



Language sampling for kindergarten children with and without SLI: mean length of utterance, IPSYN, and NDW

Lynne E. Hewitt^{a,*}, Carol Scheffner Hammer^b,
Kristine M. Yont^c, J. Bruce Tomblin^d

^a*Department of Communication Disorders, Bowling Green State University,
Bowling Green, OH 43403, USA*

^b*Department of Communication Disorders, The Pennsylvania State University,
University Park, Pennsylvania, USA*

^c*Harvard University Graduate School of Education, Cambridge,
Massachusetts, USA*

^d*Department of Communication Disorders, University of Iowa,
Iowa City, Iowa, USA*

Received 2 July 2004; received in revised form 5 October 2004; accepted 28 October 2004

Abstract

Language sample analysis measures have long been promoted as exhibiting greater ecological validity than formal testing in the assessment of language disorder in children. In practice, their use is often restricted to preschool children, owing to lack of normative information, as well as criticisms of the validity of commonly used measures for the language of older children. This study compared scores of kindergarten children (mean age 6 years) with and without specific language impairment (SLI) on three commonly used language sample analysis measures: mean length of utterance in morphemes (MLU-m), the index of productive syntax (IPSyn), and number of different words (NDWs). Mean scores of the children with SLI were significantly lower for all three measures, though not for all subtests of the IPSyn. A number of individual differences were observed; notably, several children with SLI scored as well as those without. The problems and promise of language sampling for children beyond the preschool years are discussed in light of these results.

Learning outcomes: (1) readers will gain an understanding of strengths and weaknesses of language sample measures in assessing kindergarten children with language impairment. (2) The reader will

* Corresponding author. Tel.: +1 419 372 7181; fax: +1 419 372 8089.

E-mail address: lhewitt@bgnet.bgsu.edu (L.E. Hewitt).

become aware of the utility of MLU in differentiating between young school age children with and without language impairment.

© 2004 Elsevier Inc. All rights reserved.

Keywords: Language sample analysis; Specific language impairment; Assessment; Language disorders

1. Introduction

Language sample analysis is an appealing assessment tool for several reasons. Measures derived from language sampling, such as mean length of utterance in morphemes (MLU-m), may have superior sensitivity and specificity in identifying children with language impairments (Aram, Morris, & Hall, 1993). Language sampling offers improved ecological validity relative to formal testing (Lund & Duchan, 1993; Naremore, Densmore, & Harman, 1995). There is evidence that language sample analysis may be less vulnerable to dialect and cultural variations than traditional formal tests (Stockman, 1996). Despite these advantages, a number of problems impede routine use of language sample analysis by clinicians. Eisenberg, Fersko, and Lundgren (2001) discuss problems in interpreting and applying available reference data sets for mean length of utterance, as well as a number of threats to validity of MLU-m arising from issues such as sample size and utterance identification and selection. Other language sampling issues relate to time constraints and lack of clinician knowledge in areas crucial to language sample interpretation, such as syntactic analysis. Clinicians frequently are willing to compute MLU-m for preschoolers and use those scores normatively, although the survey by Eisenberg et al. (2001) suggests that this practice is questionable. Few clinicians routinely use other types of tools, such as the index of productive syntax (IPSyn; Scarborough, 1990), to analyze preschool language samples. The use of language sampling is yet more infrequent for assessing children beyond the preschool years, and there is a tendency to rely most heavily on formal testing, both for diagnosis and development of intervention goals.

Clinical reluctance to use standard language sampling protocols to identify children with language impairments beyond the preschool years stems in part from lack of adequate reference data sets. Reference data have been reported by Leadholm and Miller (1992) up through age 13 for a variety of measures, including MLU-m and number of different words (NDW). Eisenberg et al. (2001) discuss some of the limitations of this data set, highlighting the important concern that the data set relies on as few as 27 samples per age group. The review by Eisenberg et al. (2001) highlights the need for more information about using MLU-m and other measures as normative measures for preschoolers. The available information for school age children is even less adequate. Thus, while the desirability of these measures is widely acknowledged, in practice their use is often restricted to toddler and preschool assessment. Among the difficulties that limit the utility of language sample analysis beyond the preschool years are concerns with collecting an appropriate and representative sample, and with identifying appropriate measures that will allow normative comparisons.

Critics of formal testing often advocate use of language sample analysis in order to improve the ecological validity of assessment protocols (Dunn, Flax, Sliwinski, & Aram,

1996; Miller, 1996; Lund & Duchan, 1993; Owens, 1999). Language sampling has the advantage of sampling a natural behavior of children, while formal testing may ask children to engage in activities that are foreign to their experience. Validity in assessment is a difficult issue, as there are many ways in which it may be compromised. One avenue for considering what types of information best validate a diagnosis of language impairment is to consider the contrast between normativist and neutralist definitions of disorder (see Paul, 2001, for a summary). A normativist approach calls for demonstrating that a problem exists in the life functioning of an affected person. A neutralist approach is concerned with showing that significant quantitative deviation from normality exists. These approaches align roughly with World Health Organization guidelines for performance versus capacity (World Health Organization, 2001). Whereas performance describes “what an individual does in his or her current environment,” capacity “describes an individual’s capacity to execute a task or an action in a standardized environment” (WHO, 2001, p. 4). Measures derived from language samples have the potential to unite the normativist (performance-based) and neutralist (capacity-based) approaches. When reference norms are available, language capacity can be measured against a statistical sample, thereby satisfying neutralist criteria for definition of disorder. Because the behavior measured is spontaneous and natural, it also reflects adaptive functioning in the realm of communicative success, thereby probing areas of risk for social disvalue (Fey, 1986) that are required for normativist definitions.

