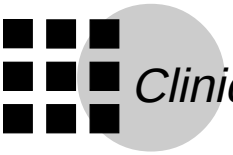

Clinical Consult

The Use of MLU for Identifying Language Impairment in Preschool Children: A Review

Sarita L. Eisenberg*
Montclair State University, Upper Montclair, NJ

Tara McGovern Fersko
Steppingstone Day School, New York, NY

Cheryl Lundgren*
Eliot Hospital, Manchester, NH

The following data were reported in an assessment of James, a boy aged 3;2 (years;months):

> James scored 28 on the Auditory Comprehension Subtest of the Preschool Language Scale. This corresponds to a standard score of 90, which is less than 1 *SD* below the mean, and is at the 25th percentile for his age. The Expressive Communication Subtest could not be scored since James did not complete all of the items. During 15 minutes of play, James produced 28 utterances that were completely intelligible. Mean Length of Utterance was 3.16, which is 1.25 *SD* below the mean for his age.

Based on these results, the speech-language pathologist concluded that James demonstrated an expressive language disorder.

Speech-language pathologists are frequently asked to determine whether or not a child has a language problem. Although a large number of norm-referenced standardized tests have been developed to answer this question, many young children, like James, cannot be tested using formal procedures. In addition, these tests have been criticized as being psychometrically inadequate (McCauley & Swisher, 1984) and unable to differentiate between typically developing children with normal language (NL children) and children with a language impairment (LI children) (Plante & Vance, 1994).

The speech-language pathologist in the example used Language Sample Analysis (LSA) as an alternative to standardized testing for evaluating James' expressive language (Bernstein & Tiegerman-Farber, 1997; Lahey, 1988; Miller, 1981; Nelson, 1998). This is consistent with the results of several surveys that suggest that the use of LSA with preschool children is increasing. On a 1993 survey, 80% of the responding speech-language pathologists reported using LSA to supplement standardized testing (Hux, Morris-Friehe, & Sanger, 1993). A 1997 survey reported a somewhat higher percentage, with 85% of the respondents indicating that they used LSA (Kemp & Klee, 1997). On a more recent survey, 93% of speech-language pathologists reported using LSA (Loeb, Kinsler, & Bookbinder, 2000). Mean length of utterance (MLU) was the most frequently listed LSA procedure, with 91% usage (Loeb et al., 2000).

On these surveys, one of the main reasons reported for using LSA was to identify a language disorder (Kemp & Klee, 1997; Loeb et al., 2000). This use for LSA is recommended in many textbooks (Bernstein & Tiegerman-Farber, 1997; Lahey, 1988; Miller, 1981; Nelson, 1998). However, not all textbooks agree that LSA is appropriate for this purpose. Paul (2000) noted the difficulty in establishing reliability for naturalistic language sampling. She suggested that LSA is best suited for describing a child's language problem, and that the more limited aim of identifying a language impairment can be accomplished more efficiently with standardized tests. Although Paul agreed that there are no standardized language tests for some age levels and populations, and that "many tests in the language area are not constructed as well as they might be" (p. 4), she concluded that "standardized testing is the *only* valid, reliable and fair way to establish that a child is significantly different from other children" (p. 43).

A number of psychometric guidelines, listed in Table 1, have been suggested as properties to look for when evaluating norm-referenced standardized tests (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985; Hutchinson, 1996; McCauley & Swisher, 1984). Test users expect to find information concerning these properties

*At the time of the study, Sarita Eisenberg was affiliated with William Patterson University in Wayne, NJ, and Cheryl Lundgren was affiliated with Columbia Presbyterian Medical Center in New York.

1058-0360/01/1004-0323

TABLE 1. Guidelines for evaluating assessment tools.

Clear definition of purpose
Sufficient description of administration and scoring procedures
Sufficient description of the normative sample
Appropriate reference data
Evidence of reliability
Evidence of validity

in a test's manual. This information is important because it allows test users to replicate test procedures, to evaluate a test's usefulness, and to determine a test's appropriateness for a particular child. These guidelines have been shown to apply to criterion-referenced tests (McCauley, 1996). In this paper, we suggest that they also apply to "formal" or "standardized" LSA procedures (terms used by Hux et al., 1993, and Kemp & Klee, 1997), such as MLU, and must be evaluated if we are to have confidence in the decisions that we make based on such analyses.

There is, however, no such manual containing this information for MLU, making it difficult for clinicians to evaluate the usefulness and appropriateness of this measure. In this paper, we summarize the available information on MLU. This information comes from two reports that have been used as reference data for MLU (Leadholm & Miller, 1992; Miller & Chapman, 1981), as well as from a number of other studies that have evaluated MLU for one of these properties. The paper is organized to address each of the properties listed in Table 1. We first delimit the specific purpose of MLU that this paper will address. The next section on administration and scoring procedures reviews studies focusing on the impact of different elicitation conditions and scoring conventions on the obtained MLU. In the next two sections, we consider the adequacy of the population samples used to develop the reference data and then interpret the data. We then review studies of reliability and discuss various aspects of validity. Our aim is to determine if there is evidence that MLU can be used for the purpose of identifying a language disorder in preschool children and how confident we can be in making that interpretation (Plante, 1996).

The case at the beginning of this paper illustrates a number of potential misuses of MLU. The clinician in this case diagnosed a language disorder based on only a single LSA measure. MLU was based on a small number of utterances obtained during a brief time period. The nature of the sampling condition was not fully specified, nor was the source of the reference data stated. By consolidating the information about MLU, we hope to enable clinicians to evaluate their own clinical use and implementation of MLU.

Purpose

The adequacy of any assessment tool can only be evaluated relative to its intended purpose. There are three aspects of purpose: domain, population, and assessment aim.

Domain

The domain or trait is what we are trying to measure. In delimiting a domain, it is important to be conservative and to not overstate the trait being measured (see, for example, Gray, Plante, Vance, & Hendrichsen, 1999). It is also important to separate the means for measuring the trait from the trait itself (Sabers, 1996). Thus, MLU should not be regarded as a measurement of morphosyntax, but should instead be viewed as one of several possible ways of measuring utterance length.

