JSLHR

Letter to the Editor

Reporting Child Language Sampling Procedures

Lizbeth H. Finestack,[a] Bita Payesteh,[a] Jill Rentmeester Disher,[a] and Hannah M. Julien[a]

Purpose: Despite the long history of language sampling use in the study of child language development and disorders, there are no set guidelines specifying the reporting of language sampling procedures. The authors propose reporting standards for use by investigators who employ language samples in their research.
Method: The authors conducted a literature search of child-focused studies published in journals of the American Speech-Language-Hearing Association between January 2000 and December 2011 that included language sampling procedures to help characterize child participants or to derive measures to serve as dependent variables. Following this search, they reviewed each study and documented the language sampling procedures reported.
Results: The authors' synthesis revealed that approximately 25% of all child-focused studies use language samples to help characterize participants and/or derive dependent variables. They found remarkable inconsistencies in the reporting of language sampling procedures.
Conclusion: To maximize the conclusions drawn from research using language samples, the authors strongly encourage investigators of child language to consistently report language sampling procedures using the proposed reporting checklist.

Language samples and their related analyses have proven to be a useful tool for researchers in the study of child language. Early seminal studies of child language used language samples to describe the development of language (e.g., Bloom, 1970; Brown, 1973; Menyuk, 1964). In current research studies, investigators use measures derived from language samples as primary outcome measures in study analyses as well as to help identify and describe study participants. In contrast to other techniques used to evaluate child language, such as experimental probes and standardized tests, language samples permit researchers to analyze language in contexts that closely resemble the child's natural environment and allow for analyses of language on multiple quantitative and qualitative dimensions (Miller, 1981).

The nature of language sampling leaves the sampling parameters to the discretion of the investigator. Using language samples, investigators may aim to evaluate any one aspect or multiple aspects of language focused on form, content, and/or use (see Miller, 1981). Thus, when designing a study, the investigator determines which aspect(s) of language to evaluate, which in turn will affect the contexts and tasks used to elicit the sample and the analyses to use for evaluation. However, this advantage of language sampling may also be a potential disadvantage, such that the nonstandardized nature of language samples may lead to extreme variability in selected language sampling contexts and related measures (see Hux, Morris-Friehe, & Sanger, 1993; Kemp & Klee, 1997). It is important to recognize this variability given that empirical investigations of language sampling procedures indicate that sampling procedures, including the sampling context, the length of the sample, and transcribing and coding procedures can lead to differential study outcomes (e.g., Chapman, 1981; Gavin & Giles, 1996; Heilmann, Nockerts, & Miller, 2010; Johnston, 2001).

Sample Context

Several empirical studies have documented how differences in language sampling contexts can lead to differences in study outcomes (Abbeduto, Benson, Short, & Dolish, 1995; Evans & Craig, 1992; Kay-Raining Bird, Cleave, White, Pike, & Helmkay, 2008; Kover, McDuffie, Abbeduto, & Brown, 2012; Miles, Chapman, & Sindberg, 2006; Southwood & Russell, 2004). For example, Evans and Craig found that children with specific language impairment produced higher mean length of utterances (MLUs), more advanced syntactic features, and more complex

[a]University of Minnesota, Twin Cities
Correspondence to Lizbeth H. Finestack: finestack@umn.edu
Editor: Rhea Paul
Associate Editor: Thomas Klee
Received April 1, 2014
Revision received August 8, 2014
Accepted August 27, 2014
DOI: 10.1044/2014_JSLHR-L-14-0093

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

semantic features in an interview sampling context compared with a free-play context. Such differences based on sampling context differences demonstrate the need for authors to report specific language sampling procedures. To further illustrate the importance of reporting language sampling context, we can look at findings from a study (Kover et al., 2012) comparing conversational and narrative language production in adolescent boys with Down syndrome, those with fragile X syndrome, and younger boys with typical development. Kover et al. (2012) found a significant interaction effect such that the boys with fragile X syndrome produced significantly higher MLUs in narration than in conversation, whereas the MLUs produced in conversational and narrative contexts by boys with typical development did not differ significantly. Thus, if a study is using language samples to match participant groups similar to those included in the Kover et al. study, matching based on conversational MLU is likely to yield a younger group of typically developing children than matching based on narrative MLU. This, in turn, is likely to yield differential outcomes with a greater likelihood of detecting group differences when samples are matched based on conversational samples than narrative samples. Thus, it is imperative that researchers using language samples to describe participants or for group matching and those using language samples to derive dependent study measures specify the sampling context. This information will not only facilitate replication but will also support interpretation of study results.

