LSHSS

Research Article

Sample Length Affects the Reliability of Language Sample Measures in 3-Year-Olds: Evidence From Parent-Elicited Conversational Samples

Ling-Yu Guo[a] and Sarita Eisenberg[b]

Purpose: The goal of this study was to investigate the extent to which sample length affected the reliability of total number of words (TNW), number of different words (NDW), and mean length of C-units in morphemes (MLCUm) in parent-elicited conversational samples for 3-year-olds.
Method: Participants were sixty 3-year-olds. A 22-min language sample was collected from each child during free play with the parent in the laboratory. Samples of 1, 3, 7, and 10 min were extracted from the 22-min samples. TNW, NDW, and MLCUm were computed from each shorter sample and the 22-min sample. TNW and NDW were adjusted by number of minutes for comparisons. The differences and correlations between each shorter sample cut and the 22-min sample on MLCUm and adjusted TNW and NDW were computed.
Results: The shorter samples and the 22-min samples significantly differed in adjusted TNW and NDW, but not in MLCUm. TNW reached an acceptable reliability level (i.e., $r = .90$) in 7-min samples. NDW and MLCUm approached the acceptable reliability level ($rs = .88$) in 7-min samples and reached it in 10-min samples.
Conclusion: For conversational language samples with similar collection procedures, samples of 7 to 10 min are desirable for calculating TNW, NDW, and MLCUm in 3-year-olds.

For decades, language sample analysis (LSA) has been considered an effective tool for differentiating children with and without language impairment and for identifying treatment goals (Bedore & Leonard, 1998; Dunn, Flax, Sliwinski, & Aram, 1996; Horton-Ikard, 2010; Lahey, 1988; Lee, 1974; Oetting et al., 2010; Paul & Norbury, 2012; Stockman, Guillory, Seibert, & Boult, 2013; Watkins, Kelly, Harbers, & Hollis, 1995). Despite its usefulness, LSA is still not incorporated into the assessment battery by some clinicians (Kemp & Klee, 1997; Westerveld & Claessen, 2014), partly because of the time required for collecting, transcribing, and analyzing the data. To reach a balance between the reliability of LSA measures and the efficiency of clinical work, language samples of 50 to 100 utterances, which take about 10 to 15 min to elicit, have been recommended (Miller, Andriacchi, & Nockerts, 2011; Paul & Norbury, 2012). There is some evidence suggesting that this sample size can generate reliable LSA measures for the purpose of screening and/or diagnosis, but the requisite sample length may vary from measure to measure (Cole, Mills, & Dale, 1989; Darley & Moll, 1960; McCarthy, 1975; Minifie, Darley, & Sherman, 1963; Rondal & DeFays, 1978). However, even with 50 to 100 utterances, LSA can still place a significant strain on speech-language pathologists (SLPs) who work with children with language impairment, given that the median caseload of a full-time school-based SLP is 47 children per week, with a range up to 240 children (American Speech-Language-Hearing Association, 2012).

To promote the use of LSA, several studies have investigated whether samples shorter than 50 utterances can generate reliable LSA measures (Brorson & Dewey, 2005; Casby, 2011; Heilmann, Nockerts, & Miller, 2010). Although these studies indicated that shorter samples were as reliable as longer samples for computing LSA measures, they were limited in their sampling procedure and the method for quantifying and/or interpreting the reliability data (see later). Thus, more evidence is needed before clinicians can confidently use language samples with fewer than 50 utterances in the assessment procedure. The present study examined the extent to which sample length affected the reliability of

[a]University at Buffalo, NY
[b]Montclair State University, Montclair, NJ
Correspondence to Ling-Yu Guo: lingyugu@buffalo.edu

Editor: Marilyn Nippold
Associate Editor: LaVae Hoffman
Received May 6, 2014
Revision received July 17, 2014
Accepted December 18, 2014
DOI: 10.1044/2015_LSHSS-14-0052

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

two global lexical measures (i.e., total number of words and number of different words) and mean length of utterances in morphemes (MLUm) in parent-elicited conversational samples for 3-year-olds. These measures were chosen because they have been previously shown to reflect developmental changes in children (Brown, 1973; Miller, 1981; Templin, 1957) and to differentiate children with and without language impairment (Bedore & Leonard, 1998; Hewitt, Hammer, Yont, & Tomblin, 2005; Rice et al., 2010; Watkins et al., 1995).

In what follows, we first review studies of sample length that support the use of conversational language samples of at least 50 utterances for LSA measures based on reliability data. We then review two studies that support the use of shorter samples for LSA measures, one using utterance cuts (Casby, 2011) and the other using time cuts (Heilmann et al., 2010). Previous studies have tended to focus on reliability of language sample measures because, in psychometrics, reliability reflects how consistent or repeatable a measure is (Bruton, Conway, & Holgate, 2000). Reliability is also a necessary, although not sufficient, condition for validity of the measure, that is, how well the measure assesses the construct that it intends to test (McCauley, 2001). Reliability of a measure can be documented across test items (i.e., internal consistency reliability) and over time (i.e., test–retest reliability; Hopkins, Stanley, & Hopkins, 1990). Furthermore, the degree of reliability of a measure can be evaluated via absolute reliability and relative reliability (Bruton et al., 2000). Absolute reliability refers to "the degree to which repeated measurements vary for individuals, i.e., the less they vary, the higher the reliability" (Bruton et al., 2000, p. 95). Relative reliability refers to the degree to which individuals maintain their relative position among others over repeated measurement. Ideally, convergent evidence from both absolute and relative reliability is preferred in order to say that a particular measure (e.g., mean length of utterance [MLU] computed from 20 utterances) is reliable. For instance, if the value of MLU computed from 20 utterances does not differ significantly from MLU based on 100 utterances (absolute reliability) and is highly correlated with MLU computed from 100 utterances (relative reliability), this would mean that MLU computed from 20 utterances is as reliable as MLU from 100 utterances.

Most of the studies we review used the magnitude of correlation coefficients as one of the indices to evaluate how reliable a measure was at a given sample length. However, different criteria have been used to interpret the magnitude of correlation coefficients, which makes it difficult to compare across studies. Following previous studies (Bogue, DeThorne, & Schaefer, 2014; Gavin & Giles, 1996; McCauley & Swisher, 1984), we consider a correlation coefficient of .90 or higher to be acceptable in interpreting the reliability data. This criterion has been adopted because a correlation coefficient of .90 means that no more than 20% (i.e., $1 - 0.90^2 = 0.19$) of the variability in children's performance on a given measure can be attributed to measurement error.

Studies Supporting the Use of at Least 50 Utterances for LSA

Darley and Moll (1960) examined how sample length affected the internal consistency reliability of mean length of response (MLR, a measure similar to mean length of utterances in words) in 150 typical 5-year-olds using different utterance cuts within the same language samples. A 50-utterance conversational sample was elicited from each child using a picture-description task administered by the examiner at the child's home. The mean MLR for the first five, 10, 15, 20, 25, and 35 utterances did not differ much from the mean MLR for the total 50 utterances, ranging from 5.54 to 5.87 words. However, variability between children was greater for the shorter samples, ranging from a standard deviation of 3.2 words for the five-utterance samples to 1.8 words for the 50 utterance samples. Darley and Moll also reported estimated reliability for MLR based on different numbers of utterances. Based on this estimation, a sample size of 50 utterances yielded a reliability of .85, with sample sizes of at least 80 utterances needed to obtain reliability at the .90 level.