The confusion about this may stem from Brown's observation that "almost every new kind of knowledge increases length" (1973, p. 53). This statement may be interpreted in different ways. It is not always the case that longer utterances are necessarily more syntactically sophisticated than shorter ones, or that grammatically more advanced utterances are necessarily longer than less advanced ones.

Consider the following two utterances:

1. want more cookies Mommy
2. I want to go home

Both utterances are five morphemes in length. However, sentence 1 is an ungrammatical simple sentence, whereas sentence 2 is a grammatical sentence containing two clauses in an embedded relationship. This dissociation between utterance length and morphosyntax has been demonstrated empirically as well. Researchers have found significantly different rates of morpheme use within a group of typically developing children at the same age and MLU (Rollins, Snow, & Willett, 1996) and between LI children and younger NL children matched on MLU (Rice, Rice, & Redmond, 2000; Rollins, 1995). MLU should, therefore, not be thought of as a measure of normal or "delayed syntactic development" (Miller & Chapman, 1984–2000), but should "be recognized for what it is: a measure of average utterance length" (Klee, 1992, p. 327).

Population

We will refer to two sets of reference data for MLU. Miller and Chapman (1981, abbreviated as MC) reported MLU data for children between the ages of 18 and 60 months. Leadholm and Miller (1992, abbreviated as LM) reported MLU data for children from 3 to 13 years of age. These latter data are the reference database for the Systematic Analysis of Language Transcripts (SALT) program (Miller & Chapman, 1984–2000)

The applicability of MLU for this entire age range has been questioned. Bernstein and Tiegerman-Farber (1997) suggested that MLU is useful only up to a ceiling of approximately four to five morphemes, corresponding to an upper age limit between 45 and 54 months for typically developing children. Bloom and Lahey (1978) questioned the applicability of MLU greater than 3.0, corresponding to an upper age limit of approximately 36 months. Above this MLU, there is increased variability reflected in larger standard deviations (Miller & Chapman, 1981) and a larger standard error of measurement (SEM; Rondal, Ghiotto, Bredart, & Bachelet, 1987). Brown (1973) suggested that by Stage V (when MLU reaches 4.0), utterance length would reflect the character of the particular interaction rather than new linguistic knowledge. Conversely, other authors have concluded that MLU is a valid developmental measure into the school years (Jones, Weismer, & Schumacher, 2000; Miller, Frieberg, Rolland, & Reves, 1992). The current paper will focus on the use of MLU for preschool children only.

Assessment Aim

There have been a number of aims suggested for MLU: to identify children in need of further language evaluation (Miller & Chapman, 1981), to diagnose or identify a language impairment (Bernstein & Tiegerman-Farber, 1997; Lahey, 1988; Miller, 1981; Nelson, 1998; Owens, 1999), to determine stage or overall level of language development (Bernstein & Tiegerman-Farber, 1997; Miller, 1981; Owens, 1999), to guide further language assessment (Paul, 2000), to select goals (Lahey, 1988; Miller, 1981; Owens, 1999), to compare language use across situations (Lund & Duchan, 1993; Owens, 1999), and to measure change in language production (Fey, 1986; Paul, 2000). The validity of MLU needs to be established separately for each of these uses if the measure is to be used clinically. The current paper focuses only on the aim of identifying a language impairment.

Administration and Scoring Procedures

Procedural standardization enables test users to replicate the procedures that were used in developing normative data. As we surveyed the procedures for language sampling that have been presented in textbooks, we were struck with the variability of these recommendations. In an effort to increase sample representativeness, most textbooks recommended eliciting language samples in at least two different interactional contexts. This is because language sampling is discussed for a variety of aims rather than solely for the aim of identifying language disorder or, more specifically, for calculating MLU. The assumption is that the same language sample can be used for making a quantitative norm-referenced comparison and for carrying out a qualitative description of the child's language production (e.g., Lahey, 1988). However, this may not be appropriate. When a child's performance will be quantified for comparison purposes, it is important to replicate the specific procedures used for gathering the comparison data (Nelson, 1998). If normative data are used for comparison without duplicating the procedures that were followed during standardization and the development of norms, "the test taker may be given an unfair advantage or may be unfairly penalized by differences in instructions, surroundings, and so forth" (McCauley & Swisher, 1984, p. 39). Even for measures without norms, it is important for the speech-language pathologist to know and follow administration and scoring procedures in order for the measure to function as it is supposed to (McCauley, 1996).

Administration

Adequate standardization for obtaining language samples requires specification of sample size, setting, participant, instructions given to interactants, the activity, and materials. Both MC and LM used a conversational sampling procedure. LM also used narrative sampling. Table 2 provides the specific procedures used by MC and LM to obtain the language samples (also the procedures for SALT). Of note is that there were significant differences in how these samples were collected, even between the two conversational procedures. This is important because the MLU values obtained from language sampling have been shown to be influenced by setting, participants, and activity (Haynes, Purcell, & Haynes, 1979; Kramer, James, & Saxman, 1979; Olswang & Carpenter, 1978; Scott & Taylor, 1978;

TABLE 2. Procedures for eliciting language samples.

Procedural Variables	MC	LM/SALT Conversational Context	LM/SALT Narrative Context
Sample size	• 50 utterances	• 100 utterances	• 100 utterances
Sampling time	• 10–15 minutes over age 2 • 20 minute maximum for age 2	• 15 minutes • provide separate data on 12-minute sampling time	• 15 minutes • provide separate data on 12-minute sampling time
Setting	• child's home, an experimental playroom, or a therapy room	• therapy room	• therapy room
Interactant	• parent	• SLP	• SLP
Instructions	• mothers instructed "to play with toys as they usually did"	• provide sample questions and prompts to facilitate child talk • "introduce at least one topic absent from the time and space of the sampling condition"	• provide questions and prompts to facilitate child talk
Activity	• play with toys	• play with clay • question about classroom and other activities	• tell a favorite story • retell an episode of a tv program • retell a familiar story
Materials	• set of toys that "varied from study to study, but always included both novel and familiar ones"	• clay	• none • for 3-yr-olds, pictures may be used for the story retelling

Note. MC = Miller & Chapman 1981 study; LM = Leadholm & Miller 1992 study; SALT = Systematic Analysis of Language Transcripts.

