is provisional for two reasons. First, there is little precedent to determine whether our values were high relative to other studies. Second, there are no established thresholds yet for determining clinically appropriate differences across most LSA measures, with the possible exception of MLU calculated using conventional methods (Brown, 1973).

Clinical Implications

Our results provide additional support for the feasibility of language samples in clinical and school settings with shorter samples and computer-assisted analysis. Heilmann et al. (2010) estimated that a 50-utterance sample could be collected, transcribed, and analyzed using SALT software in around 30 min. This estimate aligns with our time spent transcribing and analyzing the samples for this study. Finestack et al. (2020) estimated 35 min for a 50-utterance sample using Computerized Language Analysis (MacWhinney, 2000), and Pavelko and Owens (2019) estimated 20 min for collection, transcription, and SUGAR analysis. In comparison, administration time for the CELF-5 core language subtests averages 34 min for students ages 5;0–8;11 with "variable" additional time for scoring (Wiig et al., 2013). Thus, assuming other factors like individual clinician proficiency are equal, LSA may take less time than is needed to administer, score, and evaluate the CELF-5 Core Language Index, the most commonly used omnibus standardized test (Caesar & Kohler, 2009).

Future Directions

Future studies should expand the examination of reliability associated with short language samples collected on children with DLD to different ages and different groups (e.g., children who are English language learners, multilingual, or speakers of different dialects of English). Future work should also examine relative and absolute reliability associated with LSA measures based on other sampling contexts, including interview, narrative, expository, and persuasive samples. Additionally, clinically meaningful differences need to be established for individual LSA measures before absolute reliability estimates can be interpreted. In this study, we applied 1.0 and 0.5 for conventional MLU and suggested that levels of agreement within 0.5 should be used based on our analyses. Future investigations may provide data that challenge our suggestion. Finally, our research and that of others have only addressed one aspect of the potential value of short language samples. Besides being reliable, clinical measures also need to demonstrate validity to be useful. Our data show robust group differences for six LSA measures (WPM, NDW, MLU, EAO, PGU, and SI) that demonstrated adequate reliability levels. Future studies should examine more closely the diagnostic accuracy of these six measures for identifying cases of DLD.

Acknowledgments

This work was supported by the National Institute on Deafness and Other Communication Disorders Grant R01DC011023 awarded to Sean M. Redmond. Some of the data in this article were presented at the 2021 Symposium on Research in Child Language Disorders at the University of Wisconsin-Madison. The authors are grateful for the children who participated in this study and their families. Appreciation is extended to Andrea Ash and Kristin Pruett for their comments and suggestions on an earlier draft. They also appreciate the contributions to data collection of Andrea Ash and the following graduate research assistants: Anne Downhour, Austa Feller, Kirsten Hannig Russell, Jacie Meldrum, Cloey Roper, Clara Warrick, and Tina Whitehead.

References

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

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., & Donlan, C. (2005). The role of syntax in encoding and recall of pictorial narratives: Evidence from specific language impairment. *British Journal of Developmental Psychology, 23*(1), 25–46. https://doi.org/10.1348/026151004X20685

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

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet, 327*(8476), 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research, 8*(2), 135–160. https://doi.org/10.1177/096228029900800204

Brown, R. (1973). *A first language: The early stages*. Harvard University Press. https://doi.org/10.4159/harvard.9780674732469

Bruton, A., Conway, J. H., & Holgate, S. T. (2000). Reliability: What is it, and how is it measured? *Physiotherapy, 86*(2), 94–99. https://doi.org/10.1016/S0031-9406(05)61211-4

Caesar, L. G., & Kohler, P. D. (2009). Tools clinicians use. *Communication Disorders Quarterly, 30*(4), 226–236. https://doi.org/10.1177/1525740108326334

Casby, M. W. (2011). An examination of the relationship of sample size and mean length of utterance for children with developmental language impairment. *Child Language Teaching and Therapy, 27*(3), 286–293. https://doi.org/10.1177/0265659010394387

Cole, K. N., Mills, P. E., & Dale, P. S. (1989). Examination of test–retest and split-half reliability for measures derived from language samples of young handicapped children. *Language, Speech, and Hearing Services in Schools, 20*(3), 259–268. https://doi.org/10.1044/0161-1461.2003.259

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. https://doi.org/10.1007/BF02310555

Darley, F. L., & Moll, K. L. (1960). Reliability of language measures and size of language sample. *Journal of Speech and Hearing Research, 3*(2), 166–173. https://doi.org/10.1044/jshr.0302.166

DeVellis, R. F. (2012). *Scale development: Theory and applications* (3rd ed.). Sage Publications.