Rondal and DeFays (1978) reported similar results for MLUm based on 42 typically developing children between the ages of 1;8 (years;months) and 2;8. A 1-h conversational sample was collected from each child during free play with his/her mother at home. MLUm did not differ significantly for the first 25, 50, 75, 100, 125, 150, 175, and 200 utterances, but variability between children was higher for shorter samples. Rondal and DeFays reported an internal consistency reliability of .80 for MLUm based on 50-utterance samples but did not report the sample size at which reliability reached the .90 level.

Although Cole et al. (1989) did not specifically examine the effect of sample size on the reliability of LSA measures, they examined both test–retest and split-half (i.e., a type of internal consistency) reliability for MLUm for 10 children between the ages of 4;4 and 6;8 with mild to moderate developmental delays. Two separate, 100-utterance conversational language samples were collected from each child during free play with the examiner in the laboratory, with approximately 2 weeks between the two samples. Test–retest reliability was .92. Split-half reliability between the first and second 50 utterances was .95. These results indicated that MLUm may demonstrate acceptable test–retest and split-half reliabilities when based on 100-utterance samples.

Gavin and Giles (1996) further investigated the effect of sample length on the test–retest reliability of total number of words (TNW), number of different words (NDW), and MLUm for 20 children between the ages of 2;7 and 3;10 with typical language. Two 20-min conversational language samples were collected from each child during free play with his/her principal caregiver in the laboratory, approximately three to 14 days apart. The language samples were segmented by time (i.e., 12 and 20 min) and by number of utterances (i.e., 25, 50, 75, . . . 175 utterances) to explore the effect of sample length. For the time-based cuts,

none of the target measures showed acceptable test–retest reliability for the 20-min samples based on the .90 criterion (r = .49 for TNW, .72 for NDW, and .77 for MLUm). For the utterance-based cuts, TNW, NDW, and MLUm did not show acceptable test–retest reliability until sample length reached 175 utterances (r = .92 for TNW, .93 for NDW, and .97 for MLUm). However, it should be noted that only eight of the 20 children produced 175 or more utterances.

Taken together, with the exception of Gavin and Giles (1996), these studies support a conclusion that at least 80 to 100 utterances are needed for MLU to demonstrate acceptable reliability within a given sample (Cole et al., 1989; Darley & Moll, 1960) or between two samples over time (Cole et al., 1989). However, none of these studies indicate what sample size was required for total number of words or number of different words in order to demonstrate acceptable reliabilities within a given sample, although 50 utterances has been recommended in the literature as the sample size for computing these lexical measures (e.g., Miller, 1981).

Studies Supporting the Use of Fewer Than 50 Utterances

Using archival data, Casby (2011) investigated whether shorter samples were as reliable as longer samples for measuring MLUm via conversations collected from 10 children between the ages of 3;0 and 11;8 with language impairment. The total sample contained 100 to 150 utterances from each child produced during free play with the examiner in the laboratory. The MLUm values were computed for the first 10 and 20 utterances, the middle 10 and 20 utterances, the last 10 and 20 utterances, every second utterance (quasirandom 50 utterances), every fifth utterance (quasirandom 20 utterances), every 10th utterance (quasirandom 10 utterances), and the total sample. The MLUm values of the shorter samples and the total samples did not differ significantly. The correlations between the shorter samples and the total sample were .86 for the first 10, .52 for the first 20, .77 for the middle 10, .75 for the middle 20, .80 for the last 10, .92 for the last 20, .89 for the quasirandom 10, .93 for the quasirandom 20, and .94 for the quasirandom 50 utterances. MLU for all but one (i.e., the first 20 utterances) of the shorter sample cuts significantly correlated with MLU for the total sample. Casby concluded that "one can reliably and efficiently determine MLU on much smaller language samples than that typically recommended" (p. 286). However, inspecting the correlation data of the six continuous sample cuts, we find that only the last 20 utterances correlated with the total samples at the acceptable level (r = .92). It is also unclear why the middle 20 utterances did not reach a level of reliability (r = .75) similar to that of the last 20 utterances. In addition, although quasirandom 20 utterances correlated with the total samples at the acceptable level, MLU is typically not computed with every fifth utterance in the clinical setting. Thus, whether 10 or 20 utterances can generate as reliable MLUm values as 50 to 100 utterances remains an open question.

Instead of using utterance cuts, a recent study by Heilmann et al. (2010) explored whether 1- and 3-min samples were as consistent as 7-min samples in measuring words per minute (WPM, total number of complete words in the main body and in mazes per minute), number of different words per minute (NDW/m), and MLUm. They investigated both conversational and narrative samples, but we will focus only on the conversational samples for the purposes of this article. The children were divided into younger (ages between 2;8 and 5;11) and older (ages between 6;0 and 13;3) groups because Heilmann and colleagues speculated that younger children would be more variable in their production and therefore might need longer samples to obtain reliable language sample measures. The first 11 min of conversational samples, which involved the examiner interviewing or playing with the child, were used for the analysis. To avoid a warm-up effect, the samples were first divided into eleven 1-min segments, and these segments were then randomly selected for the 1-min, 3-min, and 7-min sample cuts for each child. For instance, a 3-min sample could come from minutes 1, 7, and 10 or from minutes 4, 5, and 11. On average, the total number of utterances, segmented based on the rules for C-units, was 12.5 for 1-min samples, 36.0 for 3-min samples, and 84.0 for 7-min samples in conversation. None of the target measures (i.e., WPM, NDW/m, or MLUm) differed significantly between any sample cuts. In addition, all but one of the shorter language samples demonstrated correlations with the 7-min samples at levels between .70 and .86 for all of the target measures. Heilmann et al. (2010) concluded that short samples (e.g., 1- and 3-minute samples yielding 12 to 36 C-units) may be appropriate to document children's global lexical skills and MLUm.

However, in addition to the use of random minute segments from the language samples, there were two methodological limitations in this study. First, the 1- and 3-min samples were compared to a standard sample length of 7 min to examine consistency between shorter and longer samples. Given that the 7-min samples were relatively short, there is a need to reevaluate the reliability of the 1- and 3-min samples by comparing them with longer samples that demonstrate acceptable reliability (e.g., samples longer than 20 min). Second, the nonsignificant difference in NDW/m between sample lengths may have resulted from the way NDW/m was computed. To illustrate, suppose that a child said only "They like it" in minute 4, only "We like it" in minute 5, and only "I like it" in minute 11, and that these three minutes were randomly picked to compute NDW/m. Because the three segments were considered separately when NDW was computed for each minute in Heilmann et al. (2010), NDW/m for this 3-min sample would be calculated as 3 ([3 different words + 3 different words + 3 different words]/3 min), which potentially might inflate the child's NDW. However, if each of the three segments were considered together when NDW was computed, NDW/m would be 1.67 (5 different words/3 min).

In summary, although Casby (2011) and Heilmann et al. (2010) provide some evidence to support the use of short language samples for assessment, these studies were

limited in a few ways: a lower-than-acceptable level for reliability coefficients, the use of discontinuous utterances for calculating the language measures, the lack of an established standard for comparison, and the overestimation of target measures. Thus, more evidence would be needed to verify that shorter language samples (e.g., 20-utterance or 3-min samples) are as reliable as longer samples (e.g., 100-utterance or 20-min samples) for documenting language skills such as NDW and MLU.

