LSHSS

Research Article

Sample Size for Measuring Grammaticality in Preschool Children From Picture-Elicited Language Samples

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

Purpose: The purpose of this study was to investigate whether a shorter language sample elicited with fewer pictures (i.e., 7) would yield a percent grammatical utterances (PGU) score similar to that computed from a longer language sample elicited with 15 pictures for 3-year-old children.
Method: Language samples were elicited by asking forty 3-year-old children with varying language skills to talk about pictures in response to prompts. PGU scores were computed for each of two 7-picture sets and for the full set of 15 pictures.

Results: PGU scores for the two 7-picture sets did not differ significantly from, and were highly correlated with, PGU scores for the full set and with each other. Agreement for making pass–fail decisions between each 7-picture set and the full set and between the two 7-picture sets ranged from 80% to 100%.
Conclusion: The current study suggests that the PGU measure is robust enough that it can be computed on the basis of 7, at least in 3-year-old children whose language samples were elicited using similar procedures.

Language sample analysis (LSA) is an important part of assessing a child's language. Some LSA provides a fine-grained analysis of a particular aspect of language, such as use of verb tense markers or noun phrase elaboration, for the purpose of determining strengths and weaknesses and selecting treatment goals (e.g., Eisenberg et al., 2008; Leonard, Camarata, Brown, & Camarata, 2004). Other LSA measures, such as mean length of utterance and number of different words, broadly characterize a child's linguistic performance for the purpose of identifying a language impairment (LI) or measuring change over time (e.g., Rice et al., 2010; Scott & Windsor, 2000). These are referred to as general language performance measures. The current article looks at one general language performance measure —percent grammatical utterances—that measures grammaticality (i.e., the extent to which a child's language production conforms to the conventions of English grammar and usage) using picture-elicited language samples.

One frequently given reason for not using language sample analysis in clinical work is the amount of time it takes for eliciting, transcribing, and analyzing the sample (Hux, Morris-Friehe, & Sanger, 1993; Kemp & Klee, 1997). This time could be reduced by using shorter samples. However, it needs to be established whether those shorter samples will yield the same results as longer samples. In the current article, we focused on evaluating whether percent grammatical utterances in 3-year-old children varies with the number of pictures that are used to elicit language samples.

Why Grammaticality Is Important

One aim of the common core educational standard for language is for students to demonstrate command of the conventions of standard English grammar and usage when speaking or writing (e.g., State of New Jersey Department of Education, 2010). This standard is particularly relevant for children with LI because the presence of grammatical deficits is a key characteristic of these children (Leonard, 1998). Measures of grammaticality have been shown to distinguish between preschool children with and without LI (Dunn, Flax, Sliwinski, and Aram, 1996; Eisenberg & Guo, 2013; Souto, Leonard, & Deevy, 2014). Eisenberg and Guo (2013), for instance, reported a sensitivity rate of 100% and a specificity rate of 88% for differentiating between 3-year-old children with and without LI on the basis of the percentage of utterances that were grammatical. Similarly, Souto, Leonard, and Deevy (2014) reported 93% sensitivity and 94% specificity in their study of 4-year-old children.

[a]Montclair State University, Montclair, NJ
[b]University at Buffalo–The State University of New York

Correspondence to Sarita L. Eisenberg: eisenbergs@mail.montclair.edu

Editor: Marilyn Nippold
Associate Editor: Mary Alt

Received April 24, 2014
Revision received September 19, 2014
Accepted November 25, 2014

DOI: 10.1044/2015_LSHSS-14-0049

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

Broad measures of grammaticality are also sensitive to age and language status during the school years (Fey, Catts, Proctor-Williams, Tomblin, & Zhang, 2004; Guo & Schneider, 2014; Scott & Windsor, 2000; Westerveld & Gillon, 2010). In a study by Fey, Catts, Proctor-Williams, Tomblin, and Zhang (2004), second- and fourth-grade children with LI produced a lower proportion of grammatical utterances than same-aged children with typical language for both spoken and written narratives. Likewise, school-age children with LI (mean age 11;6 [years;months]) in a study by Scott and Windsor (2000) showed significantly higher rates of grammatical errors than same-aged children with typical language for both spoken and written expository samples as well as narrative samples.

Percent Grammatical Utterances (PGU) as a Measure of Grammaticality

The most frequently noted deficit for children with LI is with usage of verb tense markers (Bedore & Leonard, 1998; Rice, Wexler, & Cleave, 1995). However, children with LI have also been reported to make other grammatical errors, such as argument omissions (Grela & Leonard, 1997), pronoun errors (Loeb & Leonard, 1991; Moore, 2001), and omissions and errors on other grammatical morphemes (e.g., articles, plurals) more often than typically developing children (Leonard, Eyer, Bedore, & Grela, 1997; Polite, Leonard, & Roberts, 2011; Watkins & Rice, 1991). Eisenberg and colleagues (Eisenberg & Guo, 2013; Eisenberg, Guo, & Germezi, 2012), therefore, suggested using percent grammatical utterances (PGU) as a broader measure that incorporates a variety of properties that can affect grammaticality rather than focusing exclusively on verb tense marking.

Eisenberg and colleagues examined PGU in 3-year-old children with and without LI using picture-elicited language samples (Eisenberg & Guo, 2013; Eisenberg, Guo, & Germezi, 2012). In those studies, language samples were collected by asking children to describe 15 pictures one at a time in response to four elicitation questions (see Appendix A). The language samples were segmented into communication units (C-units) as in studies of grammaticality in school-age children (Fey et al., 2004; Guo & Schneider, 2014; Westerveld & Gillon, 2010). A C-unit, by definition, can be one independent clause plus any number of dependent clauses (Loban, 1963). Unlike the rules in Developmental Sentence Scoring (DSS; L. L. Lee, 1974), which includes only complete utterances (i.e., utterances with both a subject and main verb), the PGU calculation includes utterances without a subject (e.g., *Washing dishes*) or a main verb (e.g., *He hungry*) if those constituents are obligatory.

A C-unit was judged as ungrammatical if it had one or more structural errors or semantic irregularities. This was broader than the criteria described by Scott and Windsor (2000) because they included only structural errors in their computation. There were three reasons for the inclusion of semantic irregularities in determining the grammaticality of C-units. First, syntax is not independent of meaning. Rather, semantics contributes to the well-formedness of sentences (Halliday, 1994; Saeed, 2009). Second, this decision is consistent with other assessments, such as DSS (L. L. Lee, 1974) and the Sentence Formulation subtest of the Clinical Evaluation of Language Fundamentals, Fourth Edition (CELF-4; Semel, Wiig, & Secord, 2003), both of which score semantic irregularities as errors. Third, we wanted consistency between scoring substitution errors on grammatical morphemes (e.g., *A boy <u>in</u> the stool is falling down*) and errors on content words. Thus, utterances such as "*They are brooming the leaves*" or "*She's gonna fall to the ceiling*" were scored as ungrammatical.