Establishing the utility of measures of language performance derived from samples requires addressing a range of problems. Measurement issues include: reliability of sampling; replicability of measurement; and deriving cut-offs for identification of disorder empirically. Logistical issues include establishing which contexts qualify as appropriate for collecting samples, and determining when a sample is unusable. See Eisenberg et al. (2001) and Johnston (2001) for a discussion of some of these issues relative to MLU-m.

Even when clinicians are confident that the samples they collect are valid and reliable, language sampling in the school age years is subject to problems arising from the assumptions under which traditional measures have been calculated. That is, MLU-m has been promoted as an indirect measure of syntactic development. Data from Brown (1973) have been used to derive five developmental stages for analyzing child language based on MLU-m. The usefulness of MLU-m for distinguishing syntactic development diminishes past Brown’s Stage V, because sentences that contain syntactic elaborations may be the same length as those that do not (Crain & Lillo-Martin, 1999; Leonard, 1998). Thus interest in applying MLU-m to older populations has been lacking. Klee (1992b) and Rollins, Snow, and Willet (1996) have challenged the validity of MLU-m for later-developing language. Scarborough, Wyckoff, and Davidson (1986) urged caution in use of MLU-m for children older than 42 months, in that Miller and Chapman’s (1981) results were not replicated. Despite the problems with currently available normative information, it can be argued that a measure primarily tapping early-developing morphosyntax may be valid for children with specific language impairment (SLI) older than 48 months, at least through the early elementary years. Persistent morphosyntactic deficits are well-documented in this population (Gopnik & Crago, 1991; King & Fletcher, 1993; Oetting & Horohov, 1997; Rice, Wexler, & Cleave, 1995;

Watkins & Rice, 1994). For English-speaking children, areas of significant morphosyntactic impairment include tense marking and agreement (see Leonard, 1998, for a review). Windsor, Scott, and Street (2000) found differences in morphosyntactic error rates in both spoken and written contexts for school-age children with language learning disability. Kemper, Rice, & Chen (1995) reported continued syntactic growth in normally developing (ND) children through age 7. The Leadholm & Miller (1992) database reveals upward trends in MLU-m through age 13. In longitudinal work examining the morphosyntactic development of children with SLI, Rice and colleagues have demonstrated that children at 8 years of age continued to show lower MLU's than age peers (Rice, Wexler, & Hershberger, 1998; Rice, Wexler, Marquis, & Hershberger, 2000). Some syntactic growth cannot be captured by simple morpheme counts, but as a measure of overall increasing linguistic competence, MLU-m may retain clinical utility beyond the earliest ages.

Because syntactic complexity is not directly reflected in utterance length, it is instructive to use measures that directly probe syntactic competence. The index of productive syntax (Scarborough, 1990) is a measure of morphological and syntactic structure developed for measuring language samples of preschool children. It was developed as a research tool to document syntactic development. It has never been normed on a large population, although Scarborough (1990) does report strong correlation between the IPSyn and MLU-m in small sample sizes of 15 per age group, up through the age of 4. The IPSyn has an advantage over MLU-m because it directly samples structures, with a rough rubric of emerging productivity whereby a given structure can receive 0 points (never occurs), 1 point (occurred once in sample), or 2 points (occurred twice or more). The four subscales examine noun phrase, verb phrase, question and negation abilities, and sentence structure. The sentence structure subscale looks at later-developing syntactic abilities, such as use of passive, relative clauses, and tag questions. Items were selected based on the normal language development literature, and then subjected to further winnowing in that structures were omitted if they did not show developmental progression over the age range sampled. For example, possessive marking on noun phrases was similar at all ages, so it was omitted from the noun phrase subscale. The particular choices made by Scarborough in developing the IPSyn have resulted in a tool that has both face and content validity as a measure of syntactic growth, at least in preschool children. The inclusion of late-developing syntactic structures and detailed verb morphology in two of the subscales suggests that it may have merit in looking at children's language beyond the preschool years.

Another measure of language complexity that has been popular in preschool language sampling has been number of different word roots (in samples of fixed length, such as 50 or 100 utterances) or NDW. Miller (1991) suggested that NDW may have better properties for investigating semantic development than type-token ratio (TTR). Although TTR shows little developmental progression (Watkins, Kelly, Harbers, & Hollis, 1995), differences in NDW have been reported between ND and SLI groups, up through the age of 5 years (the oldest children tested; Watkins et al., 1995; Klee, 1992a). Therefore, NDW, even with its limitations (Owen & Leonard, 2001; Richards & Malvern, 1997), shows promise as a means of measuring lexical development, both in the preschool years and beyond.

In summary, despite its weaknesses, language sample analysis has useful features not present in formal testing, in that it has greater ecological validity. Unlike formal tests, measures used in language sampling are applied to natural communicative behaviors. Morphosyntactic structures that can be probed by language sampling are known to be omitted by children with SLI through at least the early elementary years. Therefore, commonly used measures of preschool language may logically be extended through older ages. This study specifically investigated their relevance for children in kindergarten. The three measures discussed above were applied to samples of children with and without SLI in order to investigate their validity as identifiers of language impairment. We examined two syntactic measures—MLU-m, and the Index of Productive Syntax—IPSyn (Scarborough, 1990), and one semantic measure—NDW, number of different word roots. We predicted that kindergarten children with SLI would differ from their ND counterparts on these measures.