Stalnaker & Creaghead, 1982; Wagner, Nettelbladt, Sahlen, & Nilholm, 2000).

Setting and interactant. Two studies have reported that MLU is larger for samples that are elicited at a child's home than for samples that are elicited at a clinic. Scott and Taylor (1978) studied typically developing children between the ages of 2;1 and 5;1 with MLUs in the range of 3.5 to 6.0. Seven of the 12 children produced a higher MLU in the home condition. However, although statistically significant, the differences were small, with only 3 of the children showing an MLU difference as large as .51 to .83. The authors also noted an interaction of elicitation condition with MLU level because most of the children who achieved higher MLUs at home had an MLU from 4.0 to 5.0 in the clinic sample. Kramer et al. (1979) studied children between the ages of 3 and 5 years, with clinic MLUs between 2.5 and 5.0, who had been referred for a speech and language evaluation. Eight out of 10 children produced higher MLUs in the home condition, and these differences were large, ranging from 1.48 to 3.66. In both of these studies, it is not possible to separate the influences of setting and interactant because the home samples were elicited by the mothers and the clinic samples were elicited by an unfamiliar examiner.

A study by Olswang and Carpenter (1978) did not find any group difference in MLU between conditions in which the samples were gathered by the mother or by unfamiliar interactants when both samples were collected in the clinic. Subjects for this study included 9 LI children aged 3 to 6 with MLU ranging from 1.5 to 3.0, which is comparable to the subjects in the Kramer et al. (1979) study. Individual data were not reported, so it is not possible to determine whether individual children conformed to the general finding. Bornstein, Haynes, Painter, and Genevro (2000) also found no difference in MLU for typically developing 2-year-old subjects (n = 33) between samples that were collected either at home or in the clinic when either the mother or an unfamiliar adult elicited the samples. A limitation of the Bornstein et al. (2000) study was that, for each setting, the child first interacted with the mother while the examiner was present. This may have reduced the degree of unfamiliarity and lessened the generalizability of this study to elicitation situations in which only an unfamiliar adult interacts with the child.

The large MLU differences observed by Kramer et al. (1979) suggest that a clinical population may be more susceptible to differences in elicitation condition. We found no

study that looked at the impact of different settings with a single examiner for LI children. Thus, it is possible that the crucial factor influencing MLU for the LI children may be the setting rather than the interactant. This presents a difficulty for replicating the procedures used by MC because they combined data from several studies that varied in regard to whether the samples were collected at home or at a clinic.

An additional variable that may affect a child's language is the interactant's race relative to that of the child. A study by Bountress, Bountress, and Tonelson (1988) included 42 children, aged 2;6 to 6;9, with MLU ranging from 4.20 to 4.55, who were evaluated at their day care center. There was no significant difference in MLU for both the African American and the Caucasian children, regardless of whether the examiner's race was African American or Caucasian. As these authors noted, the children were seen in a familiar setting. Additional studies are needed both in unfamiliar clinic settings and in the children's homes.

Activity and materials. MC used free play sampling. Their report did not specify the toys that were used, but it is likely that these reflected the toys suggested by Miller (1981), including eating utensils, dolls, a barn with appropriate animals, a gas station with vehicles, people figures, a school house and bus, and a house with furniture. These are toys with multiple pieces (cf. Bernstein & Tiegerman-Farber, 1997), the types of toys "that allow the child to construct a variety of activities" (Miller, 1981, p. 11). In contrast, LM's conversational sampling involved a construction activity with clay. We found disagreement in textbooks concerning the use of this latter activity type. Nelson (1998) cautioned against the use of manipulables such as puzzles or construction items because children tend to talk less when they are engaged in activities with these toys. Owens (1999), however, recommended these materials and reported that children talk more about non-present events during play with these toys, which is an aim of the LM sampling procedure. The suggestion in most textbooks is to tailor the choice of materials to the child's developmental level, using free play with toys with younger and less talkative children and using activities that would promote decontextualized conversation with older children.

Two studies illustrate the importance of using reference data only when the corresponding collection procedures have been followed. Bain, Olswang, and Johnson (1992) manipulated predictability and variety in two play situations with 6 LI children, aged 31–35 months, with MLU around 1.0 to 1.5. The children produced fewer multi-word utterances in a clinician-directed routine play condition involving a construction activity (making an animal) than in child-directed free play with a variety of toys. Stalnaker and Creaghead (1982) found differences in both the number of utterances and the MLU between conditions involving free play with toys and questioning during play for children aged 4;0 to 5;6 with MLU around 4.5 to 5.0. The more talkative children showed no difference in number of utterances, but the "shy children" produced more utterances in the questioning during play condition than in the play only condition. This is relevant because LI children are typically less talkative than NL children. However, MLU was lower in the questioning condition. Speech-language pathologists may choose to tailor the collection procedures to child characteristics. If they do, they must use only reference data that were gathered in the same way.

For their narrative sampling, LM introduced three topics for the child to talk about. There was no contextual support, although pictures could be used with the 3-year-olds to facilitate story retelling. LM provide separate data for the narrative sampling condition. This is necessary because MLU is higher in narratives than in conversation (Leadholm & Miller, 1992; Wagner et al., 2000). Data from the Stalnaker and Creaghead (1982) study show why these separate data are necessary. These authors found MLU to be higher in a story retelling condition than during either of their two play conditions. Note, too, that procedures for eliciting narratives may also differ. The procedure used by Stalnaker and Creaghead involved first telling the child a brief story while also acting out the story with toys and then immediately asking the child to retell the story, using the toys if the child wanted. LM included story retelling as one of three narrative topics, with the child retelling a story that he or she had previously heard and with contextual support in the form of pictures available only to the 3-year-olds. As noted above, although a variety of collection procedures are available, speech-language pathologists should only reference data that were gathered in the same way.