PGU was calculated by dividing total number of grammatical C-units by the total number of C-units in the picture-elicited language samples. Fey, Catts, Proctor-Williams, Tomblin, and Zhang (2004) used an identical calculation in their study of school-aged children, dividing the percentage of C-units not containing errors by the total number of C-units. Similarly, L. L. Lee (1974) awarded one point (termed a sentence point) to utterances without errors and divided total sentence points by the total number of utterances to calculate a mean sentence point score.

PGU was significantly correlated ($r = .53$) with the Structured Photographic Expressive Language Test–Preschool 2 (SPELT-P2; Dawson et al., 2005), a standardized test of morphology and syntax (Eisenberg & Guo, 2010). Following Sackett (1991), a cutoff score for PGU was determined by first using a receiver operating characteristic (ROC) curve to generate pairs of sensitivity and specificity rates for a range of cutoff scores and then choosing the cutoff that maximized diagnostic accuracy. A PGU score of 58.32% yielded sensitivity at a 100% level and specificity at an 88% level for differentiating previously diagnosed 3-year-old children with and without LI (Eisenberg & Guo, 2013). These data supported the use of PGU for measuring children's grammaticality.

Types of Language Samples for Measuring Grammaticality

Picture-elicited language samples rather than conversational samples during free play were chosen as the context for assessing grammaticality in the studies of Eisenberg and colleagues (Eisenberg & Guo, 2013; Eisenberg et al., 2012) for two reasons. First, even when materials and topics are controlled for during conversation, utterances addressed to the child will vary. Using pictures with specific prompts to collect language samples allowed us to standardize what was said to the child so that all children responded to the same utterances. Second, children with LI are more likely to produce elliptical responses during conversation than are children with typical language (Johnston, Miller, Curtiss, & Tallal, 1993). Using pictures as the stimuli allowed us to provide prompts that would obligate the subject and reduce opportunities for ellipsis. This was important because pilot work on assessing grammaticality for conversational samples showed poor reliability for judging whether subjectless sentences and fragments were truly ungrammatical or pragmatically allowable.

Previous studies of school-aged children have calculated grammaticality on the basis of narrative and/or expository samples (Fey et al., 2004; Scott & Windsor, 2000; Westerveld & Gillon, 2010). However, a picture-description procedure was chosen over narrative or expository sampling because many younger children experience difficulty in generating narratives (Berman & Slobin, 1994; Burns-Hoffman, 1993), and similar procedures have been adopted for young children in other studies (e.g., Darley & Moll, 1960; L. L. Lee, 1974).

Sample Size for Calculating Language Sample Measures in Different Tasks

A sample size of 50 to 100 utterances has been suggested as the minimum number of utterances necessary for reliably measuring utterance length and vocabulary from a conversational sample (e.g., Miller et al., 2011; Paul & Norbury, 2012). However, in a survey by Hux, Morris-Friehe, and Sanger (1993), 25% of respondents reported using fewer than 50 utterances for LSA. Although some studies have suggested that smaller sample sizes of 20 to 30 utterances have adequate reliability for measuring utterance length and/or vocabulary (Casby, 2011; Heilmann, Nockerts, & Miller, 2010), the majority of studies have concluded that conversational samples of about 100 utterances are needed to achieve acceptable reliability (i.e., reliability at a .90 level; Bogue, DeThorne, & Schaefer, 2014; Gavin & Giles, 1996; McCauley & Swisher, 1984) for these measures (Cole, Mills, & Dale, 1989; Darley & Moll, 1960; Rondal & DeFays, 1978).

More structured speaking tasks typically yield fewer utterances than conversational sampling (e.g., Merritt & Liles, 1989; Southwood & Russell, 2004). Heilmann, Nockerts, and Miller (2010) examined narrative samples and concluded that 3-min narrative samples, yielding on average 30 utterances, would be adequate for measuring mean length of utterance in morphemes (MLUm), total number of words, and number of different words as part of a comprehensive battery for assessing language. However, although internal reliability for the 3-min sample was at an acceptable (i.e., .90) level for the vocabulary measures, reliability for MLUm was below .80. Of particular importance for the present study, reliability for measuring grammaticality, calculated as the number of omissions and errors per minute, was below .70 in the 3-min narrative samples.

Most of the previous studies determined the effect of sample size on the basis of the number of utterances. To the best of our knowledge, only one study has investigated sample size for an item-based task that involved talking about pictures in response to prompts. This was a study by Brookshire and Nicholas (1994), and it considered the number of items needed to obtain a reliable measure of percent correct information units from adults with and without brain injury. Percent correct information units measure informativeness by dividing the number of words that are accurate and relevant to the task by the total number of words produced. Their protocol included 10 items administered in random order: four single pictures, two picture sequences, two requests for personal information, and two requests for procedural information. Sample size was manipulated by controlling for number of items rather than for length of the language sample because this reflects the way clinicians control sample size. Scores for percent correct information units stabilized after participants completed four to five items, with low reliability for scores based on fewer items.

In summary, previous studies have shown that samples of 100 utterances are needed to achieve acceptable reliability for MLUm from conversational samples in children (Cole et al., 1989; Darley & Moll, 1960; Rondal & DeFays, 1978). Approximately 30 utterances are required to compute acceptably reliable vocabulary measures from narrative samples (Heilmann et al., 2010). In addition, four items (i.e., pictures and/or interview topics) are needed to calculate reliable percent correct information units in an item-based task for adults (Brookshire & Nicholas, 1994). However, the numbers of utterances or items that are needed in order to compute reliable PGU scores in children remains an open question, even though studies have shown that PGU scores differentiate children with and without LI (Eisenberg & Guo, 2013; Fey, Catts, Proctor-Williams, Tomblin, & Zhang, 2004; Guo & Schneider, 2014; Souto et al., 2014).

Purpose

As an initial step to fill in the gap in the literature, the current study examined whether PGU was affected by sample size in 3-year-old children. We used number of pictures as the measure of sample size because this coincides with clinical practice. That is, clinicians would administer a specific number of pictures to elicit the language sample rather than basing task length on reaching a specific number of utterances.

