AJSLP

Research Article

Eliciting the Language Sample for Developmental Sentence Scoring: A Comparison of Play With Toys and Elicited Picture Description

Sarita L. Eisenberg,[a] Ling-Yu Guo,[b,c] and Emily Mucchetti[d]

Purpose: This study investigated whether language samples elicited during play and description of pictured events would yield the same results for developmental sentence scoring (DSS).
Method: Two language samples were elicited from 58 three-year-olds. One sample was elicited during play with a parent, and the other sample was elicited by an examiner asking children to talk about pictured events in response to elicitation questions.
Results: DSS scores were not significantly different between the play and event description samples. However, sentence points were significantly higher for the play sample than for the event description sample. Although there was a correlation between sample types for both DSS and sentence points, the correlation for DSS (r = .52) was below an acceptable level, and the correlation for sentence points (r = .71) was at a minimally acceptable level. Agreement between sample types for pass–fail decisions on the DSS scores using the 10th percentile cutoff recommended by Lee (1974) was only moderate (78%).
Conclusion: The current study shows that type of language samples could affect DSS and sentence point scores of 3-year-olds and, hence, the passing and failing decisions for their performance on DSS.

Language sample analysis (LSA) is a widely used means of assessing language disorders in children (Caesar & Kohler, 2009; Hux, Morris-Friehe, & Sanger, 1993; Kemp & Klee, 1997; Pavelko, Owens, Ireland, & Hahs-Vaughn, 2016; Westerveld & Claessen, 2014). The proportion of speech-language pathologists (SLPs) reporting that they use LSA ranged from 94% in the Caesar and Kohler (2009) survey, 91% in Westerveld and Claessen (2014), 85% in Kemp and Klee (1997), and 67% reported by Pavelko et al. (2016). Previous research suggests that a diagnosis of language impairment (LI) in young children may be more accurately accomplished through the use of quantitative LSA measures than through standardized tests (Dunn, Flax, Sliwinski, & Aram, 1996; Rescorla, Roberts, & Dahlsgaard, 1997). The results of several surveys confirm that SLPs use LSA for this purpose. Of the SLPs using LSA, 92% of respondents in Kemp and Klee (1997), 87% of the respondents in Pavelko et al. (2016), and 79% of the respondents in Westerveld and Claessen (2014) reported using LSA for initial diagnosis of LI. The most frequently reported LSA procedure for diagnosing LI, other than mean length of utterance (MLU), was developmental sentence scoring (DSS; Hux et al., 1993; Kemp & Klee, 1997), by one third of respondents.

DSS (Lee, 1974) is a norm-referenced LSA measure for children aged 2;0 (years;months) to 6;11. The DSS analysis is based on 50 complete and unique utterances. *Completeness* means that utterances must have both a subject and a verb; utterances that do not have both a subject and verb are eliminated from the DSS analysis. *Uniqueness* means that all utterances must be different; subsequent productions of the same utterance are excluded from the DSS analysis. Each utterance is rated for the use of forms within eight grammatical categories: indefinite pronouns and noun modifiers, personal pronouns, primary verbs, secondary verbs, negatives, conjunctions, interrogative reversals, and *wh*-questions. Within each of these categories, forms either receive weighted scores for correct use, with

[a]Department of Communication Sciences and Disorders, Montclair State University, NJ
[b]University at Buffalo–The State University of New York
[c]Asia University, Taichung, Taiwan
[d]Easterseals Delaware & Maryland's Eastern Shore, Dover
Correspondence to Sarita Eisenberg: eisenbergs@montclair.edu

Editor: Krista Wilkinson
Associate Editor: Cynthia Cress
Received September 20, 2016
Revision received March 27, 2017
Accepted October 23, 2017
https://doi.org/10.1044/2017_AJSLP-16-0161

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

higher scores given to later developing forms, or receive attempt marks for incorrect use. An additional sentence point is awarded to utterances that are grammatically and semantically correct.

Lee (1974) developed DSS as a diagnostic tool for idenifying children with LI, and it continues to be recommended and used for this purpose (e.g., Fey, Cleave, Long, & Hughes, 1993; Hughes, Fey, & Long, 1992; Leonard, Camarata, Brown, & Camarata, 2004; Paul, Hernandez, Taylor, & Johnston, 1996; Paul & Norbury, 2014; Souto, Leonard, & Deevy, 2014). DSS has also been used for selecting therapy goals (Hughes et al., 1992) and for measuring progress in therapy (Feyet al., 1993; Friedman & Friedman, 1980; Leonard, Camarata, Pawtowska, Brown, & Camarata, 2006; Loeb, Stoke, & Fey, 2001). In this article, we focused on the use of DSS for making decisions about whether or not a child is meeting developmental expectations on the basis of language samples elicited during different tasks. Specifically, we examined the potential influence of sample types on the DSS scores and sentence points in the DSS analysis and, hence, on passing and failing decisions.

Language Sample Elicitation for DSS

Conversational sampling during play-based activities is generally considered the most appropriate method of eliciting a language sample from young children (Miller, 1981; Retherford, 1993). This is the sampling method used for MLU (Miller & Chapman, 1981; Rice et al., 2010) and Index of Productive Syntax (Scarborough, 1990). In recent surveys, conversation was the most frequently used type of activity for eliciting language samples from preschool children (Pavelko et al., 2016; Westerveld & Claessen, 2014), used by 95% of SLPs who reported using LSA in the Pavelko et al. (2016) survey. However, this may not be the most appropriate type of sampling for other LSA measures and for DSS in particular because the norms for DSS were not solely based on conversational samples (Koenigsknecht, 1974).

The guidelines for eliciting the language sample for DSS are somewhat open-ended. The instructions suggest using three types of stimulus materials, presented one at a time in a fixed order—toys followed by action pictures followed by a pictured story (Lee, 1974). However, Lee advocated flexibility in presenting these stimulus materials to different age groups. She reported that 2-year-old children and some 3-year-olds talked most during play with the toys, other 3-year-olds and 4-year-old children talked more about the action pictures, and 5- and 6-year-old children talked more when telling a familiar pictured story.

Lee (1974) hypothesized that the latter two activities—talking about pictures and storytelling—would yield more sophisticated language forms and earn more points on the DSS. She, therefore, used the last 50 utterances of the language sample for the normative study to ensure that utterances from these sample types would be included in the DSS analysis. However, this assumption has not been proven, and Lee advocated flexibility as well in selecting the "best" part of the corpus as long as utterances were consecutive. Accordingly, DSS has been based on a variety of sample types. DSS has sometimes been based on picture description (Allen, Kertoy, Sherblom, & Pettit, 1994; Mortimer & Rvachew, 2010) or on narrative samples (Kemper, Rice, & Chen, 1995; Loeb et al., 2001), the two activities that Lee predicted would yield more-sophisticated utterances. However, it also seems to be common for DSS to be based solely on conversational samples elicited during play (Fey et al., 1993; Friedman & Friedman, 1980; Holdgrafer, 1995; Leonard et al., 2006; Loeb, Pye, Richardson, & Redmond, 1998; Paul et al., 1996; Souto et al., 2014; Washington, 2013). Interestingly, we found no study that based DSS on a combination of sample types as recommended by Lee (1974).

This results in a lack of standardization of the procedure for eliciting language samples for DSS. It may also result in a lack of congruity between the elicitation procedure used clinically and the procedure used for developing the norms. This may be particularly problematic for 3-year-old children, who responded variably to the stimuli in the standardization studies. The current study, therefore, focused on the influence of elicitation materials and procedures on the DSS performance of 3-year-old children. In what follows, we first reviewed studies that evaluated the impact of sampling context on language production in general and then reviewed studies that specifically evaluated the impact on DSS.

Methods for Eliciting Language Samples

The terminology for characterizing language elicitation activities has varied across studies, making it difficult to compare results. Before we review studies, we will, therefore, first describe the types of tasks that have been investigated among the studies and apply a uniform label to each task.