2. Method

2.1. Participants

Language samples from 54 children, 27 normally developing and 27 with SLI, collected as part of a larger project, the Epidemiology of Specific Language Impairment Project (EpiSLI; Tomblin et al., 1997; Tomblin, Records, & Zhang., 1996) were used. Each group consisted of 16 boys and 11 girls. Children in this study were identified as specific language impaired when they met the following criteria: English as primary language; no history of mental retardation, autism, or neurologic problems; passing of hearing screening; performance IQ greater than 85 on the block design and picture completion subtests of the Wechsler preschool and primary scale of intelligence—revised (Wechsler, 1989); performance more than 1.25 S.D. below the mean on two or more of five composite scores derived from a battery of language tests, including five subtests of the Test of Language Development, 2:P (Newcomer & Hammill, 1988), and a narrative comprehension and production task based on Culatta, Page, and Ellis (1983). See Tomblin et al. (1996) for further details on the EpiSLI standard. For the present work, a further exclusionary criterion was necessary, in that children could not be included whose samples contained fewer than 50 utterances. One of the language sample measures used, Number of Different Words, required a minimum of 50 utterances, and the others are also subject to decreased validity as sample size decreases.

Participants were identified from among the 216 children with SLI identified by the EpiSLI system (identified from among a normative sample for the diagnostic system of 2009 monolingual English-speaking children). All white monolingual speakers of English who were identified as SLI were then considered for inclusion in the study. (Note that children of other ethnic backgrounds were excluded in order to control for possible dialect differences.) An attempt was made to match each to a typically developing child of the same gender and ethnicity from the database. Children were further matched on age (± 0.3 years) and years of maternal education (± 1 year). The availability of matches meeting these criteria constrained the pool of children, resulting in an initial

pool of 60 children. Six of these were eliminated, because their samples contained less than 50 utterances (of these, two were normally developing and four had SLI). The final groupings consisted of 16 boys and 11 girls with and without SLI. There were no gender differences within either of the groups on the variables studied, so groups were collapsed for gender. The mean age of children with SLI was 6.01 years (S.D. 0.35; range 5.5–6.7 years). The mean age of control children was 5.99 years (S.D. 0.33; range 5.5–6.7). Maternal education averaged 12.8 (S.D. 1.44) and 12.6 (S.D. 1.41) years, respectively (range 10–17 years).

2.2. *Data collection*

Audiotaped language samples were elicited from participants. The format involved use of two story re-tell tasks drawn from the Multilevel Informal Language Inventory (Goldsworthy & Secord, 1982). These were interspersed with conversational questions on two topics, both related to the topics of the stories (the child's experience of birthday parties and knowledge of animals and pets). Order of presentation of stories and conversational elicitations were held constant across children. After a few minutes of getting acquainted conversation, children were read the first story, about a birthday party, and asked to re-tell it. Children were then engaged in conversation about their own experience with birthdays and birthday parties. When children seemed to have no more to say, the examiner then read and had the children re-tell the second story, about a visit to a farm.

Following the second story re-telling, the children were asked about their experience with farms, animals, and pets. Most children talked about their own pets during this portion. Conversation was open-ended after this point, however at all points in the elicitation the interaction was adult-directed in character. In essence, the elicitation contexts were similar to interview-based language sample elicitations, such as are typically used by clinicians when working with school age children. Data exist supporting the use of interviewing in eliciting more utterances and more advanced language structures with school age children (Evans & Craig, 1992; see Hadley, 1998, for a summary).

2.3. *Transcription*

The audiotaped samples were transcribed using the SALT language sample analysis program (Miller & Chapman, 2000), by graduate and undergraduate students in communication disorders. Transcribers were given a tape to practice on, and their performance was probed until competence with SALT conventions was demonstrated. Following transcription, the transcriber listened to completed tapes again while viewing the transcript. Two further steps were taken to ensure accuracy: a graduate student in charge of the project monitored transcriptions for accuracy by comparing randomly selected transcripts from each transcriber against audiotapes. If accuracy was judged inadequate, all the tapes done by that transcriber were re-transcribed and re-checked. Morphological segmentation accuracy and adherence to SALT conventions were further checked by both

the senior graduate assistant and the second author, independently, for all completed transcripts as a final step.

2.4. Analysis

Because the elicitation involved question-answer exchanges in both the narrative retelling and conversational contexts, and because narratives collected were quite short, usually no more than 10 utterances, we pooled these two contexts. A *t*-test comparing MLU-m for narrative to overall MLU-m was not significant, providing some support for our contention that the discourse parameters of the “narrative” and “conversational” contexts were quite similar.

Transcripts were analyzed using the SALT program, version 6.1 (Miller & Chapman, 2000). Using options available in the SALT software, MLU-m and NDW-50 were computed for the set of complete and intelligible utterances. Transcription followed criteria outlined by Lund & Duchan (1993) that excluded exact repetitions, replies to closed-ended questions, and abandoned or partially intelligible utterances. To mitigate possible decreased complexity caused by interviewer discourse control, some replies to closed-ended questions were also eliminated, if replies were three words or fewer, unless a three-word reply contained a subject and a verb. Selection of utterances for MLU-m and other language measures is controversial (Eisenberg et al., 2001). We employed conservative utterance inclusion criteria to ensure that as high a proportion of utterances as possible represented child productions unconstrained by adult discourse. A controlling discourse context can artificially deflate MLU-m, and potentially might constrain lexical diversity by limiting children’s opportunities to generate original utterances. Our procedure has similarities to Johnston’s (2001) MLU2 discussed above.