Sample Length

Although there is some suggestion of performance stability based on measures derived from relatively short language samples (e.g., Heilmann, DeBrock, & Riley-Tillman, 2013; Heilmann et al., 2010; Tilstra & McMaster, 2007), other investigators have found that measurement reliability may be sensitive to sample length (e.g., Chapman, 1981; Gavin & Giles, 1996; Heilmann et al., 2008, 2010; Hess, Sefton, & Landry, 1986; Johnston, 2001). For example, in a study examining the effect of sample size on type-token ratio calculations, Hess et al. (1986) found that sample sizes with fewer than 100 words had inadequate reliability and that samples needed to include at least 350 words to obtain adequate reliability. Similarly, Gavin and Giles (1996) found that samples that were longer in terms of time and utterances yielded the highest reliability coefficients with adequate reliability associated with samples 20 min in length or containing at least 100 utterances. Variables that impact the reliability of language samples include context and participant age (Gillam & Pearson, 2004; Heilmann, et al., 2008) such that the samples obtained from highly controlled contexts (e.g., narrative retell) and older participants are likely to be more reliable than less controlled (e.g., conversational) samples of preschool children. Thus, it is necessary for researchers to customize language sample length for their particular study; however, sample length must be reported for others to judge the reliability of the study measures.

Transcription and Coding Procedures

If the study design demands for language samples to be transcribed or coded, details regarding the transcription and coding procedures are particularly important for both interpretation of study results and study replication. To illustrate this point, we use comparisons of utterance segmenting conventions and their relationship to MLU calculations. The Child Language Data Exchange coding system, Codes for Human Analysis of Transcripts (CHAT; MacWhinney, 2000), was designed to provide a standardized format for transcription and includes coding options for basic discourse transcription analyses as well as more detailed phonological and morphological analyses. Although CHAT leaves specific coding procedures and the level of coding detail to the discretion of the investigator, the CHAT manual offers segmenting guidance for utterances containing repeated words, lists, repetitions segmenting utterances, and tag questions. However, CHAT does not specify the number of independent clauses that may be conjoined in a single utterance. Similar to the Child Language Data Exchange System, transcription conventions of Systematic Analysis of Language Transcripts (Miller & Chapman, 2000) allow investigators extreme flexibility in coding; however, the recommended coding procedures are specified in greater detail. For example, in Systematic Analysis of Language Transcripts, utterances are segmented into *communication units* (C-units; Loban, 1976), which are defined as "an independent clause with its modifiers" (p. 9) and include incomplete sentences such as elliptical responses to questions (e.g., Where are you going? *To my house*). CHAT does not specify the utterance unit to be used, and researchers may choose to use other segmenting procedures such as terminal units (T-units; Hunt, 1965), which, like C-units, include an independent clause and its modifiers but exclude incomplete sentences. Another utterance segmenting option is to allow up to two independent clauses conjoined by a conjunction in one utterance. This procedure is specified when analyzing transcripts using Developmental Sentence Scoring (DSS; Lee, 1974). Yet another segmenting rule allows all conjunctions to be included within utterances except *and*. This procedure yields analysis units that have been specified for use when conducting Language Assessment Remediation and Screening Procedure analyses (Fletcher & Garman, 1988).

Differences in utterance segmentation are highly likely to result in different MLU values in which the total number of words or morphemes in the sample serves as the numerator and the total number of utterances serves as the denominator in the MLU formula. For example, MLUs based on C-units are likely to be lower due to the inclusion of incomplete shorter utterances than MLUs based on T-units, which only include complete sentences and are naturally longer. MLUs based on samples transcribed

according to DSS procedures are likely to be longer than either C- or T-unit–based MLUs. An investigator may choose to use any of these segmenting procedures for a particular study; however, to validly compare MLU values across studies and to appropriately interpret study results, it is necessary for investigators to specify the transcription utterance units used as well as other transcription procedures (e.g., morphemes counted, coding of repetitions). For other measures derived from language samples, it is imperative for investigators to describe transcription and coding procedures central to the measure (e.g., definition of turns, criteria for determining utterances errors, coding of dialect differences).

Transcription and Coding Reliability

Depending on the nature of the measures derived from the language sample, transcription and/or coding reliability influences interpretability of study findings. In many studies using language sampling, samples are transcribed and subsequently coded to derive study measures. For a standardized test to be viewed as valid, one criterion that the test must meet is a high level of reliability (Hutchinson, 1996). Similarly, to derive valid interpretations from study data obtained from language samples, the samples need to be reliable on relevant measures. For example, if using MLU in morphemes (total number of morphemes/total number of utterances), reliability of the number of morphemes identified and the number of utterances segmented must be reported.