Play With Toys

Although toys have been used as the stimulus materials in a number of studies, how the toys were used varied. In free play (e.g., Evans & Craig, 1992; Klein, Moses, & Jean-Baptiste, 2010; Southwood & Russell, 2004), the child was free to choose which toys to play with and how to engage with the toys. The adult followed the child's lead, commenting on the child's actions to encourage talking but limiting questions and conversational demands. Klein et al. (2010) used a variation on play, which they called *scripted play,* in which the adult set a topic on the basis of common life experiences, such as a visit to the doctor or to McDonald's, but the child was otherwise free to engage with the toys and talk. Other studies have included a condition that was called *play* but that involved the child telling a story while manipulating props and/or puppets (Sealey & Gilmore, 2008; Wren, 1985). We classified this latter elicitation condition as *event casting* rather than as *play* (see below).

Elicited Description

Description tasks focused on single objects or events and involved prompting the child with specific questions

and comments rather than following the child's lead. Object descriptions were elicited about toys in response to prompts, such as "tell me all about this" (e.g., James & Button, 1978; Longhurst & File, 1977), or about pictures of objects in response to a prompt, such as "tell me about this picture" (e.g., Longhurst & File, 1977). Event descriptions were elicited about pictures of two or more people engaged in actions in response to the prompt "tell me what's happening in the picture" (e.g., Atkins & Cartwright, 1982; Longhurst & File, 1977). Atkins and Cartwright (1982) also used additional prompts that were specific to the content of each picture.

Interviewing

Evans and Craig (1992) introduced topics, such as school, family, or free-time activities, for the child to talk about. The topics were introduced with questions and comments without accompanying pictures or props. Social continuants, such as "uh-huh" and "I see," and prompts, such as "tell me more," were used after the initial topic introduction to encourage the child to continue talking about a topic. Atkins and Cartwright (1982) referred to this sampling condition as *response to imperative requests* (pp. 33). In other studies, this type of elicitation was referred to as *conversation* (Fields & Ashmore, 1980; James & Button, 1978; Longhurst & File, 1977; Southwood & Russell, 2004).

Narrative Elicitation

Several different procedures have been used to elicit narrative samples. In story retelling, the child told a story immediately after the examiner's presentation of the story (Atkins & Cartwright, 1982; Sealey & Gilmore, 2008). For story generation, the child told either a familiar or a new story while looking at pictures (Sealey & Gilmore, 2008). Southwood and Russell (2004) used modeled storytelling for eliciting personal experience stories. These authors first modeled a story and then prompted the child to tell a story by asking whether anything similar had ever happened to the child.

Event Casting

Unlike narratives, event casts are told as the events unfold rather than afterwards (Heath, 1986). This condition involved having the child tell a story while manipulating props and/or puppets (Sealey & Gilmore, 2008; Wren, 1985). Klein et al. (2010) used an alternative procedure of prompting children to describe a series of actions as they were performed by the examiner.

The Impact of Elicitation Condition on Language Production

Many of the studies investigating the effect of language sampling condition included a play condition (Evans & Craig, 1992; James & Button, 1978; Klein et al., 2010; Longhurst & File, 1977; Sealey & Gilmore, 2008; Southwood & Russell, 2004). However, we found no studies that compared language production between play with toys and elicited event descriptions, the two activities that yielded the most talking from 3-year-old children in the DSS normative study (Koenigsknecht, 1974). Across all of the studies, children produced the most talking (i.e., a larger number of utterances) during play than in other activities, but MLU was lowest in the play condition (Sealey & Gilmore, 2008; Southwood & Russell, 2004). Southwood and Russell (2004) speculated that the lack of opportunity to produce elliptical utterances in their story generation task might have contributed to a higher MLU in this condition relative to play or interviewing. This is important because a higher number of elliptical responses in the free play and interview conditions would reduce the proportion of utterances meeting the DSS inclusion criteria because utterances without a subject are eliminated from the DSS analysis. Therefore, although play might yield the largest number of total utterances, this activity might yield fewer scorable utterances for the DSS analysis than other methods of language sampling activities.

Utterance complexity is important because it affects the opportunities for a child to earn higher point values on the DSS. Studies reported contradictory results about language complexity across sampling activities. Klein et al. (2010) compared children's production of complex sentences during free play, scripted play, event casting about actions performed by the examiner, and story retelling. Participants included four children each at ages 2, 3, and 4 years. Although it may not have been the condition with the most complex sentences, the free play condition yielded a high proportion of complex sentences from all three age groups. The event casting condition yielded the lowest proportion of complex sentences for all three age groups. The proportion of complex sentences was comparable between the two play conditions—free play and scripted play—for both the 2-year-old group and the 4-year-old group (13%–14% of utterances at age 2 years and 24%–27% at age 4 years), although the highest proportion of complex sentences for 4-year-olds was in story retelling (37% of utterances). For the 3-year-olds, free play yielded a substantially higher proportion of complex sentences (41%) than did either scripted play (27%) or story retelling (21%). This result contrasts with that of Southwood and Russell (2004), who reported a lower proportion of more sophisticated language forms during free play than for either interviewing or story generation. This latter study included somewhat older participants, 5-year-old children, suggesting that the effect of elicitation condition on utterance complexity may change with age.

Children earn points on the DSS only for correct productions. Production of errors, therefore, affects the DSS score because ungrammatical productions receive attempt marks rather than earning points. Studies have reported contradictory results about production of errors across elicitation conditions. Consistent with what Koenigsknecht (1974) reported for the DSS normative study, Southwood and Russell (2004) found no difference in number of errors produced by 5-year-old children for play, interview, and

story generation. In contrast, Sealey and Gilmore (2008) observed a difference in error rates for verb tense morphemes when comparing free play with toys to two narrative tasks—story generation about a wordless picture book and story retelling—and event casting. This study included 10 children ages 3;11 to 5;6, five identified as having LI and five with typical language (TL). The TL group had the lowest error rate for verb tense morphemes in the story retelling task among the four sampling conditions, whereas the children with LI had the lowest error rate for verb tense morphemes in the free play condition. This suggests that children with LI might earn more points on primary verbs during free play than for narrative tasks because the free play condition had the lowest error rate for verb tense errors, which, in turn, could result in a higher DSS score in the play condition. Thus, not only was there a difference between conditions, but there was a difference in how the group with LI and the TL group responded to the different elicitation conditions.

These studies suggest that elicitation condition will affect a child's ability to earn points on the DSS. Specifically, free play may yield more high scoring forms for 3-year-old children (Klein et al., 2010) and more points on main verbs (Sealey & Gilmore; 2008) for children with LI than story generation. However, there was no study that compared play to event description, so we lack data on how language performance on these two tasks compares.

Studies Investigating Elicitation Conditions for DSS

Several studies reported no difference in DSS scores among sampling conditions. Koenigsknecht (1974) found no difference in DSS scores between event description in response to pictures and story generation for a familiar story for 10 participants between the ages of 4;0 and 5;6. Evans and Craig (1992) reported no difference in DSS scores between play and interview for 10 children with LI, aged 8;1 to 9;2. James and Button (1978) also found no significant difference in DSS scores for seven children with LI, aged 4;11 to 9;2, between three conditions: object descriptions about unfamiliar toys, object descriptions about familiar toys brought from home, and an interview about school, home, and recent activities.

In contrast, other studies have reported differences in DSS scores for different sampling activities. Longhurst and File (1977) compared DSS scores for four activities: object description about toys, object description for single-object pictures, event description in response to pictures, and an interview. Participants were 20 children between the ages of 3;11 and 5;0 in a Head Start Program. DSS scores were highest for the interview condition and were also higher for object description about toys than for either object or event description about pictures. Percentile ranks for individual children were highly variable across tasks. Although all of the children scored within the normal range for the interview and object description about toys, three of the children scored below the 10th percentile for the object description about pictures, and four other children scored below the 10th percentile on the event description task. Thus, the variability in DSS scores across tasks affected pass–fail decisions about some of the children.