Sample size and sampling time. Although some textbooks recommend Brown's original sample size of 100 utterances (Lahey, 1988; Retherford, 1993), most textbooks conform to MC's recommendation for a minimum sample size of 50 utterances. Kemp and Klee (1997) reported that most speech-language pathologists use samples of 50 utterances. However, 25% of the speech-language pathologists in the

Hux et al. (1993) survey and 43% in the Loeb et al. (2000) survey indicated using samples of fewer than 50 utterances. We have spoken to clinicians who have used fewer than 25 utterances for calculating MLU. Only 15% of the speech-language pathologists in the Kemp and Klee survey used sample sizes greater than 100 utterances. These small sample sizes are problematic given the low test-retest reliability reported for sample sizes of fewer than 100 utterances (Gavin & Giles, 1996).

Twenty-eight percent of the speech-language pathologists surveyed by Hux et al. (1993) reported using sampling times of 15 minutes or less. However, these short sampling times may not yield sufficient utterances for calculating MLU. LM (1992) reported data for the number of utterances in a 12-minute sampling time after an initial warm-up. This sampling time yielded samples of fewer than 100 utterances from some of their 3-and 4-year-old typical children in the conversational context and from some children up to 7 years of age in the narrative context. This is problematic because LI children are often less talkative than NL children. Miller (1981) suggested that a 30-minute total sampling time would yield a sufficient number of utterances, and most textbooks reiterate this suggestion. Note, however, that LM used 15-minute collection times for each of their conversational and narrative samples, for a total of 30 minutes for both procedures.

Scoring

A standardized test should provide criterion for scoring items as well as designate the procedure for computing the total score. To calculate MLU, decisions must be made concerning utterance segmentation, utterance exclusion, and morpheme assignment (see Table 3). Most textbooks refer to the original rules provided by Brown (1973, p. 54) for utterance exclusion and morpheme assignment. Computer LSA programs assume that the user is familiar with the analysis procedures and so are designed only to teach the user the coding conventions of that program needed to carry out the analyses (Long, 1991).

Utterance segmentation. The calculation of MLU depends critically on how utterances are segmented. MC segmented utterances "primarily by apparent terminal intonation contour" (Miller & Chapman, 1981, p. 155). However, they report inter-examiner disagreement for 10–15% of the utterances, so this rule is not sufficient. LM also used pauses of greater than 2 seconds to determine utterance boundaries. Lund and Duchan (1993) and Owens (1999) suggested an additional strategy of using sentence structure such that word groups that would be considered as sentences are considered to be utterances. Owens also suggested using inhalation as a cue to utterance boundaries. LM added a rule for dealing with multiple conjoining in order "to avoid overly long utterances" (Leadholm & Miller, 1992, p. 28). Run-on sentences involving multiple conjoinings with *and* would be separated into utterances, each with no more than one clausal conjunction. This latter rule should not be followed when using the MC reference data. LM suggested that segmentation is relatively easy when "the child is producing only one utterance per speaking turn" (p. 27). However, we have found there to be considerable disagreement in determining whether vocatives and yes/no responses should be segmented as separate utterances or included as part of a longer utterance.

Utterance exclusion. Both MC and LM (indirectly through Miller, 1981) base their exclusionary criterion on Brown (1973). Most published textbooks also suggest using Brown's rules for determining utterance exclusion (Brown, 1973); some even reprint Brown's rules. Brown stated that only fully transcribed utterances were to be included, and that utterances "with blanks" were to be excluded. Based on this, both MC and LM exclude all totally and partially unintelligible utterances. Brown included exact utterance repetitions. However, some texts (Lund & Duchan, 1993; Owens, 1999; Paul, 2000; Retherford, 1993) suggest excluding imitations of the immediately prior adult utterance or exact self-repetitions. Lund and Duchan further suggest excluding identical utterances, elliptical responses to questions, counting and other sequences of enumeration, and single word or phrase social responses. These additional suggestions should not be followed if using the MC or LM reference data. MC did add one additional exclusion rule that must be followed when using their data. MC eliminated utterances with "a long string of conjoined words or phrases based on, for example, objects in the room" (Miller & Chapman, 1981, p. 156).

Brown (1973) stated that the utterance set for the MLU analysis should start no earlier than the second page of the sample or later "with the first recitation-free stretch" (p. 54) if the second page involved a recitation of some kind. The child's initial utterances would, therefore, be excluded. Textbooks suggest starting at the beginning of the sample (Paul, 2000) or selecting part of the sample that appears to be representative of the child's abilities (Retherford, 1993). Neither MC nor

TABLE 3. Scoring conventions for mean length of utterance.

Scoring Variables	MC	LM/SALT	Other
Utterance segmentation	• terminal intonation contour	• terminal intonation contour • document "thought completion" • pauses of greater than 2 s • only one independent clause	• sentence structure • inhalation
Utterance exclusion	• totally or partially intelligible utterances • exclude long strings of conjoined words or phrases	• totally or partially intelligible utterances	• immediate imitations of adult utterances • exact self-repetitions • identical utterances • elliptical responses to questions • counting sequences • single word or phrase social responses.
Morpheme assignment: Count as 1 morpheme	• bound inflections[a] • auxiliaries[a] • irregular past forms[a] • compound words[a] • proper names[a] • ritualized reduplications[a] • diminutives[a] • catenatives (gonna, wanna, hafta)[a] • words repeated for emphasis[a] • *no*, *hi*, and *yeah* (but not fillers)[a] • negated versions of the contracted auxiliaries *can't*, *don't*, etc., unless the non-negated forms are used[d]	• bound inflections[b] • auxiliaries[b] • irregular past forms[b] • compound words[b] • proper names[a] • ritualized reduplications[a] • diminutives[b] • catenatives (gonna, wanna, hafta)[b] • words repeated for emphasis[a] • *no*, *hi*, and *yeah* (but not fillers)[a] • multi-word titles[c]	• negative contractions unless each part of the contraction is used[d] • overgeneralizations of the regular past[c] • irregular plurals[c] • plural forms that do not have corresponding singular forms[c] • indefinite pronouns[c] • *-ing* forms (gerunds and participles) that are not part of the verb[d]

Note. MC = Miller & Chapman 1981 study; LM = Leadholm & Miller 1992 study; SALT = Systematic Analysis of Language Transcripts.