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

Eadie, P., Conway, L., Hallenstein, B., Mensah, F., McKean, C., & Reilly, S. (2018). Quality of life in children with developmental language disorder. *International Journal of Language & Communication Disorders, 53*(4), 799–810. https://doi.org/10.1111/1460-6984.12385

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

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

Eisinga, R., te Grotenhuis, M., & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman Brown? *International Journal of Public Health, 58*(4), 637–642. https://doi.org/10.1007/s00038-012-0416-3

Finestack, L. H., Rohwer, B., Hilliard, L., & Abbeduto, L. (2020). Using computerized language analysis to evaluate grammatical skills. *Language, Speech, and Hearing Services in Schools, 51*(2), 184–204. https://doi.org/10.1044/2019_LSHSS-19-00032

Finestack, L. H., & Satterlund, K. E. (2018). Current practice of child grammar intervention: A survey of speech-language pathologists. *American Journal of Speech-Language Pathology, 27*(4), 1329–1351. https://doi.org/10.1044/2018_AJSLP-17-0168

Gavin, W. J., & Giles, L. (1996). Sample size effects on temporal reliability of language sample measures of preschool children. *Journal of Speech and Hearing Research, 39*(6), 1258–1262. https://doi.org/10.1044/jshr.3906.1258

Guo, L. Y., & Eisenberg, S. (2015). Sample length affects the reliability of language sample measures in 3-year-olds: Evidence from parent-elicited conversational samples. *Language, Speech, and Hearing Services in Schools, 46*(2), 141–153. https://doi.org/10.1044/2015_LSHSS-14-0052

Guo, L. Y., Eisenberg, S., Ratner, N. B., & MacWhinney, B. (2018). Is putting SUGAR (Sampling Utterances of Grammatical Analysis Revised) into language sample analysis a good thing? A response to Pavelko and Owens (2017). *Language, Speech, and Hearing Services in Schools, 49*(3), 622–627. https://doi.org/10.1044/2018_LSHSS-17-0084

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., Schneider, P., & Harrison, W. (2021). Clausal density between ages 4 and 9 years for the Edmonton narrative norms instrument: Reference data and psychometric properties. *Language, Speech, and Hearing Services in Schools, 52*(1), 354–368. https://doi.org/10.1044/2020_LSHSS-20-00043

Hadley, P. A. (1998). Language sampling protocols for eliciting text-level discourse. *Language, Speech, and Hearing Services in Schools, 29*(3), 132–147. https://doi.org/10.1044/0161-1461.2903.132

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

Heilmann, J., Tucci, A., Plante, E., & Miller, J. F. (2020). Assessing functional language in school-aged children using language sample analysis. *Perspectives of the ASHA Special Interest Groups, 5*(3), 622–636. https://doi.org/10.1044/2020_PERSP-19-00079

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

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

IBM Corp. (2019). *IBM SPSS Statistics for Windows, Version 26.0*. https://www.ibm.com/products/spss-statistics

Kaplan, R. M., & Saccuzzo, D. P. (2018). *Psychological testing: Principles, applications, and issues*. Cengage Learning.

Kemp, K., & Klee, T. (1997). Clinical language sampling practices: Results of a survey of speech-language pathologists in the United States. *Child Language Teaching and Therapy, 13*(2), 161–176. https://doi.org/10.1177/026565909701300204

Kohnert, K. (2010). Bilingual children with primary language impairment: Issues, evidence and implications for clinical actions. *Journal of Communication Disorders, 43*(6), 456–473. https://doi.org/10.1016/j.jcomdis.2010.02.002

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Krouwer, J. S. (2008). Why Bland–Altman plots should use X, not $(Y+X)/2$ when X is a reference method. *Statistics in Medicine, 27*(5), 778–780. https://doi.org/10.1002/sim.3086

Lai, S. A., & Schwanenflugel, P. J. (2016). Validating the use of D for measuring lexical diversity in low-income kindergarten children. *Language, Speech, and Hearing Services in Schools, 47*(3), 225–235. https://doi.org/10.1044/2016_LSHSS-15-0028

Langbecker, D., Snoswell, C. L., Smith, A. C., Verboom, J., & Caffery, L. J. (2020). Long-term effects of childhood speech and language disorders: A scoping review. *South African Journal of Childhood Education, 10*(1), 1–13. https://doi.org/10.4102/sajce.v10i1.801

Lee, L. L. (1974). *Developmental sentence analysis: A grammatical assessment procedure for speech and language clinicians*. Northwestern University Press.