Fields and Ashmore (1980) compared the DSS performance of 4-, 5-, and 6-year-old children with and without LI in three conditions: interview, talking about pictures with an event description or story generation, and a telemetry condition in which a language sample was recorded in the child's home. Consistent with Longhurst and File's (1977) findings, DSS scores were highest for the interview samples for both groups of children among the three sample types.

Together, these studies (Fields & Ashmore, 1980; Longhurst & File, 1977) suggest that DSS score may vary across elicitation conditions and that these score fluctuations could affect clinical decisions about passing or failing. However, none of the studies compared play and event description, and none of the studies included children younger than age 3;11.

The Effect of Interactant on Conversational Language Samples

In the normative studies for DSS, all samples were elicited by an examiner. However, in clinical practice, the interactant for eliciting conversational samples varies, with some samples elicited during interactions with parents or another primary caregiver who is familiar to the child (Fey et al., 1993; Ryan, 2000). We found only one study that investigated the impact of interactant on DSS performance. Kramer, James, and Saxman (1979) compared language samples elicited by mothers at home with samples elicited by an unfamiliar examiner in a clinic setting. Participants included 10 children between the ages of 3 and 5 years, with MLU between 2.5 and 5.0, who had been referred for a speech and language evaluation. Although six of the 10 children produced higher DSS scores in the home condition with their mother, the difference between home and clinic samples was not significant.

In addition, several other studies also reported no difference in MLU when setting was controlled. Olswang and Carpenter (1978) compared samples elicited by the mother and by an unfamiliar examiner in a clinic setting for nine 3- to 6-year-old children with language impairment. Hannson, Nettelbladt, and Nilhom (2000) also compared clinic samples for five 5-year-old children with specific language impairment and five children with phonological impairment. These studies uniformly found that children's MLU did not differ significantly when they interacted with their mother or with an unfamiliar examiner, although the children did produce more utterances when they interacted with their mothers (but see Borstein, Haynes, Painter, & Genevro, 2000). Thus, children's performance on grammatical measures (e.g., DSS scores and MLU) in the language sample may not vary significantly with the interactants, at least for children who were 3 years old or older when they interacted with their mother or a trained examiner.

Sentence Point

Although Lee (1974) reported mean scores for the sentence point for each 1-year age group, she did not report a measure of variability (i.e., standard deviations) or suggest cutoff scores for each age. There is, therefore, no basis for making decisions about passing and failing performance for the sentence point.

Recent studies show that the sentence point differentiates between children with and without language impairment (Eisenberg & Guo, 2013; Suoto et al., 2014). Both studies converted the sentence point score into a percentage score (percent sentence point [PSP]) by dividing the total sentence points by the number of utterances and multiplying by 100%. Eisenberg and Guo (2013) suggested a cutoff score for PSP of 58% for 3-year-old children on the basis of an event description sample. Souto et al. (2014) investigated PSP for 4-year-old and 5-year-old children on the basis of conversational samples. They reported mean PSP scores—over 90% for children with TL for both ages compared to a mean for the group with LI of 60% for 4-year-olds and 70% for 5-year-olds—but did not suggest a cutoff score. In addition, Souto et al. (2014) reported better diagnostic accuracy for PSP than for DSS.

Purpose

The current study compared DSS performance for the two tasks that yielded the most talking from 3-year-old children in the DSS normative study—free play with toys and event descriptions about pictures. We computed the differences and correlations between the two sample types for DSS scores and sentence point scores. We anticipated that DSS scores might be higher for the event description task than for the play sample, given Lee's (1974) suggestion that performance for event description would yield more-sophisticated grammatical forms. We included the sentence point score in the present study because, although Lee did not provide normative data for this measure, recent studies have shown this measure to be clinically useful for differentiating children with and without LI (Eisenberg & Guo, 2013; Souto et al., 2014). We also examined the degree of agreement for making pass–fail decisions between the two samples. This was important given the finding by Longhurst and File (1977) that different contexts might yield different decisions about pass–fail performance. We asked the following questions:

1. Will children's performance on the DSS (i.e., DSS score, sentence point score) differ between language samples collected during free play with parents and during event description elicited by an examiner?
2. Will children's performance on the DSS for the parent-elicited free play sample and the examiner-elicited event description sample be significantly correlated?
3. To what extent will the two sample types yield the same pass–fail decisions when the children's DSS scores are compared with the 10th percentile cutoff suggested by Lee (1974)? Note that we did not ask this question about the sentence point score because Lee (1974) did not provide normative data for this measure.

Method

Participants

Participants for the current study were drawn from sixty-five 3-year-old children (35 boys; 30 girls) who had completed both a play activity and an event description task as part of a previous award to the first author. Parents had signed consent to have the language samples archived and used in further studies. Parents completed a questionnaire regarding their child's development, and children were excluded if the parent reported concerns or a prior diagnosis of hearing loss, premature birth, cognitive difficulty, social–emotional deficits, or neurological conditions. Children were not excluded if there was a concern or prior diagnosis of speech or language impairment. However, participants had to be producing at least three-word utterances and had to pass an articulation screening in order to be included in the study. All participants spoke mainstream English on the basis of parent report and passed an oral mechanism screening, a cognitive screening, and a hearing screening at 25 dB at 500, 1000, 2000, and 4000 Hz. Socioeconomic status (SES) was based on maternal education, with 91% having a college degree and 9% having a high school degree. The racial distribution on the basis of self-identification by the parent was 69% Caucasian, 17% African American, and 14% Asian. Sixteen percent of the participants also identified themselves as Hispanic.

Lee (1974) suggested that DSS would be an appropriate analysis for children who can produce at least 50 complete sentences within a 1-hr time period. All participants met these criteria. However, because the DSS analysis requires 50 complete and unique utterances, we excluded seven children who did not produce the requisite sample size for both sample types, one child on the free play sample and six children on the event description sample. For the current study, we reported only the data from the remaining 58 children (32 boys; 26 girls) who produced at least 50 scorable utterances for both the play and event description samples. The mean age of those 58 children was 3;6 (*SD* = 0;4, range = 3;0–3;11).

As part of the study protocol, we administered the Structured Photographic Expressive Language Test–Preschool 2 (SPELT-P 2; Dawson, Eyer, & Fonkalsrud, 2005) to document children's language ability. The mean standard score of the SPELT-P 2 for the 58 children was 105.14 (*SD* = 13.02, range = 65–125). It should be noted that four children had a standard score lower than 85 (i.e., more than 1 *SD* below the mean) and that 12 children had scores above 115 (i.e., more than 1 *SD* above the mean) on the SPELT-P 2. However, participants for the current study

were included without consideration of how they performed on the standardized language test. This was done in order to ensure that the range of DSS values would not be restricted and that the comparison of sample types would include DSS scores from children with varying language levels (see Goodwin & Leech, 2006; Pawlowska, 2014; Ukrainetz McFadden, 1996).

Language Sampling Conditions

The language samples were from archived video-recorded and audio-recorded samples of free play with toys and of elicited descriptions of pictured events (i.e., the event description task). The play samples were elicited by having each child play with his or her parent. Parents were instructed to follow their child's lead and to try not to ask questions. There were five sets of toys that included a vehicle set, a food set, a dollhouse set, a baby care set, and a farm set. The child was given one set of toys to start, and an additional set of toys was introduced every 6 min for a total of 30 min. This was done to maintain child interest throughout the 30 min of playtime. Order for introducing the play sets was randomized. Children were free to choose between the new toy set and previously introduced toys.