The Index of Productive Syntax was scored by graduate students in communication disorders following the scoring protocol outlined in Scarborough (1990). Students practiced on two to three transcripts until they demonstrated competency by having two or fewer disagreements with previously scored samples (this would be less than 1% disagreements: total categories judged equaled 49; each category could be scored as either 0, 1, or 2; thus each transcript would have 147 (3×49) possible occasions for disagreement). Graduate assistants in communication disorders calculated inter-observer agreement for the IPSyn. Each scorable cell of the IPSyn was considered separately. Total number of agreements on each IPSyn cell was divided by total number of agreements plus disagreements, for five randomly selected transcripts, yielding 96% agreement.

One modification to IPSyn scoring protocols was necessary, in line with Scarborough’s suggestions for use of the instrument. Scarborough (1990) urged investigators to consider sampling context in scoring the IPSyn, pointing out that pragmatic contextual factors may constrain a child’s production of particular forms. As previously stated, the protocol used for the samples was adult-directed, where adults asked many more questions and children few or, in some cases, none. The third subscale on the IPSyn deals with questions and negation; of the eleven items in the subscale, three are related to negation, and the rest all relate to questions. Children often waited passively for the adult to ask them questions, and they apparently felt constrained not to ask many themselves. Many children uttered two or fewer questions in the whole sample, thus leading to little variability of scores and a general

depression of this scale. For this reason, we decided that it would be inappropriate to score the question/negation subscale.

3. Results

3.1. Morphosyntactic and lexical measures

A series of group comparisons using analyses of variance (ANOVA) investigated whether scores of the morphosyntactic and lexical language measures distinguished the two groups. Children with SLI performed significantly lower than control children on each of the following grammatical and lexical measures:

- 1) MLU(m) [$F(1,53) = 12.82, p < 0.001, \eta^2 = 0.20$].
- 2) NDW-50 [$F(1,53) = 9.68, p < 0.01, \eta^2 = 0.16$].
- 3) total score on the Index of Productive Syntax [$F(1, 53) = 6.79, p < 0.01, \eta^2 = 0.12$].
- 4) IPSyn sentence structure subscale [$F(1,53) = 6.63, p < 0.01, \eta^2 = 0.11$].

Neither the IPSyn noun subscale, nor the verb subscale was significant at the 0.05 level, [IPSyn noun subscale: $F(1,53) = 2.03, p = 0.16, \eta^2 = 0.04$; IPSyn verb subscale: $F(1, 53) = 3.49, p = 0.06, \eta^2 = 0.06$]. Means and standard deviations for each group are presented in Table 1. The effect sizes for all the significant differences were medium, using Rosenthal and Rosnow (1984) definitions.

A discriminant analysis was also conducted and analyzed, following Mertler and Vannatta (2002), to determine whether six variables (IPSyn nouns, IPSyn verbs, IPSyn sentence structure, IPSyn total score, MLU-m, and NDW-50) could predict the group membership. One function was generated and was significant, $\Lambda = 0.742, p = 0.0241$, indicating that the function of predictors significantly differentiated between children with and without language impairment. Standardized function coefficients and correlation coefficients revealed that the variables of MLU-m, IPSyn total score, and NDW-50 were most associated with the function (see Table 2). Original classification results revealed that

Table 1
Syntactic and semantic language measures for kindergarten children with and without SLI

	SLI			ND		
	Mean	S.D.	Range	Mean	S.D.	Range
Total number of utterances	98	30.3	52–152	105	33.9	63–182
IPSyn nouns Score	20.11	1.18	17–22	20.51	0.89	18–22
IPSyn verbs Score	27.33	3.86	19–33	29.00	2.56	23–33
IPSyn sentences score	26.77	3.93	19–34	29.40	3.55	21–37
IPSyn total score	74.48	7.64	62–88	79.14	5.08	69–87
MLU-m	5.82	1.32	4.24–8.96	6.86	0.74	5.26–8.13
NDW-50	122.7	20.1	78–162	137.8	15.2	97–164

Note: IPSyn: index of productive syntax (Scarborough, 1990); MLU-m: mean length of utterance in morphemes; NDW-50: number of different word roots in 50 utterances.

Table 2

Correlation coefficients and standardized function coefficients for discriminant analysis of language sample measures as predictor variables for normal or disordered language group membership

	Correlation coefficients with discriminant function	Pooled within-class standardized canonical coefficients
IPSyn noun	−0.0675	−0.456
IPSyn verb	0.335	0.001
IPSyn sentence structure	0.441	−0.011
IPSyn total	0.607	0.362
MLU-m	0.848	1.585
NDW-50	0.844	−0.699

the function correctly classified 74% of cases (note: when a discriminant function is applied to equal numbers in two groups, the percentage is necessarily the same for each—that is, if it misidentifies seven in one group, it will misidentify seven in the other). Cross-validation derived 63% accuracy for the total sample.

4. Discussion

In this study we sought to find evidence to support the validity of three measures of preschool language with children beyond preschool age. Group means for children with language impairments were significantly lower than those of children who were normally developing on grammatical and lexical measures. Areas identified as significantly different between the groups were: mean length of utterance in morphemes; sentence structure, as analyzed by the IPSyn; and number of different words in a 50 utterance sample. These findings provide some preliminary support for language sample analysis as a tool for identifying language impairment. If more normative information were to become available, these measures might show promise in providing norm-referenced yet ecologically valid means of identifying impairment.