The Present Study

To address the unresolved issues regarding the results of Casby (2011) and Heilmann et al. (2010) and to explore what sample length would be needed in order to obtain acceptable internal consistency reliability for global LSA measures, this study evaluated the reliability of three global LSA measures that were calculated from 1-min, 3-min, 7-min, and 10-min conversational language samples of 3-year-old children. To this end, we compared the language measures (i.e., total number of words per minute, number of different words per minute, and mean length of C-units in morphemes) generated from these shorter samples with the same measures generated from 22-min samples (see the Method section). We used time-based cuts, instead of utterance-based cuts, to define sample length in order to compare our results with those of Heilmann et al. (2010). We included 1-, 3-, and 7-min samples because these time cuts were adopted in that study. The 10-min samples were also included because this is a common transcript cut used by clinicians (Miller et al., 2011; Paul & Norbury, 2012). The 3-, 7-, and 10-min samples were consecutive segments from larger 22-min samples. The 22-min samples were chosen as the standard sample size to compare with the shorter language samples because samples of this length typically can generate 150 or more utterances, which is close to or beyond the level suggested by previous studies (Cole et al., 1989; Darley & Moll, 1960; Gavin & Giles, 1996; Rondal & DeFays, 1978).

Although not systematically evaluated in this study, sampling contexts (e.g., partners, settings, materials/activities) of conversational samples may lead to variations in language sample measures, which in turn could potentially affect the reliability of the measures. For instance, Bornstein, Haynes, Painter, and Genevro (2000) found that typically developing 2-year-olds produced more utterances and more different words during 8 min of free play with their mothers than with the examiners, although the settings (i.e., the home or the laboratory) did not lead to differences in these measures. Hoff (2010) further found that typically developing children between the ages of 1;5 and 2;2 produced more different words in conversation with their mothers during book-reading activities than during free play. However, book-reading activities, as Miller (1981) indicates, tend to elicit routines that are not spontaneous from young children. To collect spontaneous samples that better reflect children's language skills, we chose to collect parent-elicited, instead of examiner-elicited, conversational language samples from children during free play in the laboratory. We preferred the laboratory to children's homes as the setting because those settings do not show differences in the LSA measures (Bornstein et al., 2000), and the laboratory setting allowed us to guard against potential interferences (e.g., siblings, noise from the street).

Specifically, we asked two questions. First, would total number of words per minute (TNW/m), number of different words per minute (NDW/m), and mean length of C-units in morphemes (MLCUm) generated from the 1-, 3-, 7-, and 10-min sample cuts differ significantly from those generated from the 22-min samples in 3-year-olds during free play with their parents in the laboratory? This question aimed to determine the absolute reliability (Bruton et al., 2000) of the target measures computed from shorter samples. Second, to what extent were TNW/m, NDW/m, and MLCUm generated from the 1-, 3-, 7-, 10-min sample cuts correlated with those generated from the 22-min samples in 3-year-olds? This question aimed to determine the relative reliability (Bruton et al., 2000) of the target measures computed from shorter samples. If the target measures generated from the shorter samples did not differ from, and were correlated at acceptable (or close-to-acceptable) levels with, those generated from the 22-min samples, this would suggest that shorter samples were as reliable as the 22-min samples for 3-year-olds during free play with their parents in the laboratory.

In a pilot study (Guo & Eisenberg, 2013), however, we found that NDW/m within the same language sample systematically decreased over time, possibly because of how NDW was calculated. Repeating certain words is necessary in conversations. For instance, speakers have to repeat certain closed-class words (e.g., pronouns, prepositions, determiners) across utterances to produce grammatical utterances and certain open-class words to maintain the discourse topic (Liles, 1985). Because the calculation of NDW depends on prior word productions (i.e., what has been said earlier), repeating words across utterances inevitably leads to a decrease in NDW/m with increasing sample length. Thus, the differences for NDW/m between the shorter and the 22-min samples cannot be taken as direct evidence for or against the absolute reliability of NDW in shorter samples. For this reason, although we reported the differences in NDW/m between the shorter and the 22-min samples, the results were not counted as evidence to evaluate the reliability of NDW in shorter samples. That is, we used only the correlation data (i.e., relative reliability) to evaluate the reliability of NDW/m in shorter conversational samples. In contrast, both absolute and relative reliability were used to evaluate the reliability of TNW and MLCUm in shorter samples.

Method

Participants

Sixty children (29 girls, 31 boys) between the ages of 3;0 and 3;11 (M = 3;6, SD = 0;4) participated in the current

study. They were recruited through flyers and online announcements from the Buffalo (New York) area (24 children) and the Montclair (New Jersey) area (36 children). The 36 children recruited from the Montclair area were also participants in prior studies by Eisenberg and colleagues (Eisenberg & Guo, 2013; Guo & Eisenberg, 2014) that investigated the diagnostic accuracy of percent grammatical utterances and tense usage in identifying 3-year-olds with and without language impairment. Approval for the current research was granted by institutional review boards of the University at Buffalo and Montclair State University. We focused on 3-year-olds because these children were going through a period in which their language skills were changing rapidly and were likely to show considerable variability in language production (Gavin & Giles, 1996; Heilmann et al., 2010; Tommerdahl & Kilpatrick, 2013). Thus, longer language samples might be needed for these children in order to obtain reliable measures. Although Heilmann et al. (2010) indicated that younger children, like older children, produced reliable language sample measures in short samples, further evidence is needed because of the wide age range (i.e., 2;8–5;11) for the younger group in their study. To address this need, we limited the study sample in the present investigation to 3-year-olds.

To document children's language ability, we administered the Structured Photographic Expressive Language Test–Preschool: Second Edition (SPELT-P:2; Dawson et al., 2004). A standard score of 87, which yielded a sensitivity of 90.6% and a specificity of 100% (Greenslade, Plante, & Vance, 2009), was used as the cutoff to determine whether a child had a language impairment. It should be noted that this cutoff was generated based on children between the ages of 4;0 and 5;8, instead of 3-year-olds. To the best of our knowledge, no empirical studies exist for the sensitivity or specificity of the SPELT-P:2 with 3-year-olds. We adopted this cutoff here simply for research purposes. The mean standard score on the SPELT-P:2 for all children was 101.02 (SD = 14.97, range = 65–133). Among the 60 children, 46 of them (22 girls, 24 boys) scored at or above the cutoff (standard score M = 94.29, SD = 4.08, range = 87–133), while 14 of them (seven girls, seven boys) scored below the cutoff (standard score M = 80.85, SD = 6.79, range = 65–86). However, children were included regardless of their language status (that is, whether they had typical language or language impairment) so that the participants in the current study had a wide range of language ability, which resembles the clinical setting (Peña, Spaulding, & Plante, 2006). The research assistants who collected, transcribed, and coded the samples were unaware of the children's language status.

All children passed the Articulation subtest of the Fluharty Preschool Speech and Language Screening Test–Second Edition (Fluharty-2; Fluharty, 2001). In addition, all children passed a hearing screening at 25 dB for frequencies of 500, 1000, 2000, and 4000 Hz. All children except for one child who could not be tested demonstrated cognitive ability within the typical range (standard score M = 108.38, SD = 15.58, range = 85–139) as measured by the Odd-Item-Out task of the Reynolds Intellectual Screening Test (RIST; Reynolds & Kamphaus, 2003), a task that evaluates children's nonverbal intelligence. The child who could not be tested was not compliant in pointing to the pictures with his hands or fingers as responses in the RIST. In order to follow the standardized procedure in the RIST, we did not use alternative means for the child to respond. However, he was still included in the current study, because there were no parent concerns about his cognitive development and he was able to complete the remaining experimental protocol (i.e., talking about the pictures on the SPELT-P:2 and Fluharty-2 and playing with the toys with his parent).

In addition, the parents completed a questionnaire developed by the second author of this article to provide information about their children's language background and developmental history, the parents' educational levels, and the family's ethnic and racial background. On the section about language background, the parents indicated whether their children spoke Standard American English, African American English, or other dialects of English, and whether their children also spoke a language other than English. All of the children were reported to be monolingual, native speakers of Standard American English. There was no history or current concern about cognitive, psychobehavioral, neurological, or physical development for any of the children. Socioeconomic status was determined based on maternal education, with 20% having a postcollege degree, 67% having a college degree, and 13% having a high school diploma. Based on the self-report from the questionnaire, the racial and ethnic distribution was 83% Caucasian, 8% African-American, 7% Hispanic, and 2% Asian.