For the event description task, children were shown 15 pictures, each including at least three characters. An examiner presented one picture at a time in a randomized order and prompted the child using the same series of prompts for each picture with the exception of the third prompt, which varied for each picture. The prompts, adapted from Leonard, Bolders, and Miller (1976), were as follows: (a) What is happening in the picture? (b) What else is happening in the picture? (c) Now, I will start a story, and you finish it. (d) Tell me one more thing about the picture. The third prompt was followed by a story starter specific to each picture followed by "and then" to prompt the child to complete the examiner's sentence. Alternative prompts were provided if the child did not respond or produced an off-topic utterance in response to any of the four original prompts. The total time for eliciting the event description sample ranged from 14 to 40 min (M = 27 min; SD = 7 min).

The two samples thus differed in the interactant eliciting the sample. Although this potentially introduced a confound, we thought it was important to have parents elicit the play sample because it was a common practice in a clinical setting (Fey et al., 1993; Friedman & Friedman, 1980; Holdgrafer, 1995; Leonard et al., 2006; Loeb et al., 1998; Souto et al., 2014; Washington, 2013), and it was more likely for us to obtain sufficient number of utterances for the DSS analysis in the 30-min span. In addition, previous studies (Hannson et al., 2000; Kramer, et al.; 1979; Olswang & Carpenter, 1978) have shown that children's performance on the grammatical measures in the language samples would not vary significantly when they interacted with their parents or trained examiners, at least in 3-year-olds.

Transcription and Utterance Inclusion

The language samples were transcribed following the conventions of Systematic Analysis of Language Transcripts (Miller & Iglesias, 2010), except for utterance segmentation, which was in phonological units to conform to the DSS guidelines. Utterances that could not be clearly understood after listening for a maximum of three times were transcribed as unintelligible and excluded from the analysis. Also excluded were noncompleted utterances (i.e., abandoned or interrupted) and single-word yes/no utterances and interjections. A consensus procedure (adapted from Shriberg, Kwiatkowski, & Hoffman, 1984) was used to check transcription. Each sample was transcribed by one research assistant and checked by a second research assistant and, then, by the first author. Any discrepancies that could not be resolved were excluded.

Based on the DSS guidelines, 50 consecutive complete, unique, self-generated utterances from each sample were required for the DSS analysis. A complete utterance, by definition, must have at least a subject and a verb, although a complete sentence does not have to be grammatically accurate (e.g., *He run away*). Accordingly, fragments and utterances lacking a subject or a verb were excluded from the DSS analysis. Utterances that were completely or partly unintelligible were also excluded from the DSS analysis. All utterances included in the DSS analysis must be unique, meaning that each utterance must be different. Accordingly, imitations and repeated utterances were excluded from the DSS analysis. Note that this includes utterances replicated anywhere in the sample, even if separated from the earlier utterance by multiple other utterances. However, utterances with even one word different are considered unique and are, therefore, included in the analysis; for example, *The boy wants it* and *The other boy wants it* would both be included in the DSS analysis. In addition, utterances were segmented so that they contained no more than two independent clauses conjoined by *and*. Imperative interjections (i.e., *look, lookit, see*) and sentence tags (e.g., *you know, I think*) were segmented into separate utterances and considered complete sentences. Interrater agreement for utterance inclusion and segmentation, on the basis of 10 samples (i.e., approximately 15%) of each type, was 97%.

Scoring and Analyses

Utterances were scored in accordance with the guidelines in Lee (1974) and Lively (1984). Weighted scores were given for all grammatically correct structures in each utterance using the scoring chart in Lee (1974). Attempt marks were given for productions that did not meet standard English conventions. A sentence point was given for a given utterance only when a sentence was grammatically and semantically correct. Interrater agreement for DSS scoring, on the basis of 10 samples of each type, was 94%.

The following scores were calculated for each sample type from each child: (a) the DSS score was calculated by adding the points earned for each utterance (i.e., weighted category scores plus the sentence point) and dividing by the total number of utterances (i.e., 50 utterances for each child); and (b) the sentence point score was calculated by totaling the number of utterances earning the sentence point.

One-way repeated-measures analyses of variance (ANOVAs) were adopted to examine whether DSS scores and sentence point scores differed between the play and event description samples. This allowed us to determine whether each of the target measures differed significantly between the two samples. We used the *d* value to quantify the effect size or magnitude of the differences. Following Cohen (1988), we interpreted the effect size as small ($0.2 \leq d < 0.5$), medium ($0.5 \leq d < 0.8$), or large ($d \geq 0.8$) whenever appropriate. Pearson product–moment correlations were also computed to examine the extent to which the DSS and sentence point scores for the two sample types were significantly correlated. This allowed us to evaluate the extent to which individual children's performance on a given measure was consistent, relative to other children, between two sample types (e.g., the extent to which a child who scored higher than other children on the sentence point score for the play sample also scored higher than those other children on the sentence point for the event description sample). On the basis of previous studies (Gavin & Giles, 1996), we interpreted the degree of correlation as minimally acceptable ($.71 \leq r < .90$) or acceptable ($r \geq .90$).

Lee suggested the 10th percentile as the clinical cutoff for DSS. This is equivalent to a score that is 1.25 *SD*s below the mean. In order to compare DSS scores to this cutoff, we transformed the DSS raw scores into z-scores. These were calculated by subtracting the child's DSS score from the normative mean (Lee, 1974) and then dividing the difference by the normative standard deviation (Lee, 1974). We then compared the degree of agreement for the pass–fail decisions between the play and event description samples using Cohen's kappa (κ). Following Landis and Koch (1977), we interpreted the degree of agreement as fair ($.21 \leq \kappa < .40$), moderate ($.41 \leq \kappa < .60$), substantial ($.61 \leq \kappa < .80$), or almost perfect ($.80 \leq \kappa$).

Results

The DSS analysis requires 50 complete and unique utterances. Recall that one child produced fewer than 50 scorable utterances on the play sample and six children produced fewer than 50 scorable utterances on the event description sample. Thus, the comparisons between sample types were based on the remaining 58 children.

Preliminary Analysis

Because Lee (1974) suggested that DSS be used for children who produced complete sentences 50% of the time,[1] we first examined the number of utterances needed to obtain 50 scorable utterances for the DSS analysis (hereafter, NU-50) to evaluate the appropriateness of each sample type. Note that NU-50 was determined from the analysis set, after excluding unintelligible, noncompleted, and single-word yes/no and interjection utterances. The mean of NU-50 was 79.74 ($SD = 12.92$) for the play sample and 65.07 ($SD = 16.93$) for the event description sample. One-way repeated-measures ANOVA showed that NU-50 was larger for the play sample than for the event description sample ($F = 46.09$, $p < .001$, $d = 1.80$), meaning that more utterances were needed to obtain 50 scorable utterances in the play sample than in the event description sample. The effect size was large.

Five children needed more than 100 utterances for the free play and/or event description samples to yield the requisite 50 complete and unique utterances for the DSS analysis (i.e., three children for the play sample and two children for both play and event description samples). We conducted separate analyses with and without those children. We found that analyses with and without those five children generated similar results. Therefore, we reported only the analyses with those children below.

Comparison of Sample Types

DSS Score

DSS scores for boys and girls were not significantly different for the play sample ($F = 0.271$, $p = .61$, $d = 0.14$) or for the event description sample ($F = 0.644$, $p = .43$, $d = 0.21$). Age did not significantly account for the variance in DSS scores among 3-year-olds for either sample (play: $R^2 = .020$, $p = .30$; event description: $R^2 = .046$, $p = .11$). We, therefore, collapsed the data across genders and ages for the analyses of DSS scores.