Our results are similar to those of Leadholm and Miller (1992) who reported language sample results for a group of 35 normally developing 6-year-old children, as part of a larger set of data on school age children's performance on language sampling. Our means for MLU-m were somewhat higher than theirs: they reported a mean MLU-m for normally developing 6-year olds of 5.49 (S.D. 0.97) for conversational samples, and 6.17 (S.D. 1.20) for narrative ones. Our means were 5.73 for SLI (S.D. 1.30), and 6.76 for ND children (S.D. 0.66). Based on −1 S.D. from Leadholm and Miller's means, one would expect children with language impairment to show MLU's of less than 5.0. This was not the case in our data, where some children had means below 5.0, but as the overall mean shows, more did not. As previously discussed, we did not distinguish narrative from conversational samples, owing to the conversational style with which the story retellings were elicited. Johnston (2001) demonstrated considerable variability in MLU, with more impact on children with language impairment than normally developing children, when more restrictive utterance selection is used. We believe that our conservative utterance inclusion criteria contributed to the difference

Table 3
IPSyn total score ranges by group

IPSyn score range (points)	Number of SLI children scoring within range	Number of normally developing children scoring within range
80–89	8	15
70–79	12	11
60–69	7	1
<60	0	0

between our means and those of Leadholm and Miller, as would be predicted by Johnston's (2001) results.

While our predictions of group differences in language sample measures were supported, it is important to note that not all children diagnosed with language impairments using the EpiSLI criteria scored lower than normally developing peers. Moreover, not all children identified by those criteria as normally developing scored in the higher ranges. Tables 3–5 show the numbers of children in the sample falling in various score ranges. It is evident that children labeled ND show clustering at the higher ends of the ranges, whereas children labeled SLI show more variability, with a tendency to cluster closer to the bottom. For example, when IPSyn total score ranges are compared, half as many children with SLI scored above 80 (eight children with SLI as opposed to 16 peers had total scores above 80 points). This result provides some support for the modified IPSyn total score as tapping meaningful differences between the groups. However, the eight children with SLI who did score above 80 must be accounted for in some manner. Similar issues arise for MLU-m and NDW-50. Although this descriptive listing is revealing of differences, it also reveals overlap, in both directions—children labeled ND with lower scores on these measures, and children labeled SLI with higher ones.

While our data cannot be used normatively, for purposes of discussion and exploration we engaged in an exercise in “normative” comparison, purely for the purposes of exploring the potential validity of these measures. Table 6 presents means and standard deviations for MLU-m, IPSyn, and NDW-50 for the 27 normally-developing children. If we treated the data from the ND children as normative, obviously purely for the sake of argument, cut-offs set at -1.25 S.D. derived from the data on these children would be 5.94 for MLU-m, 72 for IPSyn, and 119 for NDW-50 (with rounding for the latter two as only whole number scores are possible). (Note that the 1.25 criterion was selected to accord with that suggested by Tomblin et al., 1996.) If one applies these “criteria” to the 27 children

Table 4
MLU-m ranges by group

MLU-m	Number of SLI children scoring within range	Number of normally developing children scoring within range
8.0 and above	2	2
7.0–7.9	3	9
6.0–6.9	6	13
5.0–5.9	6	3
4.0–4.9	10	0

Table 5
NDW ranges by group

NDW-50	Number of SLI children scoring within range	Number of normally developing children scoring within range
150 and above	2	6
140–149	4	6
130–139	4	13
120–129	5	1
110–119	5	0
100–109	5	1
<100	3	1

with SLI in our sample, then MLU-m identifies 18 children (67%), NDW-50 12 children (44%), and IPSyn 10 children (37%). An informal investigation of sensitivity and specificity using the -1.25 S.D. cut-off on at least two of the three measures revealed that 12 children with SLI (44%) would be identified as impaired by this criterion, versus only two children with ND (7%). Thus while the sensitivity of these measures using these cut-offs is poor, the specificity is high. To reiterate, appropriate normative cut-offs would be derived from full-range sampling, impossible in a matched pairs data set such as this, which does not reflect the proportion of SLI to be expected in a normally-distributed population. These numbers show that some children who have been identified as language impaired using formal tests would not be identified using these language sample measures. Improved sensitivity and specificity may not be possible without alternative measures that more precisely target the deficits of children with SLI, such as the morpho-syntactic profile described by Rice et al. (2000), or possibly by eliciting discourse from highly challenging contexts (Hadley, 1998).

The results of the discriminant function analysis provide support for the use of three factors, MLU-m, IPSyn total score, and NDW-50 to distinguish between the groups. The generation of a significant function with 74% accuracy is suggestive. It is, however, preliminary, given the small sample, the lack of an independent sample for the best cross-validation of the function, and its undesirably low rate of accuracy.

Our results provide some support for use of language samples to identify kindergarten children as language impaired. Of the children in our sample who scored more than one standard deviation below the mean for conversational MLU-m for 6 year olds as reported in the Leadholm and Miller (1992) database, all were also identified as SLI using the EpiSLI criteria. Yet the majority of children labeled SLI did not fall more than one standard deviation below the Leadholm and Miller mean for 6-year olds (possibly owing to utterance inclusion differences, as discussed above). It is worth noting that the mode for

Table 6
Means and standard deviations for MLU-M, IPSyn, and NDW-50 for ND children

	MLU-M	IPSyn	NDW-50
Mean	6.86	79	138
S.D.	0.74	5.09	15.2
-1.25 S.D.	5.94	72	119

children with SLI was in the 4.0–4.9 range (12), while the mode for the ND children was in the 6.0–6.9 range (14). Our findings thus provide some support for the possibility of developing a criterion cut-off, below which kindergarten children might be considered language impaired, or at least at risk. In our data, such a cut-off would be set at below 5.0, but further studies are needed to validate any criterion. Eisenberg et al. (2001) analysis of available data for MLU-m for preschool children indicated that low MLU-m can be diagnostic of disorder, and thus our data accord with previous findings relative to younger populations. In contrast, our data do not support the use of any of these language sample measures to rule out possible language impairment. Several children diagnosed with SLI by the formal test criteria used in the EpiSLI approach scored as high as the control children on at least one of the three measures. Although the utility of mean length of utterance in morphemes as a measure of language competence beyond the early years has been repeatedly challenged (e.g. Rollins et al., 1996), our data suggest that a low MLU-m holds promise as an indicator of language difficulties in kindergartners.