Materials and Procedure

A 30-min conversational language sample was collected from each child during free play with the parent in a child-friendly test room within the laboratory. Five sets of age-appropriate toys were used in the play, such as farm animals, dolls and furniture, and vehicles. We used fixed toy sets to keep the contexts of language samples consistent across children. Before the play began, the parent was instructed by the examiner (i.e., the first author or student research assistants) to follow the child's lead and play with the child as at home. The parent then picked one set of toys to start the free play. The remaining four sets of toys were provided by the examiner one at a time in a random order every 6 min. All of the toys were placed on the floor for the child and the parent. When a new set of toys was provided, the previous sets of toys remained available on the floor to allow the child and the parent to combine different sets of toys. The entire session was video- and audio-recorded for transcription.

Language Sample Transcription and Processing

The language samples were transcribed by trained research assistants based on the conventions of Systematic

Analysis of Language Transcripts (SALT; Miller & Iglesias, 2010). Utterances that could not be fully transcribed after the research assistants listened to them three times were marked as unintelligible. To be consistent with the SALT reference database, utterances were segmented into communication units (C-units). A C-unit is typically an independent clause plus all of its dependent clauses (Loban, 1976). Nonclausal utterances that expressed complete thoughts (e.g., "Good morning, Mom") were also counted as C-units. Only intelligible, complete, and spontaneous C-units were included for analysis. Elliptical responses to questions (e.g., P. "Who wants to go?" C. "I do"; P. "What do you want to do?" C. "Put it in there"; P = parent, C = child) were counted as C-units because they represented complete thoughts, and were included in the analysis (Nippold et al., 2013).

After the transcription was completed, we marked the language samples minute by minute. Because the sixth minute for each toy set involved the change of toys and the time for changing the toys varied slightly across children, we excluded the sixth minute for each toy set, to keep the length of the language samples consistent. In addition, the change of toys created noise, which made it difficult to transcribe some of the utterances that occurred during this period. These nontranscribable utterances were thus marked as unintelligible based on the SALT transcription conventions. Excluding the sixth minute of each toy set thus avoided the use of minute segments that had relatively more utterances transcribed as unintelligible due to environmental noise than other minute segments. After this exclusion, there were 25 min remaining of the 30-min language sample. To avoid a potential warm-up effect, the first three minutes of the remaining 25-min samples were further excluded. Thus, the standard language samples for each child were 22 min (30 min − 5 min − 3 min = 22 min) in length.

Transcription Accuracy

To check the accuracy of language sample transcription, we used a consensus procedure that was adapted from Shriberg, Kwiatkowski, and Hoffman (1984). Each sample was first transcribed by one research assistant. Then a second research assistant checked the transcription by listening to the recorded language sample while reading the initial transcription. Transcription for the entire sample was then rechecked by the first or second author. Discrepancies were discussed and agreement was obtained on all transcripts. Utterances that could not be resolved were excluded from the analysis. The same consensus procedure was followed for C-unit segmentation.

Computation and Statistical Methods

All of the sample cuts started at the same time point, the beginning of the 22-min sample. Consecutive, rather than random or intermittent, minute segments of language samples were used because this is consistent with common clinical practice. The 1-min samples were extracted from minute 1, the 3-min samples were extracted from minutes 1 through 3, the 7-min samples were extracted from minutes 1 through 7, and the 10-min samples were extracted from minutes 1 through 10 of the 22-min sample for each child.

We computed TNW, NDW, and MLCUm for each sample cut (i.e., 1-, 3-, 7-, 10-, and 22-min samples) by using the SALT program. The total number of C-units for each sample cut was also computed to document the children's productivity in conversations, although it was not the measure of interest in the current study.

The raw frequency of total number of C-units, TNW, and NDW for each sample cut was generated from the SALT program. To investigate whether the measures in shorter samples (i.e., 1-, 3-, 7-, and 10-min samples) were as reliable as those in the standard samples (i.e., 22-min samples), the raw frequencies of TNW and NDW were adjusted by number of minutes for comparisons between sample cuts. That is, we computed TNW/m and NDW/m for the 1-, 3-, 7-, 10-, and 22-min samples. It should be noted that TNW/m in the present study was different from WPM in Heilmann et al. (2010). TNW/m in the present study included only intelligible words in the main body of the language sample, while WPM in Heilmann et al. included intelligible words in both the main body and mazes (e.g., false starts, revisions, repetitions). We did not include words within mazes because those words are typically excluded for analysis for the computation of TNW (e.g., Miller & Iglesias, 2010; Templin, 1957). In addition, given that mazes presumably result from problems with lexical retrieval and/or sentence formulation (Rispoli, 2003) during language production, exclusion of mazed words could avoid overestimating the lexical skill of a child who produces abundant mazes. MLCUm was computed by dividing the total number of morphemes by the total number of C-units within each sample cut, which was also available from the SALT program. Unlike the other measures, MLCUm was not adjusted by minutes, because it is inherently a ratio measure.

To document the reliability of TNW, NDW, and MLCUm in shorter samples, we first examined whether TNW/m, NDW/m, and MLCUm from shorter samples were significantly different from those from the standard (i.e., 22-min) samples via preplanned one-way repeated measures analyses of variance (ANOVAs) as the measure of absolute reliability (Bruton et al., 2000). We used preplanned ANOVAs because we focused only on comparing each of the shorter samples to the standard sample, not on comparing among the shorter samples. Given that four comparisons (i.e., the 1-, 3-, 7-, 10-min samples vs. the 22-min sample) were conducted for each dependent variable, a Bonferroni correction was adopted to control for type I error (Field, 2009), which yielded a minimum significance level of .0125. As we mentioned earlier, the differences between the shorter and the standard samples in NDW/m cannot be taken as direct evidence for or against the absolute reliability of NDW in the shorter samples. Thus, even though the differences in NDW/m between sample lengths were still computed, the results were not counted

as evidence for or against the reliability of NDW in the shorter samples.

Next we examined the extent to which target measures from shorter samples were correlated with those from the standard sample using Pearson correlations as the measure of relative reliability (Bruton et al., 2000). Following previous studies (Bogue et al., 2014; Gavin & Giles, 1996; McCauley & Swisher, 1984), we interpreted a correlation coefficient of .90 or higher as acceptable.

Results

Given that this study included 29 girls and 31 boys, we first examined whether there was a gender effect on the target measures in the standard samples (i.e., the 22-min samples). One-way ANOVAs indicated that girls and boys did not differ in TNW, NDW, or MLCUm in the standard samples, $F(1, 58) < 0.69$, $ps > .40$. Thus, the data from girls and boys were combined together in our analyses.

Table 1 presents the raw frequency of total number of C-units, TNW, and NDW across sample lengths. As expected, the raw frequency of these measures increased as the length of language sample increased. The adjusted frequency of these target measures (i.e., TNW/m and NDW/m) and MLCUm are also listed in Table 1. Preplanned one-way repeated measures ANOVAs indicated that TNW/m was significantly larger in 3-min samples than in 22-min samples, $F(1, 59) = 11.90$, $p = .001$, $\eta_p^2 = .168$. There were no other significant differences in TNW/m between shorter samples and the standard 22-min sample, $Fs < 3.44$, $ps > .069$, $\eta_p^2 < .055$ (for details of the F values, see Table 2). Consistent with our pilot study (Guo & Eisenberg, 2013), NDW/m was significantly larger for each of the shorter samples (i.e., 1-, 3-, 7-, and 10-min samples) than for the standard sample, $Fs > 164.31$, $ps < .001$, $\eta_p^2 > .736$, meaning that children produced a higher number of different words per minute in shorter samples than in the 22-min sample. In contrast, MLCUm did not differ between any of the shorter samples and the 22-min sample, $Fs < 1.15$, $ps > .29$, $\eta_p^2 < .019$ (for details of the F values, see Table 2).