The picture description procedure used to collect language samples for calculating PGU (Eisenberg & Guo, 2013) had included 15 pictures. The average number of C-units produced by the children in the picture-elicited language samples was 72 (standard deviation [*SD*] = 18) for 3-year-old children with typical language and 62 (*SD* = 15) for children with LI. Administration time ranged from approximately 1 to 3 min per picture (*M* = 1.8 min) for a total time of 14 to 40 min (*M* = 27 min; *SD* = 7 min). This wide range in administration time occurred partly because some participants needed more reinforcing activities than others during the tasks. Given that using reinforcing activities is inevitable for 3-year-old children, the entire task could be overly long for some children and may not be ideal for clinical work. Examining how sample size affects the PGU score would also allow us to determine whether we can use fewer pictures in collecting picture-elicited language samples.

The current study examined how sample size (i.e., number of pictures) affected the PGU score from two sets of analyses. First, we examined the consistency of the PGU scores between shorter samples and longer samples. To this end, we divided the 15 pictures used by Eisenberg and Guo (2013) into two seven-picture sets (Sets 1 and 2), with one remaining picture randomly administered together with Set 1 or Set 2. Each of the seven-picture sets was administered in a separate session, with the order counterbalanced. We computed the differences and correlations of the PGU scores between each of the two seven-picture sets (i.e., Sets 1 and 2) and the 15-picture set (i.e., the total sample set). We also looked at agreement for making pass–fail decisions on the basis of PGU for the seven-picture and 15-picture sets. We asked the following questions:

1. Would the PGU score for the shorter samples (i.e., the two seven-picture sets) significantly differ from the PGU score for the longer sample (i.e., the 15-picture set)?
2. Were the PGU scores for the shorter samples correlated with the PGU scores for the longer sample at an acceptable level (i.e., correlation coefficient ≥ .90)?
3. To what extent would the shorter samples yield the same pass–fail decision (i.e., score above or below a PGU cutoff of 58.23% on the basis of Eisenberg & Guo, 2013) as the longer sample?

Second, we examined the consistency of the PGU scores between the two seven-picture sets. We computed the differences and correlations of the PGU scores between the two seven-picture sets, and we looked at agreement for making pass–fail decisions on the basis of PGU scores for the two seven-picture sets. We asked the following questions:

4. Would the PGU scores significantly differ between the two shorter samples?
5. Were the PGU scores from the two shorter samples correlated with each other at an acceptable level (i.e., correlation coefficient ≥ .90)?
6. To what extent would the two shorter samples yield the same pass–fail decision?

The following results would suggest that clinicians could use fewer pictures for calculating PGU: (a) if PGU scores for the two half-sets correlated with each other and with PGU for the longer sample; (b) if PGU for the two half-sets did not differ significantly from each other or from PGU on the basis of the longer sample; and (c) if PGU for the two half-sets resulted in the same clinical decision as each other as well as the longer sample.

Method

Participants

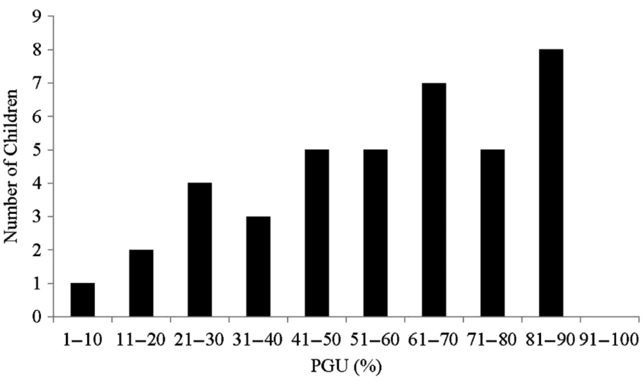
Participants included 40 children from an existing database (17 girls, 23 boys) ranging in age from 3;0 to 3;11, with a mean age of 3;5 (*SD* = 0;3). All participants were from suburban areas of northern New Jersey and had been recruited for a study about language production in young children. Approval for this research was granted by the Montclair State University Institutional Review Board, and parents had provided informed consent for the archived data to be used in further studies of language development.

The children for the current study were selected without consideration of how they had performed on a standardized language test or a conversational language sample as well as information about diagnostic status or parental concern. This was done in order to ensure that the study had the ability to obtain PGU scores from children with varying language levels (see Ukrainetz McFadden, 1996). Indeed, the resulting picture-elicited language samples yielded PGU scores ranging from 8% to 89% (see Figure 1 for the distribution of PGU on the basis of 15 pictures). The children were all monolingual English-speaking on the basis of parent report. Children were excluded if they or their parents spoke a non-mainstream dialect of English as reported by the parent. To be included in this study, children had to pass a hearing screening at 25 dB for the frequencies 500, 1000, 2000, and 4000 Hz and had to have nonverbal cognitive ability within the typical range on the basis of the Odd-Item-Out task of the Reynolds Intellectual Screening Test (RIST; Reynolds & Kamphaus, 2003). Children also had to pass the Articulation Subtest of the Fluharty Preschool Speech and Language Screening Test, Second Edition (Fluharty, 2001). Socioeconomic status (SES) was based on maternal education, with 90% having a college degree and 10% having a high school diploma as their highest level of education. Racial distribution, based on self-identification by the parent, was 75% White, 20% African American, and 5% Asian. Twenty percent self-identified as Hispanic.

Stimuli

Language samples were elicited by prompting the child to talk about 15 pictures. The pictures were selected from a file of pictures that had been collected over time by the first author to use for language sampling and therapy activities. The source for nine of the pictures was unknown. Each of the pictures included at least two people and had at least one other character that was either a person, animal,

Figure 1. Distribution of percent grammatical utterances (PGU) scores for the total sample.



or doll. Twelve of the 15 pictures were colored line drawings. The other three pictures were photographs from magazine advertisements. A description of each picture and its source, if known, is provided in Appendix B.

To evaluate whether PGU scores would vary with sample size (i.e., number of pictures), we divided the full set of 15 pictures into two seven-picture sets (Sets 1 and 2), with one remaining picture randomly administered together with Set 1 or Set 2. This decision was made to ensure that Sets 1 and 2 had an equal number of pictures. The task was administered in two sessions, with one picture set completed in each session. The order for administering the sets was randomized across sessions.

Language Sample Collection Procedure

Each child was individually tested by either the second author or a student research assistant who was trained to administer the task. The order of pictures within each session was randomized by allowing the child to select the next picture to talk about. There were four prompts, adapted from Leonard, Bolders, and Miller (1976), for each picture. The prompts were selected so that the children would be obligated to produce declarative sentences with a subject and predicate. The first, second, and last prompt were the same for all of the pictures. The third prompt involved a story starter that was specific to each picture. An alternative prompt was given if a child did not respond to the first prompt, responded with "I don't know," or produced an utterance that did not relate to the picture. A list of the prompts is provided in Appendix A. The entire procedure was audio-recorded for transcribing and coding.