DSS scores for the play and event description samples are shown in Table 1. One-way repeated-measures ANOVA showed that DSS scores from the two sample types were not significantly different, $F(1, 56) = 1.306$, $p = .29$, $d = 0.30$. The effect size was small. Despite the nonsignificant

[1]Developmental sentence analysis includes two analyses: DSS and developmental sentence types (DST). The DST analysis is based on 100 utterances and evaluates "presentences," utterances that do not include both a subject and verb. Lee (1974) first identified 100 utterances for the DST analysis. According to Lee (1966), the following utterances were excluded from the DST analysis: utterances that were partly or completely unintelligible, noncompleted utterances (utterances that were abandoned or interrupted), repeated (nonunique) utterances, and single-word interjections, including *yes* and *no*. Lee (1974) selected the sample for DSS from the 100 utterance DST sample after eliminating incomplete sentences (i.e., sentences without both a subject and verb). Lee then stipulated that if a child produced at least 50 complete utterances out of the 100 utterance DST sample, the clinician could eliminate the presentences and do only a DSS analysis. For the current study, we did not first identify 100 utterances for DST, and we kept repeated utterances in the analysis set. We excluded the repeated utterances for the DSS analysis and counted them for calculating NU-50 (i.e., the number of utterances needed to yield 50 scorable utterances).

Table 1. Means (*SD*s) of DSS scores and sentence point scores by sample type.

Measure	Play	Event description
DSS score		
All	6.65 (1.47)	6.31 (2.14)
3;0–3:5	6.35 (1.39)	5.68 (2.16)
3;6–3;11	6.85 (1.51)	6.72 (2.05)
Sentence point score		
All	40.36 (6.22)	33.33 (10.25)
3;0–3:5	38.65 (7.33)	32.57 (9.91)
3;6–3;11	41.49 (5.18)	33.83 (10.59)

Note. DSS = developmental sentence scoring.

difference between the two sample types, Table 2 shows that there was a trend for children to score higher on the play sample than on the event description sample (33 vs. 19), and this was particularly true for the younger children (16 vs. 5). The range of scores (shown in Figure 1) was wider for the event description sample (1.84 to 11.68) than for the play sample (2.76 to 9.70).

Pearson correlation indicated that DSS scores from the two sample types were significantly correlated ($r = .52$, $p < .001$). This means that children who had higher DSS scores than other children in the play sample also tended to have higher DSS scores in the event description sample. However, the magnitude of correlation was below the minimally acceptable level (i.e., $r = .71$).

Sentence Point Score

Sentence point scores for boys and girls were not significantly different for the play sample ($F = 0.136$, $p = .71$, $d = 0.09$) or for the event description sample ($F = 1.44$, $p = .24$, $d = 0.32$). Age did not significantly account for the variance in the sentence point among 3-year-olds for either sample type (play: $R^2 = .045$, $p = .11$; event description: $R^2 < .001$, $p = .99$). We, therefore, collapsed the sentence point data across genders and ages in the analyses of sentence point scores.

Table 2. Number of children scoring higher, lower, and the same on play and event description samples.

Measure	Higher on play	Higher on event description	Same
DSS score			
3;0–3:5	16	5	2[a]
3;6–3;11	17	14	4[a]
All	33	19	6[a]
Sentence point score			
3;0–3:5	18	0	5[b]
3;6–3;11	27	4	4[b]
All	45	4	9[b]

Note. DSS = developmental sentence scoring.
[a]Within ± 0.2 points. [b]Within ± 1 point.

Sentence point scores for the play and event description samples are shown in Table 1. One-way repeated-measures ANOVA showed that sentence point scores were significantly higher for the play sample than for the event description sample, $F(1, 56) = 52.73$, $p < .001$, $d = 1.93$. The effect size was large. Table 2 further reveals that most of the children earned a higher sentence point score on the play sample than on the event description sample (45 vs. 4, with nine scoring the same), and this was true for both older and younger children. It is noteworthy that, for the children earning a higher sentence point score on the play sample, the difference ranged from 1 up to 28 sentence points. In contrast, for the children who earned a higher score on the event description sample, the difference was only 1 or 2 sentence points.

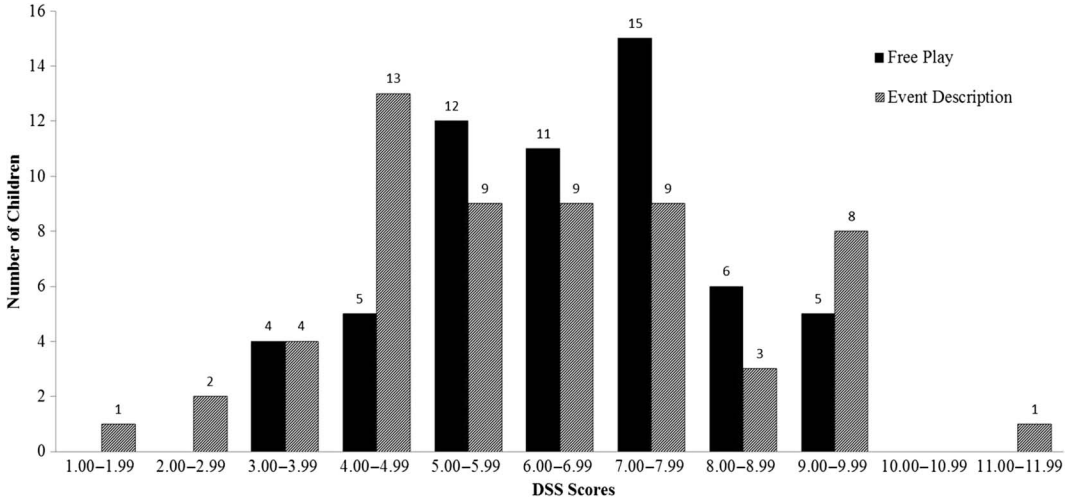
Sentence point scores from the two samples were significantly correlated ($r = .71$, $p < .001$). That is, children who scored higher than other children on the sentence point in the play sample also tended to score higher than others on the sentence point in the event description sample. The magnitude of the correlation was at a minimally acceptable level.

The DSS analysis assigns an attempt mark when production is ungrammatical, and utterances with attempt marks do not earn a sentence point. Children earned fewer sentence points for the event description sample than for the play sample. To examine whether this difference was driven by specific categories, we conducted an in-depth analysis of attempt marks, shown in Table 3. Children received more attempt marks on the event description sample than on the play sample. More than half of the attempt marks for both sample types were received on the main verb category, with slightly more on the event description than on the play sample. Children also received more attempt marks for the personal pronoun category on the event description sample than on the play sample. However, children received more attempt marks in the interrogative reversal category on the play sample than on the event description sample.

Pass–Fail Decisions

Table 4 shows the agreement for passing and failing using −1.25 *SD*s as the cutoff. This corresponds to the 10th percentile cutoff recommended by Lee (1974). Scores were rated as passing if they were at or above the −1.25 *SD* (10th percentile) cutoff and rated as failing if they were below this cutoff. Approximately 79% (46/58) of the children achieved a passing score on the play sample, and 64% (37/58) of the children achieved a passing score on the event description sample. The overall agreement for pass–fail decisions between the play and event description samples was 78% (45/58). Cohen's kappa indicated that the degree of agreement between the two sample types was significant at the moderate level ($\kappa = .47$, $p < .001$). Table 5 compares the scores for the 13 children whose pass–fail decisions for the two samples did not agree. The majority of disagreements were for children who had failed on the

Figure 1. Distribution of developmental sentence scoring (DSS) scores.



event description sample and passed on the play sample (11 children).

When the play sample was used as the reference point, there was 76% agreement for passing (35/46) and 83% agreement for failing (10/12). When the event description sample was used as the reference point, there was 95% agreement for passing (35/37) but only 49% agreement for failing (10/21). As noted above, there was an apparent difference between younger and older children in the number of children who scored higher on the play sample than on the event description sample. We therefore, also evaluated agreement for pass–fail decisions separately for the two age groups.