Table 3 displays the correlation coefficients between each of the shorter samples and the 22-min sample on each measure. It should be noted that the correlation coefficients for a given measure at a given sample length were identical when the measure was computed by raw frequency and adjusted by number of minutes. For instance, the correlation coefficients between the 7- and 22-min samples in TNW and in TNW/m were both .92. This is because adjusting the measures by number of minutes does not change the relative ranking between children.

Following previous studies (Bogue et al., 2014; Gavin & Giles, 1996; McCauley & Swisher, 1984), we interpreted a correlation coefficient of .90 or higher as acceptable. The target measures (i.e., TNW/m, NDW/m, and MLCUm) generated from 1- and 3-min samples did not correlate with any of the measures from the 22-min samples at the acceptable level, although the correlations were all significant, $ps < .05$ (two-tailed). TNW/m reached an acceptable correlation level for 7-min samples. NDW/m and MLCUm were close to the acceptable level ($r = .88$ for NDW/m and for MLCUm) for 7-min samples, and both reached acceptable correlation levels for 10-min samples.

Discussion

This study examined the extent to which sample length affected the reliability of TNW, NDW, and MLCUm for conversational samples in 3-year-old children. We defined sample length in number of minutes. Regarding our first question about absolute reliability, most of the shorter samples did not differ from the 22-min sample in TNW/m or MLCUm. NDW/m, however, was larger in shorter samples than in the 22-min sample, partly due to how NDW was calculated. Regarding our second question, about relative reliability, TNW/m, NDW/m, and MLCUm all reached an acceptable level (i.e., .90 or higher) for the 10-min samples. However, correlations for NDW/m and MLCUm were close to the acceptable level for the 7-min samples. Correlation coefficients for all of the measures in the current study were at or below .54 for the 1-min samples and at or below .83 for the 3-min samples, which are similar correlation levels to those reported by Heilmann et al. (2010). We separately discuss the results for the lexical measures and for utterance length.

Lexical Measures

In this study, we found that TNW/m was relatively more stable than NDW/m in shorter samples. TNW/m in the shorter samples, except for the 3-min sample, did not differ significantly from that in the 22-min sample. That is, TNW/m in shorter samples was generally comparable to that in the standard sample. The significant difference between the 3-min and the standard samples may indicate that TNW/m could still be somewhat variable when it is generated from a short sample (e.g., 3-min samples), even though it is relatively more stable than NDW/m. In contrast, NDW/m decreased as sample length increased. Both TNW/m and NDW/m demonstrated acceptable reliability when the sample length reached 10 min. Taken together, these results suggest that a minimum sample of 10 min (approximately 91 C-units; see Table 1) would be desirable for calculating TNW and NDW for 3-year-olds in samples collected during parent–child free play.

The current findings for NDW/m were not compatible with the study of Heilmann et al. (2010), which reported that NDW/m was similar for 1-, 3-, and 7-min conversational samples. The discrepancy may have resulted from how NDW/m was calculated in Heilmann et al. (2010). The 1-, 3-, and 7-min samples were randomly picked by minute from the total 11-min sample, and each minute was considered separately when number of different words was computed, potentially leading to an overestimation of the children's lexical skills in the 3- and 7-min samples.

As in previous studies (Owen & Leonard, 2002; Watkins et al., 1995), when NDW was calculated in raw

Table 1. Mean (standard deviation) and range of language sample measures by sample length (N = 60).

Measure	1 min		3 min		7 min		10 min		22 min	
	M (*SD*)	Range	*M* (*SD*)	Range	*M* (*SD*)	Range	*M* (*SD*)	Range	*M* (*SD*)	Range
Frequency-based measures										
Total number of C-units	9.15 (3.69)	3–16	28.83 (8.71)	13–51	63.28 (17.34)	32–107	90.90 (25.80)	40–152	192.57 (54.62)	96–344
Total number of words	29.96 (15.21)	3–83	94.63 (35.42)	36–179	204.63 (72.50)	76–358	295.63 (107.60)	126–536	630.50 (228.12)	228–1,226
Number of different words	20.62 (8.45)	3–36	50.20 (14.09)	23–87	84.07 (21.65)	38–134	107.43 (27.60)	56–162	173.67 (39.53)	102–279
Ratio-based measures										
Total number of words per minute	29.96 (15.21)	3.00–83.00	31.54 (11.81)	12.00–59.67	29.24 (10.36)	10.86–51.14	29.56 (10.76)	12.60–53.60	28.66 (10.37)	10.36–55.73
Number of different words per minute	20.62 (8.45)	3.00–36.00	16.73 (4.69)	7.67–29.00	12.01 (3.09)	5.43–19.14	10.74 (2.76)	5.60–16.20	7.89 (1.80)	4.64–12.68
Mean length of C-units in morphemes	3.43 (1.01)	1.00–6.10	3.55 (0.81)	2.05–5.83	3.50 (0.77)	2.05–5.69	3.51 (0.76)	2.12–6.00	3.55 (0.71)	2.04–5.81

Table 2. *F* values for comparisons between shorter samples and the standard sample.

Measure	1 min vs. 22 min	3 min vs. 22 min	7 min vs. 22 min	10 min vs. 22 min
Total number of words per minute	0.61	11.90*	1.12	3.44
Number of different words per minute	164.33*	394.56*	330.70*	299.28[a]
Mean length of C-units in morphemes	0.89	0.01	1.02	1.15

[a] *F* is significant at the 0.0125 level (i.e., 0.05/4 = 0.0125 due to Bonferroni corrections).

frequency, we found that NDW increased with sample length. However, when NDW was adjusted by number of minutes as in the current study, NDW/m decreased with sample length, possibly because children exhausted their active vocabulary over time (Owen & Leonard, 2002; Richards, 1987). This was in spite of the fact that the collection procedure of introducing different toy sets throughout the session may have encouraged children to use different vocabulary. A related explanation is that because the calculation of NDW depends on prior productions (i.e., what was said earlier in the language sample), repeating words across utterances inevitably leads to a decrease of NDW/m with increasing sample length. Thus, one can argue that the decrease of NDW/m with sample length in the current study may simply reflect the nature of language production and cannot be used as evidence to suggest that NDW/m is unreliable in shorter samples. To better reflect the reliability of NDW in shorter samples, different methods of adjusting for sample length, other than adjusting the samples by minutes, could be a worthwhile research pursuit (e.g., DeThorne, Deater-Deckard, Mahurin-Smith, Coletto, & Petrill, 2011; McKee, Malvern, & Richards, 2000).

Given that NDW/m decreased with sample length in the present study, an ensuing question is whether 10-min conversational samples are appropriate for documenting lexical skills in 3-year-olds. Recall that the reliability of a measure can be evaluated via absolute reliability (e.g., the degree to which NDW/m of different sample lengths varies for individual children) and relative reliability (e.g., the degree to which individual children maintain their position relative to others over different sample lengths; Bruton et al., 2000). Because samples of 10 min demonstrated a correlation with the standard sample at the acceptable level for NDW/m, a child who produced a relatively low NDW/m for a 10-min sample was also likely to show the same trend for a 22-min sample, even though the absolute NDW/m for that child would differ in the 10- and 22-min samples. Thus, we suggest that a conversational sample of 10 min can still be appropriate for computing NDW for 3-year-olds. However, when NDW is computed from 10-min conversational samples, the clinician will have to use a cutoff criterion that is generated from 10-min samples with similar collection procedures to make clinical decisions, because the frequency of NDW varies with sample length.