[a]Stated in Brown (1973).
[b]Specifically specified in Leadholm & Miller (1992).
[c]Not specified, but consistent with Brown (1973).
[d]Differs from Brown (1973).

LM say anything about this, so they may have started with the first utterance of the sample.

Morpheme assignment. Both MC and LM refer the reader to Brown (1973) for the morpheme assignment rules. Most textbooks cite these conventions as well, but there are some differences that speech-language pathologists need to be aware of (see Table 3). Particular mention needs to be made of the rules for contractions. Many speech-language pathologists follow the rule stated in Retherford (1993) that negative contractions are counted as one morpheme unless each part of the contraction is used elsewhere in the sample. However, this is not part of Brown's rules, and so should not be followed when using either the MC or the LM reference data. MC specified a different rule for negative contractions: that they be counted as one morpheme unless the non-negated forms are used. This less stringent rule should be followed when using the MC data. When typing a transcript for a computerized MLU calculation, each bound morpheme is marked so that that morpheme will be counted separately. Compound words, titles, and other elements to be counted as one morpheme are typed without any spaces or morpheme coding.

Raw score calculation. The calculation of MLU involves counting the morphemes in each utterance. Brown's rules specify that fillers and repetitions due to disfluency (referred to by LM as mazes) are not to be included in the morpheme count. All other elements, including vocatives and politeness markers, are counted.

The morphemes in each utterance are then summed and that sum is divided by the total number of utterances. Using a computer program, such as SALT, to do the calculation can save time and eliminate tabulation and mathematical errors, but it depends on the transcription having been done without any errors in morpheme marking and spacing.

Interpretation

MC suggest a cutoff point of 1 *SD* below the mean for identifying children who require further evaluation for possible delays in productive syntax. LM caution that –1 *SD* is not significant enough to identify a child as having a language impairment. They suggest that –1.5 to –2 *SD* be used, which is consistent with the cutoff used for standardized tests. Note that these are arbitrary cutoffs and they have not been empirically established.

Normative Sample

Sample Description

The normative sample must be sufficiently described so that clinicians can determine the representativeness of that sample for a particular child or for a type of child (Hutchinson, 1996; McCauley & Swisher, 1984). Important demographic information includes geographic region, socioeconomic status (as measured by parental education and income levels), race and ethnicity, gender, and normalcy.

Miller and Chapman (1981). MC combined data from five separate studies to investigate the relationship between MLU and age. This was therefore a research report rather than an attempt to develop norms. The five studies included a total of 123 children between 17 months and 5 years of age, in 3-month age intervals. The children all lived in Madison, Wisconsin, and were drawn from a predominantly middle to upper middle class population. Parental education level was unknown for one-third of the subjects. The other parents mostly had college degrees, but all had completed at least a high school level. Parental income level, race, and ethnicity were not specified. Gender composition was also not specified. Only children judged to be normally developing were included.

Leadholm and Miller (1992)/SALT (1984–2000). LM set out to develop local norms. This database included 100 children from 3 to 5 years of age. The 3- and 5-year-old groups included rural children as well as children from the Madison area. The demographic characteristics are not otherwise specified. Data are presented for whole-year age intervals. There were more boys than girls in each age group. Only typically developing children were included.

Sample Appropriateness

Sample size. The normative sample population needs to be of sufficient size to provide stable values. The parametric standards set forth by McCauley and Swisher (1984) prescribe a minimum sample size of at least 100 subjects. McCauley and Swisher further specify that any subgroup for which separate data are presented must meet this minimum sample size. Sample sizes for the age groups in both MC and LM are much smaller than this. In the MC data, the number of children per subgroup ranged from 1 to 16. Subgroups for the LM data were larger, ranging from 28 to 42, but were still well below the prescribed sample size. Given the variability in language production among typically developing children (see, for instance, Lahey, Liebergott, Chesnick, Menyuk, & Adams, 1992), we cannot be certain that these reference data adequately represent the range of normal performance. MC divided the subjects into 3-month age groups. The larger subgroup sample sizes for the LM data were at the expense of wider age intervals of 1 year. This is likely to be too wide an age range for children of this age.

Socioeconomic status. The populations sampled by both MC and LM were mostly middle class. Klee, Schaffer, May, Membrino, and Mougey (1989a) reported the same age to MLU relationship for 24 lower middle class children as MC (1981) had reported for their sample. Based on this, Klee et al. suggested that the MC data could be extended to lower middle class children. However, MC expressed concern that the MLU values that they obtained might be higher than for the general population. A recent study by Dollaghan et al. (1999) bears this out. On a normal distribution, 16% of the population score more than 1 *SD* below the mean, and 8% score more than 1.5 *SD* below the mean. Using the MC data, Dollaghan et al. found that significantly larger percentages of children fell below these cutoffs for mothers whose educational level was a high school degree or less. For mothers with a high school degree, more than 20% of their children had an MLU that was more than 1 *SD* below the mean for the MC data, and 16% were more than 1.5 *SD* below the mean. For mothers with less than a high school degree, more than 40% of their children had an MLU that was more than 1 *SD* below the mean, and almost 30% were more than 1.5 *SD* below the mean. These data show that using MLU values from children of more highly educated parents

may lead to an over-identification of children with less educated parents as having abnormally low MLU or even language impairment.

Gender. Both MC and LM combined data for boys and girls. Bornstein et al. (2000) found MLU to be higher for 2-year-old girls than for boys of the same age. Additional studies are needed to determine if this holds true for older children as well. If so, then using combined reference data could lead to an over-identification of boys as having low MLU.