Leonard, L. B. (2014). *Children with specific language impairment*. MIT Press. https://doi.org/10.7551/mitpress/9152.001.0001

MacLachlan, B. G., & Chapman, R. S. (1988). Communication breakdowns in normal and language learning-disabled children's conversation and narration. *Journal of Speech and Hearing Disorders, 53*(1), 2–7. https://doi.org/10.1044/jshd.5301.02

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk* (3rd ed.). Erlbaum.

Marinellie, S. A. (2004). Complex syntax used by school-age children with specific language impairment (SLI) in child–adult conversation. *Journal of Communication Disorders, 37*(6), 517–533. https://doi.org/10.1016/j.jcomdis.2004.03.005

McCauley, R. J., & Swisher, L. (1984). Psychometric review of language and articulation tests for preschool children. *Journal of Speech and Hearing Disorders, 49*(1), 34–42. https://doi.org/10.1044/jshd.4901.34

MedCalc. (2021). *MedCalc Statistical Software version 20.014*. https://Medcalc.org

Miller, J., & Iglesias, A. (2012). *Systematic analysis of language transcripts (SALT), research version*. SALT Software LLC. https://www.saltsoftware.com/products/software

Miller, J. F. (1981). *Assessing language production in children: Experimental procedures*. University Park Press.

Miller, J. F. (1996). The search for the phenotype of disordered language performance. In M. L. Rice (Ed.), *Toward a genetics of language* (pp. 297−314). Erlbaum.

Miller, J. F., Andriacchi, K., & Nockerts, A. (2019). *Assessing language production using SALT software: A clinician's guide to language sample analysis* (3rd ed.). SALT Software LLC.

Miller, J. F., & Chapman, R. S. (1981). The relation between age and mean length of utterance in morphemes. *Journal of Speech and Hearing Research, 24*(2), 154–161. https://doi.org/10.1044/jshr.2402.154

Naglieri, J. A. (1997). *Naglieri Nonverbal Ability Test*. The Psychological Corporation.

Nippold, M. A., Frantz-Kaspar, M. W., Cramond, P. M., Kirk, C., Hayward-Mayhew, C., & MacKinnon, M. (2014). Conversational and narrative speaking in adolescents: Examining the use of complex syntax. *Journal of Speech, Language, and Hearing Research, 57*(3), 876–886. https://doi.org/10.1044/1092-4388(2013/13-0097)

Nippold, M. A., Frantz-Kaspar, M. W., Cramond, P. M., Kirk, C., Hayward-Mayhew, C., & MacKinnon, M. (2015). Critical thinking about fables: Examining language production and comprehension in adolescents. *Journal of Speech, Language, and Hearing Research, 58*(2), 325–335. https://doi.org/10.1044/2015_JSLHR-L-14-0129

Nippold, M. A., Hesketh, L. J., Duthie, J. K., & Mansfield, T. C. (2005). Conversational versus expository discourse. *Journal of Speech, Language, and Hearing Research, 48*(5), 1048–1064. https://doi.org/10.1044/1092-4388(2005/073)

Nippold, M. A., Mansfield, T. C., & Billow, J. L. (2007). Peer conflict explanations in children, adolescents, and adults: Examining the development of complex syntax. *American Journal of Speech-Language Pathology, 16*(2), 179–188. https://doi.org/10.1044/1058-0360(2007/022)

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

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

Owens, R. E., Jr. (2016). *Language development: An introduction*. Pearson.

Paul, R., Norbury, C. F., & Gosse, C. (2017). *Language disorders from infancy through adolescence* (5th ed.). Elsevier Health Sciences.