Utterance Length

This study showed that MLCUm for the shorter (i.e., 1-, 3-, 7-, and 10-min) samples did not differ significantly from MLCUm for the standard (i.e., 22-min) sample for 3-year-olds, which is consistent with earlier studies (Brorson & Dewey, 2005; Casby, 2011; Darley & Moll, 1960; Heilmann et al., 2010; Rondal and DeFays, 1978). However, the shorter samples did not show acceptable correlations with the standard sample for MLCUm until the sample length reached 10 min, or approximately 91 C-units. This finding is compatible with those of Darley and Moll (1960) and Cole et al. (1989), which collectively suggested that 80 to 100 utterances were required in order to generate acceptably reliable values for utterance length. Taken together, the current findings suggest that a conversational sample of 10 min, or approximately 91 C-units, would be desirable for calculating MLUm.

Heilmann et al. (2010) concluded that samples of 1 and 3 min could be used for calculating MLUm. We respectfully disagree with this conclusion. Although the 1- and 3-min samples did not significantly differ from the standard sample for MLCUm (i.e., these short samples showed absolute reliability; Bruton et al., 2000) in the current study, the low level of reliability relative to a 22-min standard

Table 3. Correlation coefficients between shorter samples and 22-min samples by sample length and language sample measures.

Measure	1 min	3 min	7 min	10 min
Frequency-based measures				
Total number of words	.54**	.83**	.92**	.94**
Number of different words	.51**	.79**	.88**	.93**
Ratio-based measures				
Total number of words per minute	.54**	.83**	.92**	.94**
Number of different words per minute	.51**	.79**	.88**	.93**
Mean length of C-units in morphemes	.38*	.74**	.88**	.93**

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

suggests that these shorter sample lengths would not be appropriate for measuring MLCUm. Rather, substantial measurement error could occur when MLUm is measured from these shorter sample lengths. Consider, for instance, the data for a 1-min sample. The correlation for MLCUm between the 1- and 22-min samples for children was .38. This means that about 86% (i.e., $1 - .38^2$) of the variability in children's MLCUm from the 1-min sample can be attributed to measurement error (Gavin & Giles, 1996). Thus, although it is tempting to conclude that shorter samples of 1 or 3 min could generate reliable MLCUm based on the absolute reliability data (e.g., Brorson & Dewey, 2005), the relative reliability data do not support this conclusion.

Clinical Implications

In clinical practice, some attention has been placed on verifying the reliability and validity of standardized tests (e.g., Bogue et al., 2014; Gray, Plante, Vance, & Henrichsen, 1999; Greenslade et al., 2009; Hutchinson, 1996; McCauley & Swisher, 1984; Pearson, Jackson, & Wu, 2014; Perona, Plante, & Vance, 2005; Plante & Vance, 1994, 1995; Restrepo et al., 2006; Spaulding, Plante, & Farinella, 2006; Ukrainetz & Blomquist, 2002). However, the evidence for reliability and validity of LSA remains relatively sparse (Eisenberg, Fersko, & Lundgren, 2001). Some of the existing evidence is even contradictory. For instance, some studies have suggested that fewer than 50 utterances could reliably measure children's MLUm (e.g., Casby, 2011; Heilmann et al., 2010), while other studies have concluded otherwise (e.g., Cole et al., 1989; Gavin & Giles, 1996). Consequently, clinicians might use LSA without sufficient evidence of reliability or validity to justify their selection of particular LSA measures or particular sample lengths. As an initial step to address this clinical issue, the current study contributes empirical evidence regarding internal consistency reliability of TNW, NDW, and MLUm from conversational samples of varying lengths to guide clinicians in designing language sample collection procedures.

Traditionally, language samples with 50 to 100 utterances, which take about 10 to 15 min of recording, have been recommended for clinicians (Miller et al., 2011; Paul & Norbury, 2012). To the best of our knowledge, the current study might be the first to provide empirical evidence for the use of 10-min conversational language samples in measuring TNW, NDW, and MLUm for 3-year-olds. However, because different measures may need different sample sizes in order to obtain reliable counts (Cole et al., 1989; Darley & Moll, 1960; Heilmann et al., 2010), the use of 10-min samples can only be considered a general guideline for the same measures collected with the same procedures. Sample length may need to be customized for other measures, such as the responsiveness or assertiveness of children (Fey, 1986) and usage of tense and agreement morphemes (Gladfelter & Leonard, 2013; Tommerdahl & Kilpatrick, 2013). To make the current findings generalizable to 3-year-olds in clinical practice, clinicians may want to collect a somewhat longer sample (e.g., 13 min) from parent-elicited free play using at least two sets of toys and start to transcribe from the fourth minute to avoid a potential warm-up effect.

It should be noted that even though we suggest that a minimum sample of 10 min is desirable in order to reliably measure 3-year-old children's TNW, NDW, and MLCUm, we are not claiming that a 10-min conversational sample would be sufficient for all 3-year-old children, because some children are relatively reticent and may produce limited numbers of utterances within 10 min. For reticent children, clinicians may need to use number of utterances, instead of number of minutes, as a guideline for sample length. How many utterances, then, are desirable when the clinician wishes to compute 3-year-old children's TNW, NDW, and MLCUm at the same time from a conversational sample obtained during parent-elicited free play? The current study supports using a sample size of approximately 91 C-units, which, again, is close to the recommendation of 100 utterances made in previous studies (Paul & Norbury, 2012).

We used a correlation coefficient of .90 as the benchmark for evaluating the relative reliability of TNW, NDW, and MLUm from shorter samples. While this decision was made based on previous studies (Bogue et al., 2014; Gavin & Giles, 1996; McCauley & Swisher, 1984), we also want to point out that small quantitative differences in statistics may not necessarily mean significant distinctions in clinical work. For instance, NDW/m and MLCUm from the 7-min conversational samples both correlated with those from the standard sample at a level of .88 (see Table 3), which was just slightly below the benchmark level. Does this mean that NDW/m and MLUm from 7-min samples are clinically unreliable? We do not have a clear answer for this question. As is the case for setting confidence intervals for test interpretation (McCauley, 2001), clinicians will vary in the level of correlation that they consider desirable for interpreting LSA measures and, accordingly, will vary in decisions about sample length. If one believes that an incremental difference in correlation coefficients (i.e., .88 vs. .90) does not translate into clinically significant distinctions, collecting 7-min conversational samples (or approximately 63 utterances) to compute TNW, NDW, and MLCUm might be even more feasible in the clinical setting.

Limitations and Future Directions

In this study, we examined the effect of sample length on the internal consistency reliability of TNW, NDW, and MLUm in 3-year-olds. Although we found that 10-min parent-elicited conversational samples during free play generated reliable values for these measures in 3-year-olds, the results may not be generalizable to older children or other sampling contexts. This is because younger children tend to be variable in language production due to the rapid change of language development and thus may need longer samples than other children to obtain reliable language sample measures (Heilmann et al., 2010). It is possible that reliable results may be obtainable for older children with samples that are shorter than 10 min. Future studies are needed to examine this possibility. In addition, given that variables such

as speaking tasks (e.g., conversational, narrative, expository), communication partners (e.g., parent, examiner), and activities (e.g., book reading, free play, interview) can have an impact on LSA measures (Bornstein et al., 2000; Evans & Craig, 1992; Hoff, 2010; Nippold et al., 2013; Nippold, Hesketh, Duthie, & Mansfield, 2005; Southwood & Russell, 2004), the current findings may apply only to conversational samples collected with similar procedures.