Table 3. Number (percentage) of attempt marks (errors) by scoring category and sample type.[a]

Scoring categories	Play	Event description
Indefinite pronouns and modifiers	9 (2%)[b]	4 (< 1%)
Personal pronouns	36 (8%)	203 (22%)
Main verbs	278 (59%)	580 (64%)
Secondary verbs	44 (9%)	62 (7%)
Negatives	19 (4%)	10 (3%)
Conjunctions	12 (3%)	26 (3%)
Interrogative reversals	70 (15%)	25 (3%)
Wh-question words	7 (1%)	1 (< 1%)
Total	475 (100%)	911 (100%)

[a]Attempt marks indicate errors in specific categories in the developmental sentence scoring analysis. [b]The percentage was computed by dividing the number of attempt marks of a given scoring category by the total number of attempt marks across categories. For example, the number of attempt marks for indefinite pronouns and modifiers was nine, and the total number of attempt marks across categories was 475 in the play sample. Thus, the percentage for the indefinite pronouns and modifiers was 2% (9/475) in the play sample.

Younger 3-Year-Olds (Age 3;0–3:5, *n* = 23)

Seventy percent (16/23) achieved a passing score on the play sample, and 43% (10/23) of the younger children achieved a passing score on the event description sample. The overall agreement for pass–fail decisions between the picture and play samples was 74%. Cohen's kappa indicated that the degree of agreement between the two sample types was significant at the moderate level

Table 4. Pass–fail agreement for DSS scores using a 10th percentile cutoff (Lee, 1974).

a. All children

		Play samples		
		Pass	Fail	Total
Event description samples	Pass	35	2	37
	Fail	11	10	21
	Total	46	12	58

b. Younger 3-year-olds (3;0–3;5)

		Play samples		
		Pass	Fail	Total
Event description samples	Pass	10	0	10
	Fail	6	7	13
	Total	16	7	23

c. Older 3-year-olds (3;6–3;11)

		Play samples		
		Pass	Fail	Total
Event description samples	Pass	25	2	27
	Fail	5	3	8
	Total	30	5	35

Note. DSS = developmental sentence scoring.

Table 5. Nonagreements on pass–fail decisions.

Child	Age (months)	DSS z-score: Play	DSS z-score: Picture	z-score difference
058NR	36	0.24	**–2.00**[a]	2.74
082JN	36	–0.78	**–1.82**	1.04
013AZ	38	0.54	**–2.52**	3.06
071YV	38	0.02	**–2.20**	2.22
077EG	39	–0.66	**–2.24**	1.58
089CZ	39	–0.88	**–2.84**	1.96
066NK	42	**–3.00**	1.10	4.10
037 MB	43	–0.14	**–2.04**	1.90
043MS	43	**–1.78**	–1.10	0.68
022KS	45	–1.08	**–2.70**	1.62
024 AD	45	1.98	**–2.18**	3.84
025TS	46	–0.78	**–1.64**	0.86
042YO	47	0.26	**–2.54**	2.80

Note. DSS = developmental sentence scoring.

[a]Bold numbers indicate failing scores below the cutoff.

(κ = .50, p = .005) in younger 3-year-olds. When the play sample was used as the reference point, there was 63% agreement for passing (10/16) and 100% agreement for failing (7/7).When the event description sample was used as the reference point, there was 100% agreement for passing (10/10) and 54% agreement for failing (7/13).

Older 3-Year-Olds (Age 3;6–3;11, *n* = 35)

Eighty-six percent (30/35) achieved a passing score on the play sample, and 77% (27/35) of the older children achieved a passing score on the event description sample. The overall agreement for pass–fail decisions between the picture and play samples was 80% (28/35). Cohen's kappa indicated that, although significant, the degree of agreement between the two sample types was only fair (κ = .35, p = .03) in older 3-year-olds. When the play sample was used as the reference point, there was 83% agreement for passing (25/30) and 40% agreement for failing (2/5). When the event description sample was used as the reference point, there was 93% agreement for passing (25/27) and 38% agreement for failing (3/8).

Discussion

The current study investigated whether sample types affected DSS results for 3-year-old children. We chose to compare free play with toys and elicited descriptions about pictured events because these were the two sample types that elicited the most talking from 3-year-old children in the DSS normative study (Koenigsknecht, 1974). DSS scores were not significantly different between the play sample and the event description sample. However, the sentence point score was significantly higher for the play sample than for the event description sample, and the effect size for this difference was large. DSS scores and sentence point scores were significantly correlated between the two samples. However, the magnitude of the correlation for the sentence point score reached only a minimally acceptable level, and the magnitude of correlation for the DSS score was below a minimally acceptable level. What this means is that the rank ordering of scores of individual children varied considerably between the two samples. Overall agreement between the two sample types for pass–fail decisions was also only moderate, with different decisions between the samples for 13 (22%) of the children. We explore those findings below.

Discrepancy Between Sample Types

DSS Scores

Lee (1974) hypothesized that elicited event descriptions of pictures would yield more sophisticated utterances than play with toys and would thus earn more points on the DSS. This hypothesis was not borne out for the 3-year-old participants in the current study. The mean DSS score was not higher for the event description sample. Rather, DSS scores were not significantly different between the two sample types at the group level. For the younger 3-year-olds, the trend was even in the opposite direction from Lee's prediction at the individual level, with many of the younger children earning lower DSS scores on the event description sample than on the play sample. Although there were no previous studies comparing the play sample and the event description sample, this finding is consistent with previous studies reporting higher DSS scores for another type of conversational sample, interviewing, relative to event description (Fields & Ashmore, 1980; Longhurst & File, 1977) while finding no difference in DSS scores between interviewing and play (Evans & Craig, 1992).

To further explore the trend of higher DSS scores in the play sample, we did a qualitative analysis of category scores. Children earned more points on the play sample than on the event description sample for indefinite pronouns, main verbs, negatives, interrogative reversal, and *wh*-questions. In contrast, children earned more points on the event description sample than on the play sample for personal pronouns, secondary verbs, and conjunctions.

These results did not suggest that the higher DSS scores for the play sample were due to greater utterance sophistication. Children earned more points on the event description sample for secondary verbs and conjunctions, the two categories that reflect use of complex sentences. What seemed to largely account for the difference in DSS score was the much larger number of points earned for interrogative reversal and *wh*-questions in the play sample. That is, children asked many more questions during play than during event description and, consequently, earned a much larger number of points in these two categories.

Sentence Point Scores

Children earned significantly fewer sentence points on the event description sample than on the play sample, reflecting a higher rate of grammatical errors on the event description sample. This meant not only that children earned fewer sentence points but that points may have been withheld for ungrammatical usage of structures within the scoring categories. To explore whether this may have affected the DSS scores, we did an in-depth analysis for the number of attempt marks in each category. The results are equivocal. In the main verb category, children received more attempt marks and earned fewer points on the event description sample than on the play sample. However, for both the secondary verb and personal pronoun categories, children received more attempt marks and yet still earned more points on the event description sample than on the play sample. A higher rate of errors, therefore, did not necessarily correspond with earning fewer points.

Pass–Fail Decisions

Although not statistically significant, the difference in the DSS scores between the play and event description samples at the individual level was clinically important as it affected pass–fail decisions. More children scored below the 10th percentile cutoff on the event description sample, and the overall agreement rate for pass–fail decisions was only moderate between the two sample types. Notably, children who scored above the cutoff on the event description sample were likely to score above the cutoff on the play sample as well, whereas children scoring below the cutoff on the play sample were likely to score below the cutoff on the event description sample. These trends support Lee's suggestion that the event description is more challenging than talking during play.

Comparison to the DSS Normative Study

The DSS score range in the current study was wider than the range reported by Lee (4.60–8.60) for both the play samples (3.64–9.70) and event description samples (1.84–11.68). In spite of this, the 6.65 mean DSS score for the play sample was comparable to the 6.64 mean score previously reported by Lee (1974), whereas the mean 6.31 DSS score for the event description sample fell between the 50th and 25th percentiles in the DSS normative study. We considered differences between our procedures and the procedures for the DSS normative study that could have influenced the mean and the range of the DSS scores.