Non-mainstream speakers. Neither MC nor LM specified the racial or ethnic composition of their standardization samples. In discussing standardized tests, several authors have indicated that the inclusion of non-mainstream children in a standardization sample does not mean that the reference data are appropriate for that population (see, for instance, Vaughn-Cooke, 1986; Westby, 2000) This issue is also relevant for MLU. In a study by Bountress et al. (1988), African American children scored lower on MLU than did Caucasian children. These authors attributed this to the optional use by the African American children of third-person singular, plural, and possessive morphemes and of forms of *be*, which is a feature of African American dialect. This suggests that use of a single set of MLU data could lead to over-identification errors for some African American children as having low MLU and, therefore, a linguistic deficit.

MLU has been adapted for languages other than English (see, for instance, Arlman-Rupp, van Niekirk-de Hahn, & van de sandt-Koenderman, 1976, for Dutch; Dromi & Berman, 1982, for Hebrew; Hickey, 1991, for Irish; Linares-Orama & Sanders, 1977, for Spanish; and Thordardottir & Weismer, 1998, for Icelandic). These adaptations involve different rules for counting morphemes and result in MLU measures that are not equivalent across languages. This means that an MLU in one language cannot be compared to that same MLU in another language. For children learning two languages, there may be interacting influences of one language on the other (Kayser, 1989; Westby, 2000). Even if English is determined to be a child's dominant language, the child may show influences of his or her first language on production of English. An example of this is Spanish-influenced English, which includes features such as verb morpheme omissions and the addition of the regular plural morpheme on irregular forms (Kayser, 1989; Paul, 2000). These influences will affect a child's MLU. MLU data from monolingual English speakers would, therefore, not be appropriate for children learning English as a second language.

Normalcy. The use of norm-referenced tests that have been standardized only on typically developing children for evaluating atypical children has been questioned (Fuchs, Fuchs, Benowitz, & Barringer, 1987). The MC and LM samples included only "normal" children; children known or suspected of having language difficulties, mental retardation, sensory deficits, and/or emotional disturbance were excluded. The resulting reference data, then, reflect the performance of only typically developing children. Consequently, by definition, all MLU values must be considered normal, and only children who score less than *any* of the children in the comparison group can be considered as non-normal. A cutoff score of 1 or even 2 *SD* below the mean for a distribution including only a "normal" comparison group could result in falsely identifying a portion of the normal population as having a language impairment (McFadden, 1996).

Reference Data and Interpretation

Both MC and LM reported means and standard deviations. These data were not derived into standard scores but represent actual MLU values, equivalent to the raw scores on standardized tests. As is true for raw scores on a standardized test, the obtained MLU is not interpretable by itself. For norm-referenced interpretation, we need to have measures of central tendency and variability (McCauley & Swisher, 1984). To use these data to make a norm-referenced interpretation, we would compare a child's MLU to the distribution for that child's age. We look to see where the obtained child's MLU falls on the distribution relative to the mean and express this in standard deviations. For example, an MLU of 1.29 for a 30-month-old would be more than 2 *SD* below the mean for that age when compared to the MC data. Alternatively, we could derive a *z* score as a way to more precisely express the child's performance relative to the group as some number of standard deviations from the mean. The *z* score can be calculated using the formula:

$$z = \frac{(\text{obtained MLU} - \text{expected MLU})}{SD}$$

$$= \frac{(1.29 - 2.54)}{.57} = -2.17$$

It is common clinical practice to use –1.5 to –2 *SD* as a diagnostic cutoff. Optimally, however, we would want to empirically establish a cutoff point on the distribution that maximally distinguishes between LI children

and NL children (Plante & Vance, 1994). Based on data from Klee, Schaffer, May, Membrino, and Mougey (1989b), the highest efficiency rate for MLU in distinguishing between LI children and NL children is achieved at –1 *SD*. At this point on the distribution, there is both 80% sensitivity and 80% specificity (see Table 10 for definitions of these terms). This is the cutoff that MC (1981) suggested as a screening cutoff for identifying children who need further language evaluation. Note that if we use this cutoff for screening, we will end up missing some of the children who would have needed further evaluation (an under-identification error) as well as evaluating some children who didn't need to be (an over-identification error). The latter error type is acceptable if we are using MLU as a screening. It is also predictable because a cutoff at –1 *SD* is within the range of typical performance and is the point along the distribution below which you expect 16% of the population to fall. However, the under-identification error means that we cannot conclude that children with an MLU above the cutoff do not have an impairment and do not need to be evaluated.

We picked a different point on the distribution at which there was at least 90% specificity, the level recommended by Plante and Vance (1994). For the Klee et al. (1989a) subjects, this was achieved at a cutoff of –1.5 *SD*. At this cutoff, the sensitivity was only 63%, so we still could not conclude that children who have an MLU higher than the cutoff have normal language. However, because the over-identification errors have been reduced, an MLU below this cutoff may serve as evidence supporting a diagnosis of language impairment.

An alternative way of interpreting MLU would be to base the interpretation on criterion referencing. In criterion-referenced interpretation, a child's performance is compared to a performance standard—the criterion—rather than to the performance distribution of a peer group. The criterion cutoff separates individuals into two groups, typically a group that meets the criterion—the mastery group—from other individuals who do not. The criterion can be established empirically by comparison to the performance of a reference group. In contrast to norm-referenced interpretation, this type of interpretation uses the raw score data. A different criterion level would be determined for each subgroup based on the actual data. One possibility is to base the criterion on range data, setting the criterion as the lowest MLU in the range, as shown in Table 4. Ranges are not available in the MC data but are provided by LM.