Language Sample Preparation and Coding

Student research assistants were trained to transcribe the samples following the conventions of Systematic Analysis of Language Transcripts (SALT; Miller & Iglesias, 2010). Utterances were segmented into C-units. As described by Loban (1963), a C-unit consists of an independent clause plus any number of dependent clauses. That is, a C-unit can include a maximum of one independent clause. A clause, by definition, must have at least a subject and a predicate (e.g., a verb). Phrasal utterances (e.g., *a boy*, *a dog*, and *a mom*) and utterances without subjects (e.g., *have a book*) or verbs (e.g., *The dog mad*) can also be counted as C-units as long as they are complete thoughts (Miller & Iglesias, 2010). Utterances that are not completed, either because the child was interrupted or because the child discontinued the utterance, are not counted as C-units.

Only completely intelligible and on-topic C-units (that is, only language focused on the pictures in contrast to external comments to the examiner) were included in the analysis. Any questions that the children asked about the pictures were excluded (e.g., *What he got in his hand?*). C-units that were not complete clauses were included in the PGU analysis if a full sentence response with both a subject and predicate were obligated. Thus, fragments, defined as C-units with only a phrasal constituent, were included if the examiner's prompt required a complete sentence. Utterances without a subject or without a main verb were also included if these elements were obligatory. However, any elliptical responses (i.e., utterances in which nonproduction of the subject or predicate was judged to be pragmatically allowable) to examiner requests for repetition or clarification (e.g., *The boys are what? Throwing snowballs*) were excluded from the analysis. Self-corrections after examiner queries (e.g., *What did you say?*) were also excluded.

All C-units with one or more errors in grammatical morphology, syntax, and/or semantics were marked as ungrammatical. Errors included (a) omissions of obligatory constituents, such as the subject or object of a verb, (b) substitution errors for subject (e.g., *her is mad*), object, reflexive and possessive pronouns and possessive determiners, (c) omissions and usage errors for verb tense markers including copula, auxiliary *be*, auxiliary *do*, third person singular *-s*, and regular and irregular past tense (Rice, Wexler, & Cleave, 2001), (d) omissions or substitutions of other bound or free morphemes (e.g., *-ing*, article *a*/*the*), (e) production of fragments in response to questions and prompts that obligated a complete sentence (e.g., *a boy*, *a dog*, and *a mom*), and (f) other syntactic errors or semantic irregularities (L. L. Lee, 1974) that did not fall into another error category. Pronouns and nouns were scored as gender errors only when there was inconsistency in the child's usage (e.g., referring to the same character as "girl" and "he"). Instances where the child's usage did not match the pictured character were not counted as errors (e.g., when the child consistently referred to a picture of a girl as "he" or "boy").

During pilot studies with the PGU measure, reliability for judging the acceptability of turn initial conjunctions produced after the elicitation question or prompt was low. An example of this would be the child saying, "*because they're not sharing*" when the examiner asked, "*what else is happening in this picture?*" In our original design, we, therefore, decided to eliminate these conjunctions from the analysis and did not judge their grammaticality. However, the samples in the current study did not include any utterances in which responses to prompts started with a conjunction.

To check the reliability of transcription, a consensus procedure (adapted from Shriberg, Kwiatkowski, & Hoffman, 1984) was used. Each sample was first transcribed by one of the trained student research assistants. Utterances that could not be fully transcribed after the research assistant listened three times were marked as unintelligible and excluded from the analysis. Transcription for the entire sample was rechecked by a second trained student research assistant and then by the first author. Any discrepancies that could not be resolved were excluded. The same consensus procedure was followed for utterance segmentation, utterance inclusion, and coding. There were no disagreements for utterance segmentation or utterance inclusion. All instances of error coding that could not be resolved were considered to be acceptable and were not coded as errors.

Analyses

PGU was calculated for all 15 pictures (i.e., total PGU) and for each of the seven-picture sets (i.e., PGU Set 1 and PGU Set 2) by subtracting the number of ungrammatical C-units from the total number of C-units and then dividing by the total number of C-units. The PGU score of 58.32% from Eisenberg and Guo (2013) was used as the standard cutoff score for decision making across picture sets. Scores at or above this cutoff were considered to be passing; scores below this cutoff were considered to be failing.

The first set of analyses compared shorter and longer samples. One-way repeated-measures analysis of variance (ANOVA) models were adopted to examine whether PGU scores differed between each of the seven-picture sets and the total sample. Pearson product-moment correlations were computed to examine the extent to which PGU scores from each seven-picture set correlated with PGU scores from the total sample. We also compared the degrees of agreement for the pass–fail decisions between each seven-picture set and the total sample.

In the second set of analyses, we compared the two shorter samples (i.e., Sets 1 and 2). One-way repeated-measures ANOVAs were adopted to examine whether PGU scores differed between Set 1 and Set 2. Pearson product-moment correlations were computed to examine the extent to which PGU scores were correlated for the two seven-picture sets. We then compared the degree of agreement for the pass–fail decisions between Set 1 and Set 2.

Results

Preliminary Analyses

Table 1 presents the total number of utterances and mean length of C-units in morphemes (MLCU-m) for each set. One-way repeated-measures ANOVAs indicated that the total number of utterances did not differ significantly between Set 1 and Set 2, $F(1, 39) = 2.03$, $p = .16$, $\eta_p^2 = .049$. However, MLCU-m was longer for the total sample than for Set 1 or Set 2, Fs > 4.24, ps < .04, $\eta_p^2 > .19$. In addition, MLCU-m was also longer for Set 1 than for Set 2, $F(1, 39) = 4.59$, $p = .04$, $\eta_p^2 = .105$. Furthermore, because the administration of Sets 1 and 2 was randomized across sessions, we wanted to know if the PGU scores varied with sessions due to the potential practice effect. The PGU score in Session 1 (M = 56.55%, SD = 22.77%) did not significantly differ from that in Session 2 (M = 56.60%, SD = 24.28%), $F(1, 39) = 0.001$, $p = .98$, $\eta_p^2 < .001$. Thus, the factor of session order is not further discussed here.

Comparisons Between PGU Scores Based on Longer and Shorter Samples

The mean PGU scores for the total sample and for each seven-picture set are presented in Table 1. We first compared PGU for the total sample to PGU for the two seven-picture sets (i.e., Sets 1 and 2). There were no significant differences between PGU from the total sample and PGU from Set 1 or Set 2, Fs < 0.59, ps > .45, $\eta_p^2 < .02$. To further explore whether the PGU scores were consistent between longer and shorter sets, we computed Pearson correlation coefficients between these measures. Table 2 presents the correlation coefficients between PGU scores by picture set. PGU for each half-set was significantly correlated with PGU for the total sample (rs > .95, ps < .01).