Participants

Lee (1974) included only participants judged to have language skills in the average range. She eliminated children with test scores more than 1 *SD* above or below the mean on a standardized test for receptive vocabulary. In contrast, the current study included children with a wider range of language ability. Removing the 16 children who scored more than 1 *SD* above or below the mean on the SPELT-P 2 did not change how the DSS scores in the present study compared with the normative DSS data. That is, the mean for the play sample (6.59; range = 3.64–9.70) remained at the DSS mean, and the mean for the event description sample (6.12, range = 1.84–10.0) remained between the 50th and 25th percentiles. The inclusion of participants with higher and lower standardized test scores did not, therefore, appear to have affected the results of the current study.

SES of participants was similar between this study and the DSS normative study. All but three of the participants in the DSS normative study were from middle-income families, on the basis of the father's income. We used maternal education to categorize participants, and 91% had a college degree. This imposes a limit on the applicability of the results for lower SES children but would not affect the results relative to the DSS normative procedures.

Interactants

In the current study, parents elicited the play sample and examiners elicited the event description sample. This differed from the DSS normative study, in which an examiner elicited the entire sample, including both of these sample types and storytelling. We know of only one study (Kramer et al., 1979) that examined the impact of interactant on DSS scores. That study did not find a significant difference between DSS scores elicited by the mother at home versus examiners in a clinic setting. However, more children earned higher DSS scores for the home sample elicited by the mother. We found no study that compared DSS scores for parent and examiner samples elicited in the same setting, and it remains possible that the difference in the interactant affected DSS performance.

Sample Elicitation

For the play sample, we used similar toys to those used by Lee (1974), although we included five instead of three different toy sets to maintain interest. We followed Lee's recommendation to introduce materials gradually rather than all together to add novelty. For the event description sample, we used similar pictures to those used by Lee (1974). The prompts, however, were different. In addition to the "what is happening?" prompt, we added three other prompts to encourage more talking. According to Lee, the clinician's priority is to keep the child interested

and talking and that this is more important than standardizing the procedures. The procedures that we used for the event description task were successful in eliciting sufficient utterances for DSS analysis even from younger 3-year-olds, who talked less in this condition during the DSS normative study. However, this may have resulted in a more structured event description task than was used for the DSS normative study.

The biggest procedural difference was in the samples used for calculating DSS. Lee (1974) combined different types of language samples when calculating the DSS normative data, whereas the current study calculated DSS separately for each sample type. The language samples we compared with the reference data were not, therefore, matched to the language samples used for developing the normative data. In fact, it would not have been possible to accomplish this matching because the relative proportion of utterances that was included from each activity type was not specified, and this varied among the children (Lee, 1974). Notably, the mean for the play sample was comparable to the normative mean. The range for the play sample was also comparable when we restricted the samples to average-performing children on the standardized test.

Clinical Implications

The current study provides information about reliability of DSS performance across sample types. It does not provide information about the diagnostic accuracy of DSS. This is because we do not have sufficient information to determine the language status for all of the children in the current study. Although we do have SPELT-P 2 scores for all of the children, there is reason to believe that scores on the SPELT-P 2 may overestimate language in 3-year-old children and, consequently, underidentify children with language impairment at this age (Oetting & Hadley, 2009). We did not follow up with the children with low DSS scores (i.e., below the 10th percentile cutoff) to see whether or not they were subsequently evaluated and diagnosed by an SLP as having a language impairment. We cannot, therefore, make conclusions about the best sample to use for computing DSS.

It is common for clinicians to use only play samples for conducting the DSS analysis rather than combining utterances from different activities when assessing 3-year-old children. This may be because 3-year-olds are likely to produce more utterances during play than during other activities (Sealey & Gilmore, 2008; Southwood & Russell, 2004). However, the differences between DSS results (i.e., DSS scores and sentence points) for play and event descriptions suggest that choice of elicitation task could affect pass–fail decisions for children in this age range. This does not mean that clinicians should not use play samples for conducting the DSS analysis for 3-year-old children. It does suggest caution in interpreting results from the DSS analysis based only on play samples.

Note that most of the children who failed DSS on the play sample (i.e., who scored below the cutoff on the DSS score) also failed DSS on the event description sample and that most of the children who passed DSS on the event description sample also passed DSS on the play sample. This suggests that low DSS performance on a play sample and passing DSS performance on an event description sample may be representative of a child's language abilities. In contrast, over one quarter of the children who passed DSS on the play sample (i.e., scored at or above the cutoff) failed on the event description sample. A passing DSS score on a play sample might possibly, therefore, overestimate children's language abilities. Similarly, half of the children who failed the DSS on the event description sample passed on the play sample. A failing DSS score on an event description sample might, therefore, underestimate children's language abilities. In these cases, clinicians might want to recalculate DSS using a sample that includes utterances from both sample types.

However, we are not suggesting that clinicians use DSS as the sole measure for assessing children's language skills. Clinicians should never base decisions about language skills on a sole measure, and all measures must be technically sound. This is true for standardized tests and language sample analyses (Individuals with Disabilities Education Act, 2004).

Concluding Thoughts

When a child's performance will be compared to normative data, it is important to replicate the specific procedures used for gathering that comparison data (Eisenberg, Fersko, & Lundgren, 2001; McCauley & Swisher, 1984). However, this is not possible for DSS. The DSS guidelines suggest eliciting three types of language samples in the following order—play with toys, elicited description of pictures, and storytelling—but selecting utterances in reverse order to give priority to utterances from the last two sample types. The intent of these guidelines was to standardize data collection and utterance selection (Koenigsknecht, 1974). However, the reality is that the composition of samples used for DSS analysis will vary considerably across children and that this will affect pass–fail decisions.

The current study shows only moderate agreement for pass–fail decisions on the basis of DSS scores from different sample types for 3-year-olds. Studies are needed to establish separate norms and cutoff scores for different sample types. In addition, the DSS normative data (Lee, 1974) are grouped into 1-year age intervals with a single cutoff score. It is more common for LSA measures for younger children to use smaller age intervals—for instance, Rice et al. (2010) used 6-month intervals for MLU and Scarborough (1990) used 3-month intervals for Index of Productive Syntax. Although there is overlap between age groups, this results in lower cutoff scores for younger children. In the current study, the lower DSS scores for younger 3-year-olds relative to the normative mean resulted in a higher fail rate. Interestingly, Lee and Canter (1971) originally reported DSS performance in 6-month age intervals and provided separate 10th percentile cutoff scores for younger and older

3-year-olds. However, these data would not be applicable to the current reweighted DSS scoring system. Thus, future studies might also consider separate norms with smaller age intervals (e.g., 6 months) for 3-year-olds for a given sample type.

As previously noted, recent studies suggest that the sentence point score is a clinically useful measure (Eisenberg & Guo, 2013; Souto et al., 2014). Future studies are needed to provide normative data and cutoff scores for this measure. The current study showed the sentence point score to be affected by sample types. As for the DSS score, there will need to be separate norms for the sentence point for each sample type.

Acknowledgments

Data for this study were from archival data originally collected with support to the first author from the National Institute of Deafness and Other Communication Disorders, Award R21DC009218. The authors are grateful to the children who participated and to their parents who allowed them to participate and to the research assistants who collected and transcribed the samples. Portions of this study were presented at the 2014 Symposium for Research on Child Language Disorders in Madison, WI, and at the 2014 Convention of the American Speech-Language-Hearing Association in Orlando, FL.

References

Allen, M. S., Kertoy, M. K., Sherblom, J. C., & Pettit, J. M. (1994). Children's narrative productions: A comparison of personal event and fictional stories. *Applied Psycholinguistics, 15,* 149–176.

Atkins, C. P., & Cartwright, L. R. (1982). An investigation of the effectiveness of three language elicitation procedures on Head Start children. *Language, Speech, and Hearing Services in Schools, 13,* 33–36.