In cases where the language sample is subsequently coded to derive another measure, accuracy of the transcription that the coding is based on, as well as of the secondary coding scheme, must be high. For example, if using a coding scheme such as DSS or the Index of Narrative Complexity (Petersen, Gillam, & Gillam, 2008), reliability of the language sample transcription plus reliability of the DSS or Index of Narrative Complexity coding should be reported. These dual reliability reports are necessary to reveal the total amount of error associated with the variable(s) of interest. As noted by Whittington (2003), when reliability information is missing, "the meanings of a study's results are jeopardized, regardless of the importance of the research question, the quality of the research design, or the sophistication of the data analysis techniques" (p. 187).

Given the variability associated with language sampling contexts, lengths, and transcription procedures, it is necessary for investigators to meticulously document and report such details. Such information confirms whether the language sampling measures employed in the study are valid and reliable and also allows for study replication. To determine whether such reporting was common practice, we conducted a literature search of studies that included language sampling procedures to help characterize child participants and/or to derive measures to serve as dependent variables. Following this search, we reviewed each study and documented whether the article reported the specific sampling context, sampling procedures (i.e., length of the samples in units of time and/or utterances), transcription procedures used, and transcription reliability.

Literature Search and Review

To examine the reporting of language sampling procedures used by researchers, we searched all of the journals of the American Speech-Language-Hearing Association (ASHA; i.e., *American Journal of Audiology*; *American Journal of Speech-Language Pathology*; *Journal of Speech, Language, and Hearing Research*; *Language, Speech, and Hearing Services in Schools*) from January 2000 through December 2011 (http://journals.pubs.asha.org). We confined our search to include only ASHA journals to gain a sense of the reporting of language sample procedures within the field of speech-language pathology. Our initial search indicated that, from January 2000 through December 2011, ASHA published 2,532 articles. In this time frame, there were approximately 1,178 studies focused on children published in ASHA journals. We derived this estimate by searching all journals for the key word "child" in abstracts and titles.

To identify articles reporting the use of child language sampling in study procedures, we completed four separate advanced searches in which the following phrases and key words were searched for in the text, abstract, and title: "language sample," "narrative," "mean length of utterance," and "MLU." The "language sample" search yielded a total of 254 articles; the "narrative" search yielded 490 articles. The "mean length of utterance," and "MLU" searches yielded 314 and 263 articles, respectively. Of these 1,321 articles, 595 were duplicate articles, appearing in more than one search result. This reduced the sample to 726 articles.

These 726 articles were analyzed to ensure that they met the inclusion criteria for the review. Specifically, each article was required to meet the following criteria: (a) be characterized as a traditional empirical study (surveys, literature reviews, tutorials, and meta-analyses were excluded); (b) include child participants between the ages of 0 and 18 years; and (c) use at least one measure derived from a language sample that reflects language microstructure or macrostructure (i.e., phonological variables were excluded) to describe study participants and/or to serve as a dependent measure for analysis. For articles that met these criteria, we documented whether investigators reported the sampling context, sample length, transcription or coding procedures, and reliability. The authors completed all reviews by consulting the article abstracts and the full-text publication when necessary. Each author reviewed approximately one fourth of the articles. Once this review was completed, approximately 25% of the nonduplicate articles were independently reviewed by one of the four authors as a reliability check. Upon comparison of inclusion decisions, a total of seven discrepancies (0.05%) were identified. These discrepancies were discussed among the authors and resolved.

A total of 290 studies met the inclusion criteria. Of these studies, 125 used language samples to help describe participants' language independent of measures used for analyses (i.e., MLU based on a language sample to match participant groups), and 238 used language samples to derive a dependent variable for study analyses (i.e., MLU based on a language sample to compare study groups). Thus, we estimate that the rate of inclusion of child language samples in studies published in ASHA journals between January 2000 and December 2011 to help characterize participants is approximately 5% (125/2,532) and to derive dependent study variables is approximately 10% (238/2,532).