Our data offer limited validation for the Index of Productive Syntax (Scarborough, 1990) as a potential tool for children older than 4 years. The lack of significant results for the Noun and Verb subscales suggests items used on the IPSyn for these scales are at ceiling by age 6. It should be noted, however, that findings for the Verb Phrase Subscale indicated a weakness in at least some of the children. The significant results for the Sentence Structure Subscale suggest that measures concentrating on grammatical development are promising as a means to identify language disorder in school age children. In future work, we hope to examine in more detail syntactic structures used by the children in the study, including an examination of the sensitivity and specificity of items on the IPSyn. Such an analysis might yield a subset of items with greater utility in identifying children at risk for disorders. This approach might be enhanced by identifying morphosyntactic structures already shown to be difficult for English-speaking children with SLI, such as tense marking (Rice & Wexler, 1996; Rice et al., 2000). An instrument specifically targeting known grammatical weaknesses in preschool and young school age children has recently been published (Rice & Wexler, 2001).

The lexical measure used in this study, NDW-50, has been criticized in recent years, despite some success in studies that demonstrated it distinguished between children with and without language impairments (Klee, 1992a; Watkins et al., 1995). Richards and Malvern (1997) argued that NDW suffers from a serious threat to validity, in that children with shorter MLU's will tend to have shorter NDW's, because they will produce fewer words overall in any given sample of fixed length. Richards and Malvern propose an alternative metric using a mathematical formula that would use all available data but would be impervious to sample size. Recent work by Owen and Leonard (2001) did not support the claim that the alternative measure is not affected by sample size. Nonetheless, it would be instructive to investigate lexical diversity in our samples further using the Richards and Malvern measure, in order to exert some degree of control for utterance length. Indeed, previous work has found NDW-50 to be correlated with MLU-m (Klee, 1992a; Watkins et al., 1995). Thus, there is reason to be concerned that it does not tap lexical diversity as directly as might be wished, and we may need to look further to find a language sample measure that assesses semantic competence independent of morphosyntax. Candidate measures might look at lexical richness (e.g., comparing words used to lists of common

vocabulary to derive a lexical rarity metric—Paul, 2001); presence of derivational morphology; or attempt a semantic-pragmatic analysis, looking at use of nominals for referencing (e.g., post-nominal modification, and pronoun use).

4.1. *Limitations*

One limitation to these results stems from our inclusion criterion. Six children who produced fewer than 50 utterances were eliminated from the pool of candidates for matching, of whom 4 were children with SLI. This was done to ensure that sufficient utterances were present for analysis. It would be instructive to conduct a separate investigation of all children in the larger data set who produced fewer than 50 utterances, to discover whether the proportion of children with SLI who are unforthcoming is higher than the proportion of children with typical language development.

In looking at individual differences in scores on these measures, we found variability, with some children scoring high on at least one of the measures, despite having been identified as SLI using formal testing. There are a number of potential sources of the variability found in score ranges. A child identified by the EpiSLI criteria as SLI might score high on our measures, because either the language sample measures or the formal tests used in the EpiSLI protocol were in error. Sources of error that affect all assessments might have affected the results. In particular, measurement error in the eliciting of appropriate samples cannot be ruled out as a source of variability. Individual instances of poor samples may exist (perhaps explaining the few ND children with low NDW-50 scores, for example). Measurement error is difficult to minimize as much as would be desirable in language sampling, given the fluid nature of the interactions being sampled. There is evidence that language sampling can result in unacceptable variability of results. Chabon, Kent-Udolf, and Egolf (1982), using a picture description task, found considerable variability in MLU-m obtained over three days (although use of picture description as a sampling context could be considered a pragmatically questionable context, certainly quite different from either the play or conversational elicitations typically used). Eisenberg et al. (2001) discuss the serious threats to validity of MLU posed by reliability issues. The validity of language sample analysis relative to formal testing is challenged by these measurement issues.

Problems with sensitivity of the measures used in this study are signaled by the several children with SLI who scored in the upper ranges. In contrast, children with low scores on the language sample measures were much more likely to be classified as SLI, indicating that these measures have the potential to exhibit specificity, if empirically derived criteria for cut-off scores were to become available. For example, looking at Table 3, it can be seen that only one ND child scored below 70 on the IPSyn total score, compared to seven children with SLI. The IPSyn in particular appears to lack sensitivity, given that eight children with SLI had scores at 80 or above. As discussed previously, deriving a subset of items would likely improve its sensitivity in identifying older children with language impairments.

We have discussed error sources from the measures and from the samples, but there is also a possible third source of variability. Conflicting findings may in fact reflect real differences caused by linguistic profiles with scatter. For example, a child might do poorly on a formal test because it taps metalinguistic skills that are an area of weakness, whereas spontaneous measures highlight an area of strength. This issue gets to the heart of the

identification of disorder using norm-referenced approaches (Burroughs and Tomblin, ND). Even if a cut-off can be derived, either a criterion set by age (Eisenberg et al., 2001), or a certain number of standard deviations below the mean, it is still difficult to decide which measures are most accurate when language sampling and formal test scores do not agree. It is for this reason that it would appear that the most conservative approach would be to investigate further any finding of potential disorder on any one measure, whether of language sampling or formal testing. When a child scores below his or her peers, a full investigation into the child's communicative competence in meeting linguistic demands of his or her environment is warranted. Whatever normative measure is used, whether sample-derived or test-derived, it must be validated by examining real world child adaptive functioning, of which even sampling is at best an indirect and incomplete probe (Paul, 2001).