Bornstein, M. H., Haynes, O. M., Painter, K. M., & Genevro, J. L. (2000). Child language with mother and stranger at home and in the laboratory: A methodological study. *Journal of Child Language, 27,* 407–420.

Caesar, L. G., & Kohler, D. (2009). Tools clinicians use: A survey of language assessment procedures used by school-based speech-language pathologists. *Communication Disorders Quarterly, 30,* 226–236.

Cohen, J. (1988). The effect size index: d. *Statistical Power Analysis for the Behavioral Sciences, 2,* 284–288.

Dawson, J., Eyer, J. A., & Fonkalsrud, J. (2005). *Structured Photographic Expressive Language Test–Preschool 2*. 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., Fersko, T., & Lundgren, C. (2001). Use of MLU for identifying language impairment in preschool children: A review. *American Journal of Speech-Language Pathology, 10,* 323–342.

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

Evans, J. L., & Craig, H. K. (1992). Language sample collections and analysis: Interview compared to freeplay assessment contexts. *Journal of Speech and Hearing Research, 35,* 343–353.

Fey, M. E., Cleave, P. L., Long, S. H., & Hughes, D. L. (1993). Two approaches to the facilitation of grammar in children with language impairment: An experimental evaluation. *Journal of Speech and Hearing Research, 36,* 141–157.

Fields, T. A., & Ashmore, L. L. (1980). Effect of elicitation variables on analysis of language samples for normal and language-disordered children. *Perceptual and Motor Skills, 50,* 911–919.

Friedman, P., & Friedman, K. A. (1980). Accounting for individual differences when comparing the effectiveness of remedial language teaching methods. *Applied Psycholinguistics, 1,* 151–170.

Gavin, W. J., & Giles, L. (1996). Temporal reliability of language sample measures. *Journal of Speech and Hearing Research, 39,* 1258–1262.

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

Hannson, K., Nettelbladt, U., & Nilhom, C. (2000). Contextual influences on the language production of children with speech/language impairment. *International Journal of Language & Communication Disorders, 35,* 31–47.

Heath, S. B. (1986). Taking a cross-cultural look at narratives. *Topics in Language Disorders, 7,* 84–94.

Holdgrafer, G. (1995). Language abilities of neurologically normal and suspect preterm children now in preschool. *Perceptual and Motor Skills, 80,* 1251–1262.

Hughes, D. L., Fey, M. E., & Long, S. H. (1992). Developmental sentence scoring: Still useful after all these years. *Topics in Language Disorders, 12,* 1–12.

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.

Individuals with Disabilities Education Act, 20 U.S.C. § 1400 (2004).

James, S. L., & Button, M. (1978). Choosing stimulus materials for eliciting language samples from children with language disorders. *Language, Speech, and Hearing Services in Schools, 9,* 91–97.

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.

Kemper, S., Rice, K., & Chen, Y. J. (1995). Complexity metrics and growth curves for measuring grammatical development from five to ten. *First Language, 15,* 151–166.

Klein, H. B., Moses, N., & Jean-Baptiste, R. (2010). Influence of context on the production of complex sentences by typically developing children. *Language, Speech, and Hearing Services in Schools, 41,* 289–302.

Koenigsknecht, R. A. (1974). Statistical information on developmental sentence analysis. In *Developmental sentence analysis: A grammatical assessment procedure for speech and language clinicians* (pp. 222–268). Evanston, IL: Northwestern University Press.

Kramer, C., James, S., & Saxman, J. (1979). A comparison of language samples elicited at home and in the clinic. *Journal of Speech and Hearing Disorders, 44,* 321–330.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Lee, L. L. (1966). Developmental sentence types: A method for comparing normal and deviant syntactic development. *Journal of Speech and Hearing Disorders, 31,* 311–330.

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

Lee, L. L., & Canter, S. M. (1971). Developmental sentence scoring: A clinical procedure for estimating syntactic development in young children. *Journal of Speech and Hearing Disorders, 36,* 315–340.

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., Camarata, S. M., Pawtowska, M., Brown, B., & Camarata, M. (2006). Tense and agreement morphemes in the speech of children with specific language impairment during intervention: Phase 2. *Journal of Speech, Language, and Hearing Research, 49,* 749–770.

Lively, M. A. (1984). Developmental sentence scoring: Common scoring errors. *Language, Speech, and Hearing Services in Schools, 15,* 154–168.

Loeb, D. F., Pye, C., Richardson, L. Z., & Redmond, S. (1998). Causative alternations of children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 41,* 1103–1114.

Loeb, D. F., Stoke, C., & Fey, M. E. (2001). Language changes associated with Fast ForWord-Language: Evidence from case studies. *American Journal of Speech-Language Pathology, 10,* 216–230.

Longhurst, T. M., & File, J. J. (1977). A comparison of developmental sentence scores from Head Start children collected in four conditions. *Language, Speech, and Hearing Services in Schools, 8,* 54–64.

McCauley, R., & Swisher, L. (1984). Psychometric review of language and articulation tests for preschool children. *Journal of Speech and Hearing Disorders, 49,* 34–42.

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

Miller, J., & Iglesias, A. (2010). *Systematic Analysis of Language Transcripts (Research version)* [Computer software]. Middleton, WI: SALT Software.

Miller, J. F., & Chapman, R. S. (1981). The relation between age and mean length of utterance in morphemes. *Journal of Speech and Hearing Research, 24,* 154–161.

Mortimer, J., & Rvachew, S. (2010). A longitudinal investigation of morpho-syntax in children with speech sound disorders. *Journal of Communication Disorders, 43,* 61–76.

Oetting, J. B., & Hadley, P. A. (2009). Morphosyntax in child language disorders. In R. G. Schwartz (Ed.), *Handbook of child language disorders*. (pp. 341–364). New York, NY: Psychology Press.

Olswang, L. B., & Carpenter, R. L. (1978). Elicitor effects on the language obtained from young language-impaired children. *Journal of Speech and Hearing Disorders, 43,* 76–88.

Paul, R., Hernandez, R., Taylor, L., & Johnston, K. (1996). Narrative development in late talkers: Early school age. *Journal of Speech, Language, and Hearing Research, 39,* 1295–1303.

Paul, R., & Norbury, C. F. (2014). *Language disorders from infancy through adolescence* (4th ed.). St Louis, MO: Elsevier Mosby.

Pavelko, S. L., Owens, R. E., Ireland, M., & Hahs-Vaughn, D. L. (2016). Use of language sample analysis by school-based SLPs: Results of a nationwide survey. *Language, Speech, and Hearing Services in Schools, 47,* 246–258.

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.

Rescorla, L., Roberts, J., & Dahlsgaard, K. (1997). Late talkers at 2: Outcome at age 3. *Journal of Speech, Language, and Hearing Research, 40,* 556–566.

Retherford, K. S. (1993). *Guide to analysis of language transcripts* (2nd ed.). Eau, WI: Thinking Publications.

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 with and without language impairment. *Journal of Speech, Language, and Hearing Research, 53,* 333–349.

Ryan, B. P. (2000). Speaking rate, conversational speech acts, interruption, and linguistic complexity of 20 pre-school stuttering and non-stuttering children and their mothers. *Clinical Linguistics & Phonetics, 14,* 25–51.

Scarborough, H. (1990). Index of productive syntax. *Applied Psycholinguistics, 11,* 1–22.

Sealey, L. R., & Gilmore, S. E. (2008). Effects of sampling context on the finite verb production of children with and without delayed language development. *Journal of Communication Disorders, 41,* 223–258.

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.

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.

Washington, K. N. (2013). The association between expressive grammar intervention and social and emergent literacy outcomes for preschoolers with SLI. *American Journal of Speech-Language Pathology, 22,* 113–125.

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

Wren, C. (1985). Collecting language samples from children with syntax problems. *Language, Speech, and Hearing Services in Schools, 16,* 83–102.