Our review reveals that investigators reported far fewer details regarding language sampling procedures when the language sample was used to describe study participants than when the sample was used to derive a primary study measure. More specifically, when language samples were used to characterize participants, 60% of studies specified the sampling context compared with 97% of studies using samples for study measures. Similarly, when language samples were used to characterize study participants, less than 30% of articles included a description of the language sample length in units of time or number of utterances. However, when language samples were used to derive dependent measures, the majority (62%) of studies included a description of length. In respect to the reporting of transcription and/or coding procedures, approximately 57% of articles using language samples to describe participants described the procedures and approximately 90% of studies using language samples for dependent measures included descriptions of the transcription/coding procedures. In terms of reliability, few studies (less than 6%) using language samples to describe participants included any information regarding transcription or coding reliability. In contrast, well over half (68%) of articles using language samples to derive dependent measures reported some estimate of transcription and/or coding reliability. Despite the finding that more sampling details are reported in studies using language samples to derive dependent measures than when samples are used to help describe participants, examination of the sampling reporting habits associated with samples used to derive dependent measures reveals a considerable lack in reporting. This is particularly true for details regarding length, procedures, and reliability associated with the sample.

Proposal

Our review and synthesis documents the relatively high prevalence of language sample use in child-focused research to describe participants or obtain key study variables. This reveals that researchers are taking advantage of the inherent flexible nature of language samples. However, we found substantial discrepancies in the reporting of specific language sampling procedures, particularly when samples were used to characterize study participants. Thus, to uphold the integrity of language samples as tools for research, we propose a checklist to guide the reporting of language sampling procedures. Similar checklists have been adopted by ASHA for use when reporting clinical studies in its journals, including Consolidated Standards of Reporting Trials for the reporting of randomized controlled trials (Schulz, Altman, & Moher, 2010), Transparent Reporting of Evaluations with Nonrandomized Designs for reporting nonrandomized trials (Des Jarlais, Lyles, Crepaz, & the TREND Group, 2004), and Standard for Reporting Diagnostic Accuracy for reporting studies of diagnostic accuracy (Bossuyt et al., 2003). Additionally, the American Psychological Association has adopted a set of standards to use when reporting studies using new data: Journal Article Reporting Standards (JARS; Cooper, 2010). The standards recommended by JARS include separate modules for reporting particular types of studies (e.g., intervention) or elements of studies (e.g., use of random or nonrandom assignment).

In Table 1, we propose a checklist to guide reporting in journal articles when language samples are used in a study either to characterize study participants or as a study measure. Such a checklist could be used only when applicable, similar to the JARS modules. Minimally, we propose that it is necessary for authors to report the context and length of the language sample. Such information allows readers to make some broad judgments regarding the type of sample used and how this may impact reliability and validity of the sample. Reporting the context and length would be similar to authors' reporting the name of the standardized assessment used rather than stating "expressive language test" or "vocabulary test." In addition to information regarding the context and length, if study measures (e.g., MLU or type-token ratio) are derived from the sample, transcription and transcription reliability should be reported. If transcripts are further coded, the specific procedures used for the coding and coding reliability should be reported. Authors should be able to report all of this information in a short paragraph included in the article's method section.

Conclusion

Provision of the information suggested in the language sampling reporting checklist is critical to interpretation and evaluation of study findings. Without this information, readers are left wondering whether the reported study was properly conducted and are limited in their ability to make soundly reasoned judgments of the study results (Whittington, 1998). Thus, to allow for study replication and adequate interpretation of study results, it is imperative that authors provide the details of their language sampling procedures, including the sampling context, length of the sample, transcription procedures, and reliability of transcription and language sample coding procedures. This information ensures the integrity of the language sample(s). It is the collective responsibility of authors, reviewers, and editors to ensure that these standards are upheld to strengthen the impact of our research.

Table 1. Proposed checklist of information to include when reporting a study using language samples.

Topic	✓	Description
Sampling context		• Type of context (e.g., free play, narrative, expository, conversation) • Procedure for eliciting sample (e.g., retell, self-generated) • Who participated in the sample (e.g., participant only, parent, peer)
Sample length		• Targeted and actual length of sample in terms of time and/or utterances, depending on the unit that is most meaningful to measures reported
Transcription procedures (if applicable)		• Who transcribed the samples • Training of transcribers • Procedures used, including segmenting procedures and other transcription procedures
Transcription reliability (if applicable)		• Who completed the reliability • The number of samples on which the reliability is based • How reliability comparisons were made and on which variables • How reliability was calculated between samples (e.g., point-by-point, intraclass coefficients based on totals)
Coding procedures (if applicable)		• Who coded the samples • Training of coders • Procedures used for coding
Coding reliability (if applicable)		• Who completed the reliability • The number of samples on which the reliability is based • How reliability comparisons were made and on which variables • How reliability was calculated between samples (e.g., point-by-point, intraclass coefficients based on totals)

References

Abbeduto, L., Benson, G., Short, K., & Dolish, J. (1995). Effects of sampling context on the expressive language of children and adolescents with mental retardation. *Mental Retardation, 33,* 279–288.