5. Conclusion

In summary, this study provides evidence that mean scores on language sample measures are lower for 6-year-old children with SLI than they are for their peers who are normally developing. This accords with the many studies that have found children with SLI to have disorders of morphosyntactic development, persisting well beyond the preschool years (e.g., Gopnik & Crago, 1991; King & Fletcher, 1993; Oetting & Horohov, 1997; Rice et al., 1995; Watkins & Rice, 1994). Our results indicate that traditional language sample analysis measures have the potential to be used with older children in diagnostic protocols. Given the individual differences and lack of sensitivity of these measures, additional and novel types of analyses must also be sought. Further investigation of reliability of samples, and how best to select utterances to ensure appropriateness for measurement (Johnston, 2001), is needed. The issues raised by Hadley (1998) pose further challenges for language sampling with school age children, in that appropriately challenging contexts must be created in order to uncover language deficits. Finally, the full diagnostic potential of any language sample measure will not be realized until we have access to normative information on a large scale.

6. Self-study questions

1. Compared to formal tests, language sample analysis has been argued to demonstrate:
 - A. More ecological validity.
 - B. Less ecological validity.
 - C. More ethnographic validity.
 - D. Improved test-re-test reliability.
 - E. More time efficiency.
2. Research on the language abilities of children with specific language impairment has consistently identified which area of language to exhibit the most persistent impairments?
 - A. Lexical development.
 - B. Pragmatic ability.
 - C. Fast mapping.

- D. Morphosyntactic ability.
 - E. Literacy.
3. The results of this study offer preliminary evidence that:
- A. Children with SLI older than 48 months do not show differences from age-matched typical peers in language sample measures.
 - B. Language sample analysis holds promise in the assessment of children over the age of 5.
 - C. Most children with SLI do not respond to conversational prompting in language sample contexts.
 - D. Noun phrase development is an important clinical marker for SLI.
 - E. Language sample measures are superior to formal testing in the identification of kindergarten children with SLI.
4. In the population studied, children with SLI:
- A. Sometimes scored below the typical children on measures of language.
 - B. Always scored below the typical children on measures of language.
 - C. Scored about the same as typical children on measures of language.
 - D. Scored the same as typical children on syntactic but not semantic measures.
 - E. Scored the same as typical children on semantic but not syntactic measures.
5. The results of this study support:
- A. Use of the IPSyn Questions and Negation Subscale as a standard diagnostic tool for kindergarten children.
 - B. Use of the NDW in preference to the PPVT in diagnosing SLI in 5-year olds.
 - C. Use of measurements of conversational pragmatics, such as turn-taking, to identify children with SLI.
 - D. Limiting use of formal testing for identifying SLI in kindergarteners.
 - E. A continued role for language sampling in assessing kindergarten children with suspected language disorders.

Acknowledgements

This study was supported by a contract NIH-DC-19-90 from the National Institute on Deafness and Other Communication Disorders and a Clinical Research Center NIDCD PODC02746. The assistance of Alahna Cogburn, Heather Corbett, Shannon Ross, and Kerry Frey in data management and analysis is gratefully acknowledged.

References

- Aram, D., Morris, R., & Hall, N. (1993). Clinical and research congruence in identifying children with specific language impairment. *Journal of Speech and Hearing Research*, 36, 580–591.
- Brown, R. (1973). *A first language*. Cambridge, MA: Harvard University Press.
- Chabon, S., Kent-Udolf, L., & Egolf, D. (1982). The temporal reliability of Brown's mean length of utterance measure with post-stage V children. *Journal of Speech and Hearing Research*, 25, 124–128.
- Crain, S., & Lillo-Martin, D. (1999). *An introduction to linguistic theory and language acquisition*. Malden, MA: Blackwell Publishers.