Table 3 presents the pass–fail decisions for each participant for the total sample and for each seven-picture set on the basis of the criterion PGU level of 58.32% obtained from Eisenberg and Guo (2013). Overall agreement between Set 1 and the total set was 88%, with 95% agreement for passing (i.e., scoring at or above a PGU of 58.23%) but only 81% agreement for failing (i.e., scoring below a PGU of 58.23%). Agreement between Set 2 and the total set was 95%, with 95% agreement for both pass and fail decisions. Disagreements for pass–fail decision between longer and shorter sets were all for children who had scored between 52% and 62% for PGU for the total sample. There were no pass–fail disagreements for children who scored below 52% and above 65%.

Comparisons Between PGU Scores for the Two Shorter Samples

The mean PGU scores for the two seven-picture sets are presented in Table 1. We first compared PGU between the two seven-picture sets. PGU for Set 1 was not significantly different from PGU for Set 2, $F(1, 39) = 0.46$, $p = .50$, $\eta_p^2 = .012$. Next, we computed Pearson correlation coefficients between PGU for Set 1 and Set 2 to further examine whether the PGU scores were consistent between the two seven-picture sets (see Table 2). PGU for Set 1 and PGU for Set 2 were significantly correlated with each other ($r = .83$, $p < .01$), although this correlation was not as strong as

Table 1. Means (standard deviations) for language sample measures and percent grammatical utterances (PGU).

Stimuli	No. pictures	No. utterances	MLCU-m	PGU
Total sample	15	66.83 (13.14)	5.72 (1.12)	57% (23%)
Set 1	7	31.80 (7.76)	5.52 (1.05)	57% (23%)
Set 2	7	29.85 (7.46)	5.25 (1.08)	56% (23%)

Note. No. utterances = total number of utterances; MLCU-m = mean length of C-units in morphemes.

the correlations between each seven-picture set and the total sample.

The pass–fail decisions for each picture set on the basis of the criterion PGU level of 58.32% obtained from Eisenberg and Guo (2013) are also included in Table 3. Overall agreement between Set 1 and Set 2 was 83%. When Set 1 was used as the reference point, there was 89% agreement for failing and 77% agreement for passing. When Set 2 was used as the reference point, there was 90% agreement for failing and 75% agreement for passing. Disagreements for pass–fail decisions between the two sets occurred only for the seven children who had scored between 42% and 65% on one of the two seven-picture sets.

Table 2. Correlation coefficients between percent grammatical utterances (PGU) scores on the basis of longer and shorter samples.

Variable	1	2	3
1. PGU-Total	—	.95**	.95**
2. PGU-Set 1		—	.83**
3. PGU-Set 2			—

**Correlation is significant at the .01 level (1-tailed).

Table 3. Diagnostic decisions on the basis of the same cutoff percent grammatical utterances (PGU) score (58.32%) by picture set and child.[a]

Child	PGU Total (15)[b]	PGU Set 1 (7)	PGU Set 2 (7)
108MP11	1	**0**	1
103OL11	1	**0**	1
111EG11	1	**0**	1
067EE10	1	**0**	1
101BM11	1	1	**0**
073CB11	1	1	1
071YV10	1	1	1
068SW10	1	1	1
072AG10	1	1	1
064JG10	1	1	1
066NK1	1	1	1
083BB10	1	1	1
084GD10	1	1	1
085NL10	1	1	1
100LW11	1	1	1
095ML11	1	1	1
107KD11	1	1	1
116NG11	1	1	1
109AM11	1	1	1
115NO11	1	1	1
069MM10	0	0	0
087MW10	0	0	0
078BW10	0	0	0
112TG11	0	0	0
110AR11	0	0	0
092GA10	0	0	0
079MH10	0	0	0
080WB10	0	0	0
077EG10	0	0	0
089CZ10	0	0	0
114JK11	0	0	0
099JS11	0	0	0
074MD10	0	0	1
081JC10	0	0	0
082JN10	0	0	0
113SA11	0	0	0
098MH11	0	0	0
091JC10	0	0	0
065JM10	0	1	0
102HL11	0	0	0
% agreement with PGU Total		88%	95%

[a] In Columns 2 to 4, the number "0" indicates the child was classified as failing on the basis of scoring below the 58.32% criterion, whereas the number "1" indicates that the child was classified as passing on the basis of scoring above the 58.32% cutoff. Boldfaced areas mark the discrepancy in the classifications using the full set of pictures and using a half set of pictures or pictures of a given type.

[b] The number within parentheses indicates the number of pictures for PGU computation.

Discussion

The current study investigated the utility of using smaller samples, on the basis of the number of pictures used to elicit the language samples, to calculate PGU. Overall, we found that PGU did not change significantly when fewer (i.e., seven) pictures were used, and there was no difference in PGU between longer and shorter samples or between the two shorter samples. In addition, PGU for each of the seven-picture sets was highly correlated with PGU for the longer 15-picture set and with each other. Taken together, these findings suggest that PGU scores are not affected by reducing the sample size (i.e., number of pictures), possibly because rate-based measures are less influenced by sample size than are token-based measures (e.g., Owen & Leonard, 2002).

To further evaluate the consistency between the shorter and longer samples, we also compared pass–fail decisions (i.e., children scoring above or below a previously established PGU cutoff score of 58.23%) between these samples. Overall agreement for these decisions was moderate (i.e., between 80% and 89%) to good (i.e., between 90% and 100%). Disagreements for pass–fail decisions were not randomly distributed throughout the PGU score range. Rather, all pass–fail disagreements between longer and shorter samples and between the two shorter samples clustered within the PGU range of 42% and 65%.

One of the seven-picture sets (Set 2) showed good agreement (i.e., 90% or better) with the total set for both passing and failing decisions on the basis of PGU scores. The other seven-picture set (Set 1), however, showed a moderate level of agreement (i.e., 80% to 89%) with the total sample as well as with Set 2 for failing decisions. That is, more children scored below the PGU cutoff score on Set 1 than on either Set 2 or the total sample.