A related issue about sampling contexts is how we introduced the toys to children. Recall that the examiner provided one set of toys for the parents and children every 6 min. This decision was made because we wanted to maintain children's interest during the 30-min free play. This procedure, however, might have interrupted the flow of parent–child interaction and promoted use of different vocabulary items when a new toy set was provided. These variables made it difficult to compare the current study with previous studies that involved children in free play with the same toys throughout the session (e.g., Cole et al., 1989; Gavin & Giles, 1996). Another consideration is that we instructed parents to follow their child's lead and play with their child as they do at home. Although this is a common guideline for language sampling (e.g., Miller & Iglesias, 2010; Paul & Norbury, 2012), such instructions might be contradictory for parents who have a directive style in interacting with their children. It may be preferable just to instruct parents to follow their child's lead. Studies are needed that systematically manipulate how toys are introduced to children (e.g., one set every 6 min, five sets all at once in the beginning) and how parents are instructed, to determine whether these variables should also be controlled in the language sampling.

In addition, the current study examined the effect of sample length on language sample measures based on amount of time instead of number of utterances or C-units. This was done in order to compare the results to those of the previous study by Heilmann et al. (2010). However, young children vary in their amount of talking, and some children may require longer sampling times to achieve the requisite number of utterances. Further studies that directly look at reliability as a function of number of utterances are needed to examine this issue, although the current study suggests that a minimum of 91 C-units are desirable in order to obtain reliable measures for TNW, NDW, and MLCUm in 3-year-olds.

Last, the current study only evaluated how sample length affected reliability of language sample measures, and did not investigate validity. Recall that reliability is a necessary, although not sufficient, condition for validity of a measure. A reliable measure could potentially have low validity (McCauley, 2001). For instance, although MLCUm in 10-min language samples may be reliable, whether it can accurately identify 3-year-olds with and without language impairment remains unknown. Future studies that determine how sample length might affect diagnostic accuracy (e.g., how accurately a measure can identify a child with language impairment or typical language at different sample lengths) are needed. Together, reliability and validity data of language measures at varying lengths would enable clinicians to determine the length of language samples in the assessment process.

Conclusions

Sample length affects the reliability of measuring TNW, NDW, and MLCUm for children in parent-elicited conversational samples. Although shorter samples may seem more feasible for clinical practice, samples that are too short (e.g., 10-utterance or 1-min samples) may generate unreliable language sample measures, which in turn may lead to misdiagnosis. Based on the current findings, we suggest that if parent-elicited free play is used as the context to collect conversational samples, a minimum sample of 10 min (or 91 C-units) is desirable for computing global lexical measures and MLCUm for 3-year-old children. However, given that the correlations between the 7-min and standard samples for TNW, NDW, and MLCUm were above or close to the acceptable level, parent-elicited conversational samples of 7 min (or 63 C-units) could also be used to generate these measures for 3-year-olds, especially in clinical settings. In addition, conversational samples of varying lengths (e.g., 3 min) might provide information for other clinical purposes, such as use of specific linguistic forms, functions, or dialect features. Thus, the clinician may choose to use language samples of varying lengths, depending on the purpose or measure.

Acknowledgments

This project was supported by a Language Learning Research Grant (from Blackwell Publishing) awarded to Ling-Yu Guo and by National Institute on Deafness and Other Communication Disorders (NIDCD) Grant R21DC009218 awarded to Sarita Eisenberg. The content is solely the responsibility of the authors and does not necessarily represent the official views of Blackwell Publishing, NIDCD, or the National Institutes of Health. We are grateful to the children who participated as well as to their parents who allowed them to participate, and to the research assistants who collected and transcribed the samples. We also thank Yow-wu Bill Wu for statistical consultation and Amy Briggs and Sanjana Nair for data analysis. Portions of this study were presented as part of a poster session at the 2013 Convention of the American Speech-Language-Hearing Association in Chicago, Illinois.

References

American Speech-Language-Hearing Association. (2012). *ASHA 2012 schools survey: SLP caseload characteristics report*. Rockville, MD: Author.

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,* 1185–1192.

Bogue, E., DeThorne, L., & Schaefer, B. (2014). A psychometric analysis of childhood vocabulary tests. *Contemporary Issues in Communication Science and Disorders, 41,* 55–69. doi:1092-5171/14/4101-0055

Bornstein, M. H., Haynes, O. M., Painter, K. M., & Genevro, J. L. (2000). Child language with mother and with stranger at home

and in the laboratory: A methodological study. *Journal of Child Language, 27,* 407–420.
Brorson, K., & Dewey, C. (2005). Effect of language sample size on MLUw. *Hearsay, 17,* 46–56.
Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.
Bruton, A., Conway, J. H., & Holgate, S. T. (2000). Reliability: What is it, and how is it measured? *Physiotherapy, 86,* 94–99.
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,* 286–293. doi: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,* 259–268.
Darley, F. L., & Moll, K. L. (1960). Reliability of language measures and size of language sample. *Journal of Speech and Hearing Research, 3,* 166–173.
Dawson, J., Stout, C., Eyer, J., Tattersall, P., Fonkalsrud, J., & Croley, K. (2004). *Structured Photographic Expressive Language Test–Preschool: Second Edition*. DeKalb, IL: Janelle.
DeThorne, L. S., Deater-Deckard, K., Mahurin-Smith, J., Coletto, M.-K., & Petrill, S. A. (2011). Volubility as a mediator in the associations between conversational language measures and child temperament. *International Journal of Language & Communication Disorders, 46,* 700–713. doi:10.1111/j.1460-6984.2011.00034.x
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,* 643–654.
Eisenberg, S. L., Fersko, T. M., & Lundgren, C. (2001). The use of MLU for identifying language impairment in preschool children: A review. *American Journal of Speech-Language Pathology, 10,* 323–342. doi:10.1044/1058-0360(2001/028)
Eisenberg, S. L., & Guo, L.-Y. (2013). Differentiating children with and without language impairment based on grammaticality. *Language, Speech, and Hearing Services in Schools, 44,* 20–31. doi:10.1044/0161-1461(2012/11-0089)
Evans, J. L., & Craig, H. K. (1992). Language sample collection and analysis: Interview compared to freeplay assessment contexts. *Journal of Speech and Hearing Research, 35,* 343–353. doi:10.1044/jshr.3502.343
Fey, M. E. (1986). *Language intervention with young children*. San Diego, CA: College-Hill Press.
Field, A. (2009). *Discovering statistics using SPSS* (4th ed.). Thousand Oaks, CA: Sage.
Fluharty, N. (2001). *Fluharty Preschool Speech and Language Screening Test–Second Edition*. Austin, TX: Pro-Ed.
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,* 1258–1262.
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,* 542–552. doi:10.1044/1092-4388(2012/12-0100)
Gray, S., Plante, E., Vance, R., & Henrichsen, M. (1999). The diagnostic accuracy of four vocabulary tests administered to preschool-age children. *Language, Speech, and Hearing Services in Schools, 30,* 196–206. doi:10.1044/0161-1461.3002.196
Greenslade, K., Plante, E., & Vance, R. (2009). The diagnostic accuracy and construction validity of the Structured Photographic Expressive Language Test–Preschool: Second Edition. *Language, Speech, and Hearing Services in Schools, 40,* 150–160. doi:10.1044/0161-1461(2008/07-0049)
Guo, L.-Y., & Eisenberg, S. (2013, November). *Are shorter samples as good as longer samples*? Poster presented at the 2013 Convention of the American Speech-Language-Hearing Association, Chicago, IL.
Guo, L.-Y., & Eisenberg, S. (2014). The diagnostic accuracy of two tense measures for identifying 3-year-olds with language impairment. *American Journal of Speech-Language Pathology, 23*(2), 203–212. doi:10.1044/2013_AJSLP-13-0007
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,* 393–404. doi:10.1044/0161-1461(2009/09-0023)
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,* 197–213. doi:10.1016/j.jcomdis.2004.10.002
Hoff, E. (2010). Context effects on young children's language use: The influence of conversational setting and partner. *First Language, 30,* 461–472. doi:10.1177/0142723710370525
Hopkins, K. D., Stanley, J., & Hopkins, B. (1990). *Educational and psychological measurement and evaluation* (7th ed.). Upper Saddle River, NJ: Pearson Education.
Horton-Ikard, R. (2010). Language sample analysis with children who speak non-mainstream dialects of English. *Perspectives on Language Learning and Education, 17*(1), 16–23. doi:10.1044/lle17.1.16
Hutchinson, T. A. (1996). What to look for in the technical manual: Twenty questions for users. *Language, Speech, and Hearing Services in Schools, 27,* 109–121. doi:10.1044/0161-1461.2702.109
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,* 161–176. doi:10.1177/026565909701300204
Lahey, M. (1988). *Language disorders and language development*. Needham, MA: Macmillan.
Lee, L. (1974). *Developmental sentence analysis: A grammatical assessment procedure for speech and language clinicians*. Evanston, IL: Northwestern University Press.
Liles, B. Z. (1985). Cohesion in the narratives of normal and language-disordered children. *Journal of Speech and Hearing Research, 28,* 123–133. doi:10.1044/jshr.2801.123
Loban, W. (1976). *Language development: Kindergarten through grade twelve*. Urbana, IL: National Council of Teachers of English.
McCarthy, D. A. (1975). *The language development of the preschool child*. Westport, CT: Greenwood Press.
McCauley, R. J. (2001). *Assessment of language disorders in children*. Mahwah, NJ: Erlbaum.
McCauley, R. J., & Swisher, L. (1984). Psychometric review of language and articulation tests for preschool children. *Journal of Speech and Hearing Disorders, 49,* 34–42.
McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing, 15*(3), 323–338. doi:10.1093/llc/15.3.323
Miller, J. F. (1981). *Assessing language production in children: Experimental procedures*. Baltimore, MD: University Park Press.