- Culatta, B., Page, J., & Ellis, J. (1983). Story retelling as a communicative performance screening tool. *Language, speech, and Hearing Services in Schools, 14*, 66–74.
- Dunn, M., Flax, J., Sliwinski, M., & Aram, D. M. (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). The use of MLU for identifying language impairment in preschool children: A review. *American Journal of Speech-Language Pathology, 10*, 323–342.
- Evans, J., & Craig, H. (1992). Language sample collection and analysis: Interview compared to freeplay assessment. *Journal of Speech and Hearing Research, 35*, 343–353.
- Fey, M. (1986). *Language intervention with young children*. San Diego, CA: College-Hill Press.
- Hadley, P. (1998). Language sampling protocols for eliciting text-level discourse. *Language, Speech, and Hearing Services in Schools, 29*, 132–147.
- Goldsworthy, C. & Secord, W. (1982). *MILI: Multilevel informal language inventory*. San Antonio, TX: The Psychological Corporation.
- Gopnik, M., & Crago, M. (1991). Familial aggregation of a developmental language disorder. *Cognition, 39*, 1–50.
- Johnston, J. (2001). An alternate MLU calculation: Magnitude and variability of effects. *Journal of Speech Language Hearing Research, 44*, 156–164.
- 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.
- King, G., & Fletcher, P. (1993). Grammatical problems in school-age children with specific language impairment. *Clinical Linguistics and Phonetics, 7*, 339–352.
- Klee, T. (1992a). Developmental and diagnostic characteristics of quantitative measures of children's language production. *Topics in Language Disorders, 12*(2), 28–41.
- Klee, T. (1992b). Measuring children's conversational language. In S. Warren & J. Reichle (Eds.), *Causes and effects in communication and language intervention* (pp. 315–330). Baltimore, MD: Paul H. Brookes Publishing.
- Leadholm, B. J., & Miller, J. F. (1992). *Language sample analysis: The Wisconsin guide*. Madison, WI: Wisconsin Department of Public Instruction.
- Leonard, L.B. (1998). *Children with specific language impairment*. Cambridge, MA: MIT Press.
- Lund, N. J., & Duchan, J. F. (1993). *Assessing children's language in naturalistic contexts*. Englewood Cliffs, NJ: Prentice Hall.
- Mertler, C., & Vannatta, R. (2002). *Advanced and multivariate statistical methods* (2nd ed.). Los Angeles, CA: Pyrczak Publishing.
- Miller, J. (1991). Quantifying productive language disorders. In J. F. Miller (Ed.), *Research on child language disorders: A decade of progress* (pp. 211–220). Austin, TX: Pro-Ed.
- Miller, J. (1996). Progress in assessing, describing, and defining child language disorder. In K. Cole, P. Dale, & D. Thal (Eds.), *Assessment of communication and language* (pp. 309–324). Baltimore, MD: Brookes.
- Miller, J., & Chapman, R. (1981). The relation between age and mean length of utterance in morphemes. *Journal of Speech and Hearing Research, 24*, 154–161.
- Miller, J. & Chapman, R. (2000). *Systematic analysis of language transcripts* (Vol. 6.01). Madison, WI: Language Analysis Lab, University of Wisconsin.
- Naremore, R., Densmore, A., & Harman, D. (1995). *Language intervention with school-aged children*. San Diego, CA: Singular Publishing.
- Newcomer, P., & Hammill, D. (1988). *Test of Language Development-2 Primary*. Austin, TX: Pro-Ed.
- Oetting, J. B., & Horohov, J. E. (1997). Past tense marking by children with and without specific language impairment. *Journal of Speech and Hearing Research, 40*, 62–74.
- Owen, A. & Leonard, L. (2001). Lexical diversity in the speech of normally developing and specific language impaired children. *Poster presented at the Symposium for Research in Child Language Disorders*, Madison, WI.
- Owens, R. (1999). *Language disorders: A functional approach to assessment and intervention* (3rd ed.). Boston, MA: Allyn and Bacon.
- Paul, R. (2001). *Language disorders from infancy to adolescence* (2nd ed.). St. Louis, MO Mosby.

- Rice, M. L., & Wexler, K. (1996). Toward tense as a clinical marker of specific language impairment in English-speaking children. *Journal of Speech and Hearing Research, 39*, 1239–1257.
- Rice, M. L. & Wexler, K. (2001). *Rice-Wexler test of early grammatical impairment*. San Antonio, TX: The Psychological Corporation.
- Rice, M. L., Wexler, K., & Cleave, P. (1995). Specific language impairment as a period of extended optional infinitive. *Journal of Speech and Hearing Research, 38*, 850–863.
- Rice, M., Wexler, K., & Hershberger, S. (1998). Tense over time: The longitudinal course of tense acquisition in children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 41*, 1412–1431.
- Rice, M., Wexler, K., Marquis, J., & Hershberger, S. (2000). Acquisition of irregular past tense by children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 43*, 1126–1145.
- Richards, B. & Malvern, D.D. (1997). *Quantifying lexical diversity in the study of language development*. Reading, UK: University of Reading, New Bulmershe Papers.
- Rollins, P. R., Snow, C., & Willett, J. (1996). Predictors of MLU: Semantic and morphological developments. *First Language, 16*, 243–259.
- Rosenthal, R., & Rosnow, R. L. (1984). *Essentials of behavioral research: Methods and data analysis*. New York: McGraw-Hill.
- Scarborough, H. (1990). Index of productive syntax. *Applied Psycholinguistics, 11*, 1–22.
- Scarborough, H., Wyckoff, J., & Davidson, R. (1986). A reconsideration of the relation between age and mean utterance length. *Journal of Speech and Hearing Research, 29*, 394–399.
- Stockman, I. (1996). The promises and pitfalls of language sample analysis as an assessment tool for linguistic minority children. *Language, Speech, and Hearing Services in Schools, 27*, 355–366.
- Tomblin, B., Records, N. L., Buckwalter, P. R., Zhang, X., Smith, E., & O'Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research, 40*, 1245–1260.
- Tomblin, B., Records, N. L., & Zhang, X. (1996). A system for the diagnosis of specific language impairment in kindergarten children. *Journal of Speech and Hearing Research, 39*, 1284–1294.
- Watkins, R., Kelly, D., Harbers, H., & Hollis, W. (1995). Measuring children's lexical diversity: Differentiating typical and impaired language learners. *Journal of Speech and Hearing Research, 38*, 1349–1355.
- Watkins, R., & Rice, M. L. (1994). *Specific language impairments in children*. Baltimore, MD: Paul H. Brookes Publishing Co.
- Wechsler, D. (1989). *WPPSI-R manual: Wechsler preschool and primary scale of intelligence—revised*. New York: Psychological Corporation.
- Windsor, J., Scott, C., & Street, C. (2000). Verb and noun morphology in the spoken and written language of children with language learning disabilities. *Journal of Speech, Language, and Hearing Research, 43*, 1322–1336.
- World Health Organization (2001). *ICF Checklist, Version 2.1a, for International Classification of Functioning, Disability, and Health*. Available on-line at <http://www.who.int/classification/icf/checklist/icf-checklist.pdf>. (Accessed 10/15/03).