We considered several possible factors to explain why the failure rate for PGU for Set 1 was higher than the failure rate for the total set and for Set 2. One possibility was utterance length. We reasoned that children might produce more errors in longer utterances than in shorter utterances. The preliminary analysis showed that MLCU-m was higher for Set 1 than for Set 2. However, MLCU-m was also higher for the total sample than for either of the smaller samples. In addition, children who scored below the PGU cutoff varied in their MLCU-m and did not cluster at either the high or low MLCU-m range. It, therefore, did not seem likely that utterance length would account for the difference in failure rates between sets.

A second possibility was the number of C-units produced. We reasoned that children might produce proportionally more errors in shorter samples than in longer samples. However, the preliminary analysis showed that there was no difference in the number of C-units produced on the two seven-picture sets. There was also no difference in PGU between the shorter seven-picture samples and the longer sample of 15 pictures. Additionally, children who scored below the PGU cutoff varied in their sample length and did not cluster at either the high or low range for number of C-units. Number of C-units, therefore, also did not account for the difference in failure rates between sets.

A third possibility was the type of pictures included in each set. In a study by Shapiro and Hudson (1991), preschool children responded differently to picture sequences showing a problem and resolution than to picture sequences that depicted a series of events that were not causally related. The problem-based picture sequences yielded not only better stories but also more varied and complex language forms than the action sequences. We reasoned that a similar effect on utterance form might have occurred in response to single pictures and that this might affect not only utterance complexity but also grammaticality. That is, it may be that problem-based pictures promoted use of more complex language forms and that these were more likely to contain errors.

To explore this, we did a post hoc analysis for picture type. We administered a Likert-type scale for picture type to 18 master's students majoring in speech-language pathology. We classified pictures as problem-based if they were rated as definitely or probably a problem picture by at least 80% of respondents and classified pictures as action-based if they were rated as definitely or probably an action picture by at least 80% of respondents. On the basis of these ratings, five of the pictures were classified as problem-based, and five were classified as action-based. The remaining five pictures were considered to be nonclassifiable.

We then checked the pictures included in each of the seven picture sets based on these ratings. Set 1 included three problem-based pictures, three action pictures, and one unclassifiable picture. Set 2 included one problem-based picture, two action pictures, and three unclassifiable pictures. The distribution of picture types was, thus, quite different between sets, with Set 1 including more problem-based pictures than Set 2. In spite of this, PGU did not differ significantly between sets. That is, the higher number of problem-based pictures in Set 1 did not result in more grammatical errors or lower PGU overall. Inspections of the individual data indicated that although 18 of the children had a lower PGU on Set 1 than Set 2, another 18 children had a higher PGU on Set 1, and four children had the same PGU on both sets. Picture type did not, therefore, seem to account for the difference in failure rates between picture sets. However, the variability in picture type between sets may have affected the internal consistency of the language sampling task and, therefore, may have contributed to measurement error. Further studies are needed in order to determine whether and to what extent picture type might affect PGU scores.

Clinical Implications

Grammaticality, as measured by PGU, appears to be unaffected by reducing the sample size (i.e., number of pictures). Therefore, smaller samples can be used, shortening the time for administration, transcription, and scoring. To measure children's performance on PGU, we recommend using an activity, such as having a child talk about pictures in response to prompts, which obligates complete sentences with both a subject and a predicate. In the current study, we used sets of seven pictures that took seven to 20 min, including reinforcement time, to administer.

For the current study, we used a previously determined PGU cutoff score of 58.23% for making pass–fail decisions. That cutoff score was based on children who were previously identified as having either LI or typical language. The disagreement pattern in the current study suggests that applying that cutoff to make decisions about children who are not previously identified, as would be the case in clinical assessments, could result in both over- and underidentifying LI. In the current study, pass–fail disagreements between picture sets clustered in the PGU range between 42% and 65%, with no disagreements above or below this range. This suggests that we can have confidence in passing children whose PGU score is above 65% and in failing children whose PGU score is below 42%. However, PGU scores between 42% and 65% would warrant using additional pictures and longer samples for calculating PGU.

We are not suggesting that PGU should be used as the only measure in making diagnostic decisions. Instead, PGU should be used as one piece of evidence in the assessment battery that provides supplementary information in addition to standardized tests and other language sample analyses.

Strengths and Limitations of the Current Study

Consistent with other studies investigating sample size for language analyses, we compared the mean PGU scores between samples and calculated reliability correlations. To the best of our knowledge, the present study is the only investigation that examines the effect of sample size on grammaticality measures (i.e., PGU) and considers its impact on clinical decision making.

Our previous study (Eisenberg & Guo, 2013) about the diagnostic accuracy of PGU found high sensitivity and moderate specificity for differentiating between typical language and LI using a 58.23% cutoff score. However, that result was based on pre-identified groups of children, as is common for studies of diagnostic accuracy (Pawlowska, 2014; Souto et al., 2014) and which inflates group differentiation accuracy rates (Goodwin & Leech, 2006; Pawlowska, 2014). In contrast, in the current study, we included children with varying language abilities and were masked to the language status of participants. This allowed us to better evaluate the interpretation of PGU scores.

A limitation of the current study is that we used two fixed sets of seven pictures each with pictures randomly assigned to each set. However, further inspection revealed that the composition of the picture sets was not the same. Although the PGU scores in Set 1 and Set 2 were not significantly different, Set 1 yielded more children who scored below the cutoff scores than the total sample. Thus, we cannot be certain that picture selection will not influence the child's PGU score and the use of PGU in clinical decision making. Further studies that manipulate picture type and use different pictures are needed to investigate whether clinicians can use varied pictures for PGU or whether the pictures need to be standardized. Note, however, that the issue of a possible impact of stimulus materials is not unique to picture-elicited language samples. Miller (1981), for instance, raised a similar concern about the potential impact on MLU of the toys used for eliciting language samples during play.

Another limitation was that the SES range of the participants was not fully representative. This is important because SES has been shown to influence performance on language assessments. For instance, M. W. Lee et al. (2008) reported that a group of children for whom only 5% of mothers completed college scored significantly lower on a standardized language test than a group of children whose mothers all had college degrees. Thus, we cannot be certain that the recommended PGU interpretation is applicable to children from lower SES backgrounds.

Conclusion

PGU scores from picture-elicited language samples have been found to be a sensitive tool in identifying children with LI (Eisenberg & Guo, 2013). In this study, we investigated whether a shortened version (i.e., seven pictures) of this language sample collection procedure would yield a PGU score similar to the full version of the task (i.e., 15 pictures). PGU scores for the shortened version were similar to, and were correlated with, those for the full version, suggesting that PGU scores were not affected by reducing the number of pictures. Taken together, the results showed that PGU scores can be computed using seven pictures, at least for 3-year-old children. However, the results also suggest caution in interpreting PGU scores relative to the previously determined cutoff when different pictures are used.