We wanted to determine the sensitivity and specificity for the Klee et al. (1989a) data using these cutoffs. However, we had a concern about the composition of the comparison subgroups in the LM data. LM included only children at the lower end of each age and even children below the group age. The intent seems to have been to have the mean age be at the specified age level. Children in the 3-year-old group thus ranged in age from 2;7 to 3;4, with a mean of 3;1. Children in the 4-year-old group ranged in age from 3;7 to 4;3, with a mean of 4;0. Children in the 5-year-old group ranged in age from 5;2 to 5;5, with a mean of 5;4. Because of this, it is not clear which comparison group should be used for children at some ages. LM labeled the groups by age, suggesting that they intended for speech-language pathologists to compare all 3-year-olds to the 3-year age group, all 4-year-olds to the 4-year age group, and all 5-year-olds to the 5-year age group. We felt that this might not be appropriate considering the actual composition of each age group.

TABLE 4. Criterion cutoffs for MLU by age.

Age Group	MLU Range[a]	Criterion Cutoff
3	2–5	<2
4	3–7	<3
5	4–7	<4

[a]Based on ranges for conversational samples provided by Leadholm and Miller (1992).

We thus did two separate comparisons. We first compared all of the 3- and 4-year-old subjects in the Klee et al. (1989a) study to the 3-year and 4-year age cutoffs in Table 4. For this comparison, there was 100% specificity but only 36% sensitivity. We then compared only the subjects who fell within the age ranges of the LM reference groups, that is, children between the ages of 31 to 40 months and 43 to 51 months, to the cutoffs. Specificity dropped to 88%, but sensitivity increased to 61%. These results suggest that the LM subgroups should not be defined by age level (that is, as 3-year or 4-year subgroups) but, instead, that the subgroups should be based on the actual ages of the children in the normative sample.

Reliability

Reliability is the "dependability or reproducibility of test scores or other data" (Cordes & Ingham, 1994, p. 265). There are several subtypes of reliability, each of which must be established separately. These subtypes include internal consistency, temporal stability, and examiner reliability.

Internal Consistency

Internal consistency (shown in Table 5) involves the extent to which different parts of a test converge on the same result. High correlations have been reported for sample sizes of 100 utterances for children with mild to moderate language delays (Cole, Mills, & Dale; 1989) and for younger typically developing children (Casby, 1984). A study by Darley and Moll (1960) of mean length of response (MLR)—average utterance length measured in words—suggests that reliability for samples fewer than 100 utterances may be insufficient. These authors predicted that a reliability of .90 could not be reached with fewer than 90 utterances.

Temporal Stability

Temporal stability (shown in Table 6) is determined by comparing the results from two independent administrations of the same item set separated by a brief time period and with no intervening training experience. For MLU, this involves a comparison of two language samples obtained by the same examiner under the same conditions and with the same materials. Of course, the two samples will not involve identical utterance sets and so it is not actually possible to get a pure measure of temporal stability for MLU.

Although Minifie, Darley, and Sherman (1963) reported moderate reliability for MLR with sample sizes of 50 utterances, Gavin and Giles (1996) found low reliability for MLU based on 50-utterance samples for younger children and only minimally acceptable reliability for samples of 100 to 150 utterances. A test-retest correlation of at least .90 was only achieved for samples of 175 utterances. In contrast, Cole et al. (1989) obtained a high correlation for MLU based on samples of 100 utterances from older LI children.

Examiner Reliability and Agreement

Examiner reliability and agreement are concerned with the consistency of administration and scoring. Two studies (listed in Table 7) showed low agreement between samples that were elicited by the mother at home and those that were elicited by an unfamiliar examiner in the clinic (Kramer et al., 1979; Scott & Taylor, 1978). Two other studies did not find a significant difference in MLU for a group of children between samples that were elicited by either the mother or an unfamiliar adult (Bornstein et al., 2000; Olswang & Carpenter, 1978) or for samples that were gathered at either the child's home or in the clinic (Bornstein et al., 2000). These latter two studies, however, only provide group data, so it is not possible to determine the

TABLE 5. Internal consistency reliability.

Study	Subjects	Age	MLU	Sample Size	Correlation
Cole et al. (1989)[a]	LI; *n* =10	52–80 mos	2.00–5.61	100	.94
Casby (1984)[b]	NL; *n* = 9	24–44 mos	2.90–4.37	100	.97
Darley & Moll (1960)[c]	NL; *n* =150	5;6 yrs	5.54–5.87	100	.92
				90	.90
				50	.85
				25	.74

[a]Odd/even split-half correlations.
[b]Correlation between total and partial samples of every 3rd and every 5th utterance.
[c]Estimated reliability of MLR based on analysis of variance as a function of sample size.

TABLE 6. Temporal stability reliability for MLU.

Study	Subjects	Age	Sample Size	Correlation
Minifie et al. (1963)[a]	NL	5;6 yrs	50	.82
Gavin & Giles (1996)	NL	24–44 mos	175	.90
			150	.83
			100	.78
			50	.64
Cole et al. (1989)	LI	52–80 mos	100	.92

[a]Determined MLR.

TABLE 7. Inter-examiner reliability for mean length of utterance (MLU) with different elicitation conditions.

Study	Subjects	Age	MLU	Sample Size	Agreement[a]
Kramer et al. (1979)	LI	3–5 yrs	2.5–6.42	10	20%
Scott & Taylor (1978)	NL	2;1–5;1 yrs	3.5–6.0	12	25%

[a]Calculated as the percentage of children who showed no difference in MLU between sampling conditions.

extent of agreement for MLU across the children.

Although we found no studies that investigated reliability or agreement for MLU related to scoring issues, research studies that use MLU as the basis for subject selection or language matching do report reliability or agreement. For example, MC (1981) reported inter-examiner agreement for utterance segmentation ranging from 85% to 90% and for morpheme counts ranging from 85% to 95%. A more recent study by Thal, O'Hanlon, Clemmons, and Fralin (1999) reported a mean agreement of 91% (with a range from 84% to 100% for individual samples) for utterance segmentation and a mean agreement of 88% (with a range from 82% to 93%) for word transcription for samples that were elicited from children with specific language impairment aged 39–49 months. Dollaghan et al. (1999) reported correlations of .99 for comparisons of the MLU based on two independent transcriptions, with a mean difference in MLU of .14 morphemes. However, these numbers may overestimate agreement if the individuals involved in the research project received the same training. Klee (1992) found considerably more variability in utterance segmentation (measured by total number of utterances) and total MLU for beginning graduate students. There has not been a study looking at agreement or reliability among experienced speech-language pathologists who graduated from different programs.