Pavelko, S. L., & Owens, R. E., Jr. (2017). Sampling Utterances and Grammatical Analysis Revised (SUGAR): New normative values for language sample analysis measures. *Language, Speech, and Hearing Services in Schools, 48*(3), 197–215. https://doi.org/10.1044/2017_LSHSS-17-0022

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

Pavelko, S. L., Price, L. R., & Owens, R. E., Jr. (2020). Revisiting reliability: Using Sampling Utterances and Grammatical Analysis Revised (SUGAR) to compare 25- and 50-utterance language samples. *Language, Speech, and Hearing Services in Schools, 51*(3), 778–794. https://doi.org/10.1044/2020_LSHSS-19-00026

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

Pieretti, R. A., & Roseberry-McKibbin, C. (2016). Assessment and intervention for English language learners with primary language Impairment. *Communication Disorders Quarterly, 37*(2), 117–128. https://doi.org/10.1177/1525740114566652

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)

Ratner, N. B., & MacWhinney, B. (2016). Your laptop to the rescue: Using the Child Language Data Exchange System archive and CLAN utilities to improve child language sample analysis. *Seminars in Speech and Language, 37*(2), 74–84. https://doi.org/10.1055/s-0036-1580742

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

Retherford, K. S., Schreiber, L. R., & Jarzynski, L. R. (2019). *Guide to analysis of language transcripts*. Pro-Ed.

Rice, M. L., Redmond, S. M., & Hoffman, L. (2006). Mean length of utterance in children with specific language impairment and in younger control children shows concurrent validity and stable and parallel growth trajectories. *Journal of Speech, Language, and Hearing Research, 49*(4), 793–808. https://doi.org/10.1044/1092-4388(2006/056)

Rice, M. L., Smolik, F., Perpich, D., Thompson, T., Rytting, N., & Blossom, M. (2010). Mean length of utterance levels in 6-month intervals for children 3 to 9 years with and without language impairments. *Journal of Speech, Language, and Hearing Research, 53*(2), 333–349. https://doi.org/10.1044/1092-4388(2009/08-0183)

Roseberry-McKibbin, C. (2007). Assessment and intervention guidelines for service delivery to low-SES children. *SIG 16 Perspectives on School-Based Issues, 8*(3), 4–9. https://doi.org/10.1044/sbi8.3.4

SALT Software. (2021). *Self-paced online courses*. https://www.saltsoftware.com/training/self-paced-online-training

Scott, C. M., & Windsor, J. (2000). General language performance measures in spoken and written narrative and expository discourse of school-age children with language learning disabilities. *Journal of Speech, Language, and Hearing Research, 43*(2), 324–339. https://doi.org/10.1044/jslhr.4302.324

Semel, E. M., Wiig, E. H., & Secord, W. A. (2003). *Clinical evaluation of language fundamentals (CELF-4)*. The Psychological Corporation.

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: A practical guide to their development and use*. Oxford University Press. https://doi.org/10.1093/med/9780199685219.001.0001

Thordardottir, E. T., & Weismer, S. E. (2002). Content mazes and filled pauses in narrative language samples of children with specific language impairment. *Brain and Cognition, 48*(2–3), 587–592. https://doi.org/10.1006/brcg.2001.1422

Tilstra, J., & McMaster, K. (2007). Productivity, fluency, and grammaticality measures from narratives: Potential indicators of language proficiency? *Communication Disorders Quarterly, 29*(1), 43–53. https://doi.org/10.1177/1525740108314866

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

Tyack, D., & Gottsleben, R. (1974). *Language sampling, analysis and training: A handbook for teachers and clinicians*. Consulting Psychological Press.

Von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *Bulletin of the World Health Organization, 85,* 867–872. https://doi.org/10.2471/blt.07.045120

Watkins, R. V., Kelly, D. J., Harbers, H. M., & Hollis, W. (1995). Measuring children's lexical diversity: Differentiating typical and impaired language learners. *Journal of Speech and Hearing Research, 38*(6), 1349–1355. https://doi.org/10.1044/jshr.3806.1349

Westerveld, M. F., & Claessen, M. (2014). Clinician survey of language sampling practices in Australia. *International Journal of Speech-Language Pathology, 16*(3), 242–249. https://doi.org/10.3109/17549507.2013.871336

Wiig, E. H., Semel, E. M., & Secord, W. (2013). *CELF 5: Clinical Evaluation of Language Fundamentals*. Pearson.

Wilder, A., & Redmond, S. (2021). Spontaneous productions of infinitive clauses by English-speaking children with and without specific language impairment. *Clinical Linguistics & Phonetics, 35*(1), 43–64. https://doi.org/10.1080/02699206.2020.1740323