Acknowledgments

This project was supported by National Institute on Deafness and Other Communication Disorders (NIDCD) Grant R21DC009218, awarded to the first author, and by the Language Learning Research Grant (Blackwell Publishing), awarded to the second author. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIDCD, National Institutes of Health, or Blackwell Publishing. The authors 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, transcribed, and analyzed the samples. Portions of this study were presented at the 2013 convention of the American Speech-Language-Hearing Association in Chicago, Illinois.

References

Bank Street College of Education. (1968a). *"Give me my pail!"* New York, NY: MacMillan.

Bank Street College of Education. (1968b). *The cat in the tree.* New York, NY: MacMillan.

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.

Berman, R. A., & Slobin, D. I. (1994). *Relating events in narrative: A cross-linguistic developmental study*. Hillsdale, NJ: Erlbaum.

Bogue, E., DeThorne, L., & Schaefer, B. (2014). A psychometric analysis of childhood vocabulary tests. *Contemporary Issues in Communication Science and Disorders, 41,* 55–69.

Brookshire, R. H., & Nicholas, L. E. (1994). Speech sample size and test–retest stability of connected speech measures for adults with aphasia. *Journal of Speech and Hearing Research, 37,* 399–407.

Burns-Hoffman, R. (1993, March). *Scaffolding children's informal expository skills.* Paper presented at the Biennial Meeting of the Society for Research in Child Development, New Orleans, LA, retrieved from the ERIC database (ED362292).

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 and Therapy, 27,* 286–293.

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 samples. *Journal of Speech and Hearing Research, 3,* 166–173.

Dawson, J., Stout, C., Eyer, J., Tattersall, P., Fonkalsrud, J., & Croley, K. (2005). *Structured Photographic Expressive Language Test–Preschool 2 (SPELT-P2)*. DeKalb, IL: Janelle Publications.

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., & Guo, L. (2010, June). *How grammatical are 3-year-olds*? Paper presented at the Symposium on Research in Child Language Disorders, University of Wisconsin–Madison.

Eisenberg, S., & Guo, L. (2013). Differentiating children with and without language impairment based on grammaticality. *Language, Speech, and Hearing Services in Schools, 44,* 20–31.

Eisenberg, S., Guo, L., & Germezi, M. (2012). How grammatical are three-year-olds? *Language, Speech, and Hearing Services in Schools, 43,* 36–52.

Eisenberg, S. L., Ukrainetz, T. A., Hsu, J. R., Kadaverak, J. N., Justice, L. M., & Gillam, R. B. (2008). Noun phrase elaboration in children's stories. *Language, Speech, and Hearing Services in Schools, 39,* 145–157.

Fey, M. E., Catts, H. W., Proctor-Williams, K., Tomblin, J. B., & Zhang, X. (2004). Oral and written story composition skills of children with language impairment. *Journal of Speech, Language, and Hearing Research, 47,* 1301–1319.

Fluharty, N. B. (2001). *Fluharty Preschool Speech and Language Screening Test* (2nd ed.). 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.

Goodglass, H., Kaplan, E., & Barresi, B. (2000). *Boston Diagnostic Aphasia Examination* (3rd ed.). Austin, TX: Pro-Ed.

Goodwin, L. D., & Leech, N. L. (2006). Understanding correlation: Factors that affect the size of *r*. *The Journal of Experimental Education, 74,* 251–266.

Grela, B., & Leonard, L. (1997). The use of subject arguments by children with specific language impairment. *Clinical Linguistics & Phonetics, 11*(6), 443–453.

Guo, L., & Schneider, P. (2014, June). Differentiating school-aged children with and without language impairment using grammaticality measures from a narrative task. Poster presented at The Symposium on Research in Child Language Disorders, University of Wisconsin, Madison, WI.

Halliday, M. A. K. (1994). *Introduction to functional grammar* (2nd ed.). London, United Kingdom: Edward Arnold.

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.

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., Miller, J. F., Curtiss, S., & Tallal, P. (1993). Conversations with children who are language impaired: Asking questions. *Journal of Speech and Hearing Research, 36,* 973–978.

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

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

Lee, M. W., Guillot, K., Spencer, E., Arndt, K. B., Rosenthal, J., Lineback, A., . . . Schuele, C. M. (2008, November). SES influences on preschoolers' performance on the Preschool Language Scale and the Peabody Picture Vocabulary Test. Paper presented at the Convention of the American Speech-Language-Hearing Association, Chicago, IL.

Leonard, L. B. (1998). *Children with specific language impairment,* Cambridge, MA: MIT Press.

Leonard, L. B., Bolders, J. G., & Miller, J. A. (1976). An examination of the semantic relations reflected in the language usage of normal and language-disordered children. *Journal of Speech and Hearing Research, 19,* 371–392.

Leonard, L. B., Camarata, S. M., Brown, B., & Camarata, M. N. (2004). Tense and agreement in the speech of children with specific language impairment: Patterns of generalization through intervention. *Journal of Speech, Language, and Hearing Research, 47,* 1363–1379.

Leonard, L. B., Eyer, J. A., Bedore, L. M., & Grela, B. G. (1997). Three accounts of the grammatical morpheme difficulties of English-speaking children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 40,* 741–753.

Loban, W. D. (1963). *The language of elementary school children*. Research Report No. 1. Urbana, IL: National Council of Teachers of English.

Loeb, D., & Leonard, L. (1991). Subject case marking and verb morphology in normally developing and specifically language-impaired children. *Journal of Speech and Hearing Research, 34,* 340–346.

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.

Merritt, D. D., & Liles, B. Z. (1989). Narrative analysis: Clinical application of story generation and story retelling. *Journal of Speech and Hearing Disorders, 54,* 429–438.

Miller, J. (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: SALT Software.

Miller, J., & Iglesias, A. (2010). *Systematic analysis of language transcripts (SALT), Research version 2010* [Computer software]. Madison, WI: SALT Software.

Moore, M. E. (2001). Third person pronoun errors by children with and without language impairment. *Journal of Communication Disorders, 34,* 207–228.

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.

Paul, R., & Norbury, C. F. (2012). *Language disorders from infancy through adolescence* (4th ed.). St Louis, MO: Mosby Elsevier.

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

Polite, E. J., Leonard, L. B., & Roberts, F. D. (2011). The use of definite and indefinite articles by children with specific language impairment. *International Journal of Speech-Language Pathology, 13,* 291–300.

Reynolds, C. R., & Kamphaus, R. W. (2003). *Reynolds' Intellectual Screening Test (RIST)*. 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.