Miller, J., Andriacchi, K., & Nockerts, A. (2011). *Assessing language production using SALT software: A clinician's guide to language sample analysis*. Madison, WI: Language Analysis Laboratory, Waisman Center, University of Wisconsin–Madison.

Miller, J., & Iglesias, A. (2010). Systematic Analysis of Language Transcripts (Version 10). Madison, WI: Language Analysis Laboratory, Waisman Center, University of Wisconsin–Madison.

Minifie, F. D., Darley, F. L., & Sherman, D. (1963). Temporal reliability of seven language measures. *Journal of Speech and Hearing Research, 6,* 139–148.

Nippold, M. A., Frantz-Kaspar, M. W., Cramond, P. M., Kirk, C., Hayward-Mayhew, C., & MacKinnon, M. (2013). Conversational and narrative speaking in adolescents: Examining the use of complex syntax. *Journal of Speech, Language, and Hearing Research, 57,* 1–11. doi:10.1044/1092-4388(2013/13-0097)

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

Oetting, J. B., Newkirk, B. L., Hartfield, L. R., Wynn, C. G., Pruitt, S. L., & Garrity, A. W. (2010). Index of productive syntax for children who speak African American English. *Language, Speech, and Hearing Services in Schools, 41,* 328–339. doi:10.1044/0161-1461(2009/08-0077)

Owen, A. J., & Leonard, L. B. (2002). Lexical diversity in the spontaneous speech of children with specific language impairment: Application of D. *Journal of Speech, Language, and Hearing Research, 45,* 927–937. doi:10.1044/1092-4388(2002/075)

Paul, R., & Norbury, C. F. (2012). *Language disorders from infancy through adolescence: Listening, speaking, reading, writing, and communicating* (4th ed.). St. Louis, MO: Elsevier.

Pearson, B. Z., Jackson, J. E., & Wu, H. (2014). Seeking a valid gold standard for an innovative, dialect-neutral language test. *Journal of Speech, Language, and Hearing Research, 57,* 495–508. doi:10.1044/2013_JSLHR-L-12-0126

Peña, E. D., Spaulding, T. J., & Plante, E. (2006). The composition of normative groups and diagnostic decision making: Shooting ourselves in the foot. *American Journal of Speech-Language Pathology, 15,* 247–254. doi:10.1044/1058-0360(2006/023)

Perona, K., Plante, E., & Vance, R. (2005). Diagnostic accuracy of the Structured Photographic Expressive Language Test–Third Edition (SPELT-3). *Language, Speech, and Hearing Services in Schools, 36,* 103–115. doi:10.1044/0161-1461(2005/010)

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

Plante, E., & Vance, R. (1995). Diagnostic accuracy of two tests of preschool language. *American Journal of Speech-Language Pathology, 4*(2), 70–76. doi:10.1044/1058-0360.0402.70

Restrepo, M. A., Schwanenflugel, P. J., Blake, J., Neuharth-Pritchett, S., Cramer, S. E., & Ruston, H. P. (2006). Performance on the PPVT–III and the EVT: Applicability of the measures with African American and European American preschool children. *Language, Speech, and Hearing Services in Schools, 37,* 17–27. doi:10.1044/0161-1461(2006/003)

Reynolds, C., & Kamphaus, R. (2003). *Reynolds Intellectual Screening Test*. Lutz, FL: Psychological Assessment Resources.

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,* 333–349. doi:10.1044/1092-4388(2009/08-0183)

Richards, B. (1987). Type/token ratios: What do they really tell us? *Journal of Child Language, 14,* 201–209. doi:10.1017/S0305000900012885

Rispoli, M. (2003). Changes in the nature of sentence production during the period of grammatical development. *Journal of Speech, Language, and Hearing Research, 46,* 818–830. doi:10.1044/1092-4388(2003/064)

Rondal, J. A., & DeFays, D. (1978). Reliability of mean length of utterance as a function of sample size in early language development. *The Journal of Genetic Psychology: Research and Theory on Human Development, 133,* 305–306.

Shriberg, L. D., Kwiatkowski, J., & Hoffman, K. (1984). A procedure for phonetic transcription by consensus. *Journal of Speech and Hearing Research, 27,* 456–465.

Southwood, F., & Russell, A. F. (2004). Comparison of conversation, freeplay, and story generation as methods of language sample elicitation. *Journal of Speech, Language, and Hearing Research, 47,* 366–376. doi:10.1044/1092-4388(2004/030)

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

Stockman, I. J., Guillory, B., Seibert, M., & Boult, J. (2013). Toward validation of a minimal competence core of morphosyntax for African American children. *American Journal of Speech-Language Pathology, 22,* 40–56. doi:10.1044/1058-0360(2012/11-0124)

Templin, M. C. (1957). *Certain language skills in children: Their development and interrelationships* [Child Welfare Monograph No. 26]. Minneapolis: University of Minnesota Press.

Tommerdahl, J., & Kilpatrick, C. (2013). Analyzing frequency and temporal reliability of children's morphosyntactic production in spontaneous language samples of varying lengths. *Child Language Teaching and Therapy, 29,* 171–183. doi:10.1177/0265659012459528

Ukrainetz, T. A., & Blomquist, C. (2002). The criterion validity of four vocabulary tests compared with a language sample. *Child Language Teaching and Therapy, 18,* 59–78. doi:10.1191/0265659002ct227oa

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,* 1349–1355.

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