Bloom, L. (1970). *Language development: Form and function in emerging grammars*. Cambridge, MA: The MIT Press.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glasziou, P. P., Irwig, L. M., … Lijmer, J. G. (2003). The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. *Annals of Internal Medicine, 138,* W1–12.

Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.

Chapman, R. (1981). Computing mean length of utterance in morpheme. In J. Miller (Ed.), *Assessing language production in children* (pp. 22–25). Baltimore, MD: University Park Press.

Cooper, H. M. (2010). *Reporting research in psychology: How to meet journal article reporting standards*. Washington, DC: American Psychological Association.

Des Jarlais, D. C., Lyles, C., & Crepaz, N., & the TREND Group. (2004). Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: The TREND statement. *American Journal of Public Health, 94,* 361–366.

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.

Fletcher, P., & Garman, M. (1988). LARSPing by numbers. [Comment/reply]. *British Journal of Disorders of Communication, 23,* 309–321. doi:10.3109/13682828809011940

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.

Gillam, R., & Pearson, N. A. (2004). *Test of Narrative Language*. Austin, TX: Pro-Ed.

Heilmann, J., DeBrock, L., & Riley-Tillman, T. C. (2013). Stability of measures from children's interviews: The effects of time, sample length, and topic. *American Journal of Speech-Language Pathology, 22,* 463–475. doi:10.1044/1058-0360(2012/11-0035)

Heilmann, J., Miller, J. F., Iglesias, A., Fabiano-Smith, L., Nockerts, A., & Andriacchi, K. D. (2008). Narrative transcription accuracy and reliability in two languages. *Topics in Language Disorders, 28,* 178–188.

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)

Hess, C. W., Sefton, K. M., & Landry, R. G. (1986). Sample size and type-token ratios for oral language of preschool children. *Journal of Speech and Hearing Research, 29,* 129–134.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels* (NCTE Research Report No. 3). Champaign, IL: NCTE.

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.

Hux, K., Morris-Friehe, M., & Sanger, D. D. (1993). Language sampling practices: A survey of nine states. *Language, Speech, and Hearing Services in Schools, 24,* 84–91.

Johnston, J. R. (2001). An alternate MLU calculation: Magnitude and variability of effects. *Journal of Speech, Language, and Hearing Research, 44,* 156–164. doi:10.1044/1092-4388(2001/014)

Kay-Raining Bird, E., Cleave, P. L., White, D., Pike, H., & Helmkay, A. (2008). Written and oral narratives of children and adolescents with Down syndrome. *Journal of Speech, Language, and Hearing Research, 51,* 436–450.

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

Kover, S. T., McDuffie, A., Abbeduto, L., & Brown, W. T. (2012). Effects of sampling context on spontaneous expressive language in males with fragile X syndrome or Down syndrome. *Journal of Speech, Language, and Hearing Research, 55,* 1022–1038. doi:10.1044/1092-4388(2011/11-0075)

Lee, L. (1974). *Developmental sentence analysis*. Evanston, IL: Northwestern University Press.

Loban, W. (1976). *Language development: Kindergarten through grade twelve*. Urbana, IL: National Council of Teachers of English.

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

Menyuk, P. (1964). Comparison of grammar of children with functional deviant and normal speech. *Journal of Speech and Hearing Research, 7,* 109–121.

Miles, S., Chapman, R., & Sindberg, H. (2006). Sampling context affects MLU in the language of adolescents with Down syndrome. *Journal of Speech, Language, and Hearing Research, 49,* 325–337.

Miller, J. (1981). *Assessing language production in children: Experimental procedures*. Boston, MA: Allyn and Bacon.

Miller, J., & Chapman, R. (2000). *SALT: Systematic Analysis of Language Transcripts* [Computer software]: Language Analysis Laboratory, Waisman Center, University of Wisconsin-Madison.

Petersen, D. B., Gillam, S. L., & Gillam, R. B. (2008). Emerging procedures in narrative assessment: The index of narrative complexity. *Topics in Language Disorders, 28,* 115–130. doi:10.1097/01.TLD.0000318933.46925.86

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *Trials, 11,* 32. doi:10.1186/1745-6215-11-32

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)

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

Whittington, D. (1998). How well do researchers report their measures? An evaluation of measurement in published educational research. *Educational and Psychological Measurement, 58,* 21–37.

Whittington, D. (2003). How well do researchers report their measures? An evaluation of measurement in published educational research. In B. Thompson (Ed.), *Score reliability: Contemporary thinking on reliability issues* (pp. 173–190). Thousand Oaks, CA: Sage.