Rice, M. L., Wexler, K., & Cleave, P. L. (1995). Specific language impairment as a period of extended optional infinitive. *Journal of Speech and Hearing Research, 38,* 850–863.

Robinson, H. M., Monroe, M., & Artley, A. S. (1962a). *Before we read*. Glenview, IL: Scott Foresman.

Robinson, H. M., Monroe, M., & Artley, A. S. (1962b). *We read pictures*. Glenview, IL: Scott Foresman.

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.

Sackett, D. (1991). *Clinical epidemiology: A basic science for clinical medicine* (2nd ed.). Boston, MA: Little Brown.

Saeed, J. (2009). *Semantics* (3rd ed.). Malden, MA: Wiley-Blackwell.

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

Semel, E., Wiig, E., & Secord, W. (2003). *Clinical Evaluation of Language Fundamentals* (4th ed.; CELF-4). San Antonio, TX: Pearson.

Shapiro, L. R., & Hudson, J. A. (1991). Tell me a make-believe story: Coherence and cohesion in young children's picture-elicited narratives. *Developmental Psychology, 27,* 960–974.

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.

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

State of New Jersey Department of Education. (2010). Core curriculum content standards. Retrieved from http://www.state.nj.us/education/cccs/

Ukrainetz McFadden, T. (1996). Creating language impairments in typically achieving children: The pitfalls of "normal" normative sampling. *Language Speech and Hearing Services in Schools, 27,* 3–9.

Watkins, R., & Rice, M. (1991). Verb particle and preposition acquisition in language-impaired preschoolers. *Journal of Speech and Hearing Research, 34,* 1130–1141.

Westerveld, M. F., & Gillon, G. T. (2010). Profiling oral narrative ability in young school-aged children. *International Journal of Speech-Language Pathology, 12,* 178–189.

Ziefert, H. (1987). *Jason's bus ride*. New York, NY: Puffin Books.

Appendix A

Prompts for the Picture Task

The following were used as prompts (or alternative prompt if the child did not respond, responded with "*I don't know*," or produced an off-topic utterance; Eisenberg & Guo, 2013).

1. What is happening in the picture? (PROMPT: POINT TO DIFFERENT PARTS OF PICTURE AND SAY: "Just tell something about the picture.")

2. What *else* is happening in the picture? (PROMPT: POINT TO DIFFERENT PARTS OF THE PICTURE AND SAY: "Tell me something else about the picture.")

3. SAY "NOW I'LL START THE STORY AND YOU FINISH IT." PROVIDE STORY STARTER AND SAY: "And then ~" (PROMPT: REPEAT STORY STARTER AND SAY: "And then what happens in the story?")

4. Tell me one more thing about the story. (PROMPT: POINT OUT PARTS OF THE PICTURE THE CHILD HAS NOT TALKED ABOUT AND SAY: "Just tell me anything else about the picture.")

Story starters for prompt 3:

- COOKIE: The boy is trying to get the cookies and then ~
- CAKE: Oh no! The dog ate some of the cake and then ~
- BUS: The dog is in front of the bus and the bus can't move and then ~
- SANDBOX: The boys are fighting and here comes the mom and then ~
- CAT: The children see the cat. The cat is stuck up in the tree and then ~
- DOLL: The daddy is hiding a doll behind his back and then ~
- DONUT: The children are taking the donuts from the bag and then ~
- SUPERMARKET: The boy knocked the boxes off the shelf and then ~
- WASHING: The children are trying to wash the dog and then ~
- BUBBLES: Oh no, the bubbles spilled and then ~
- SNOW: The boys are throwing snowballs and then ~
- DRESSING: The dog and the girl have the daddy's shoes and then ~
- BREAKFAST: The little girl is still in her pajamas and then ~
- LEAVES: They raked the leaves into a big pile and then ~
- SCISSORS: The boy is taking his grandma's scissors and then ~

Appendix B

Description of Pictures Used in the Task by Set

Not Used for Computing PGU for Set 1 or Set 2

BUS: dog in front of bus; man and woman pulling on dog's leash; driver and boy in front of bus (Ziefert, 1987, pp. 19–20).

Set 1

1. BREAKFAST: table with bowls, glasses, spoons, and cereal box; woman pouring juice into a glass; girl holding spoon; another girl in pajamas yawning and rubbing her eyes; man putting bread in toaster.

2. CAKE: partially eaten cake on table with footprints leading to dog under couch; woman holding a broom and boy crying, both facing the couch; woman with boy holding present and woman with girl holding present entering the room through an open door.

3. CAT: cat in tree; boy and girl looking up at the tree (Bank Street College of Education, 1968b, pp. 1–2).

4. DRESSING: man looking in mirror and buttoning his shirt; boy holding a tie; girl with a man's shoe on each hand; another girl with man's hat on her head; dog with slipper in his mouth (Robinson, Monroe, & Artley, 1962b, p. 23).

5. LEAVES: girl putting leaves in bag; another girl and a boy playing in the leaves; man and boy raking leaves; another boy holding a rake and waving at woman in car; dog barking at squirrel in tree.

6. SANDBOX: two boys in sandbox tugging at a bucket; another bucket in the sand; woman running over to the sandbox (Bank Street College of Education, 1968a, pp. 3–4).

7. WASHING DOG: dog jumping out of tub and splashing a girl; boy with a bucket of water (Robinson, Monroe, & Artley, 1962b, p. 2).

Set 2

1. BUBBLES: boy leaning over wagon; girl tripping over his feet with bubble jar and wand flying out of her hand; bucket of water spilling over; another girl sitting down holding bubble jar and wand.

2. COOKIE: boy on stool that is tipping over, reaching for cookie in cookie jar on shelf and holding cookie in other hand; girl reaching for cookie; woman standing by overflowing sink drying a plate (Goodglass, Kaplan, & Barresi, 2000).

3. DOLL: girl running out of house towards man; man kneeling down with one arm reaching towards girl and the other holding a doll behind his back.

4. DONUT: woman at wheel of car; donut bag on front seat; boy in front seat handing donut to girl in back seat; both children looking at woman with laughing expressions on their faces.

5. SCISSORS: woman sitting in chair and knitting; boy on his hands and knees taking scissors out of basket at side of chair; girl and man kneeling behind the chair, man with finger to lips.

6. SNOW: man and girl building a snowman; another girl shoveling snow; two boys throwing snowballs at each other.

7. SUPERMARKET: boy reaching for box on top shelf and several boxes falling down; girl reaching for fruit in bin; another girl with shopping cart looking at her and about to push cart into the boy; woman with list in her hand looking at the shelves of boxes (Robinson, Monroe, & Artley, 1962a, p. 23).