SEM

SEM is the estimated amount of error for an obtained score. It is calculated from the reliability coefficient *(r)* and standard deviation *(SD)* using the formula [SEM = $SD \div 1 - r$]. For norm-referenced tests, SEM provides a range of scores, based on confidence intervals, in which there is a high probability that the true score will be included. The SEM can help us to avoid an over-interpretation of performance differences, either between an individual test taker and the distribution mean for the normative sample or between two scores for a single test taker.

Meline and Meline (1981) reported an SEM of .50 for the MLU calculated for 50 children, aged 39–67 months, with an MLU ranging from 3.28 to 6.00. Although most tests report a single SEM, the magnitude of the SEM will actually be different for each set of raw scores (Hutchinson, 1996). Rondal et al. (1987) reported an increase in SEM with age and MLU for children between the ages of 1;8 and 2;8, with an MLU between 1.05 and 3.06.

We calculated SEM for the MC reference and for the LM conversational data up to age 5 using the test-retest reliability coefficients reported by Gavin and Giles (1996). These SEM values are provided in Table 8. Consistent with Rondal et al. (1987), there is an increase in the SEM with age and MLU for both sets of data because of the larger standard deviations. Noteworthy are the smaller SEM values for the LM data. This is due to the higher reliability for the larger sample size used by LM, in spite of larger standard deviations.

Confidence intervals can be calculated from the SEM by using *z* scores, representing some number of standard deviations from the mean, in the formula [$z \times SEM \pm MLU$]. Because 95% of the population falls within 2 *SD* of the mean, we would use a *z* score of 2 in order to be 95% confident that the obtained MLU was representative of the child's true MLU. We can show the importance of the SEM by calculating the confidence interval using the SEM from both the MC and LM data, as shown in Table 9. The smaller score interval for the LM data means that we can be more certain in the obtained MLU. This is solely due to the higher reliability with larger sample sizes. Note that in this particular case, this might affect the decision to further evaluate a child's language if the –1 *SD* cutoff recommended by MC (1981) were to be used.

Validity

Validity is often defined as the extent to which a test measures what it claims to measure (e.g., Haynes & Pindzola, 1998). This definition implies that validity is some inherent property that a test either does or does not have. An alternate view is that validity is less a matter of the test itself and more a matter of how test results are used (Hutchinson, 1996). In this view, what needs to be validated is not the

TABLE 8. Standard error of measurement for mean length of utterance (MLU) as a function of age.

	Miller & Chapman (1981)		Leadholm & Miller (1992)	
Age (Months)	MLU *(SD)*	SEM[a]	MLU *(SD)*	SEM[b]
18	1.18 (.32)	.19	—	—
21	1.39 (.39)	.23		
24	1.87 (.45)	.27		
27	2.40 (.51)	.31		
30	2.75 (.57)	.34	3.38 (.59)	.28
33	2.67 (.63)	.38		
36	3.66 (.69)	.41		
39	4.16 (.76)	.46		
42	3.74 (.82)	.49	4.22 (1.02)	.48
45	4.24 (.88)	.53		
48	4.33 (.94)	.56		
51	4.54 (1.00)	.60		
54	4.70 (1.06)	.64	5.71 (.91)	.43
57	5.17 (1.12)	.67		
60	5.25 (1.19)	.71		

[a]Based on a reliability of .64 for a 50-utterance sample size (Gavin & Giles, 1996).
[b]Based on a reliability of .78 for a 100-utterance sample size (Gavin & Giles, 1996).

TABLE 9. 95% confidence intervals for a 3-year-old child with a mean length of utterance of 2.18.

	MC	LM
Confidence interval	1.36–3.00	1.62–2.74
SD from mean	–.96	–1.08

test itself, but the interpretations that are made about test performance and the actions that are recommended based on those interpretations.

Whereas there are distinct types of reliability, each of which must be separately evaluated as a property of a test, this is not the case for validity. Although traditionally, several types of validity have been discussed, these are not in actuality distinct properties. Rather, these are multiple sources of evidence about validity. Messick (1989) defined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment" (p. 5). What must be considered is the extent and quality of the available evidence concerning validity (Hutchinson, 1996). The evaluation for validity begins with a consideration of its psychometric properties. We should question the validity of any assessment measure if it has an insufficiently defined purpose, inadequately described and non-replicable procedures, an insufficiently described population sample, insufficient or inappropriate reference data, or poor reliability. In addition, there are several other issues to consider when evaluating validity.

Content

What has typically been thought of as content validity relates to the operationalization and sampling of a domain. Content validity delimits the trait by which individuals will be differentiated and how that trait will be measured (Sabers, 1996). Content validity includes the two notions of content relevance and content coverage.

Content relevance. Content relevance involves specification of the test domain (Lieberman & Michael, 1986). There are definitional problems in specifying the domain of MLU that relate to the notions of *utterance* and *morpheme*. Brown (1973) did not define utterance, nor did he provide operational criteria for identifying utterances. As noted above, MC (1981) experienced 10–15% disagreement in determining utterances based primarily on terminal intonation contour. LM (1992) suggested additional guidelines (shown in Table 3), as have other authors. It is not

known to what extent use of these additional guidelines can resolve the problem of determining utterances. This is a crucial concern because the number of utterances is part of the MLU calculation.

The notion of morpheme is also not defined. Instead, Brown (1973) provided a set of decision rules for determining what should and should not be counted as a separate morpheme. For any particular child, these rules may not reflect what is actually a morpheme in that child's grammar (Arlman-Rupp et al., 1976; Crystal, 1974). Several authors have questioned the choice of morpheme as the unit for computing utterance length (Crystal, 1974; Hickey, 1991). Hickey concluded that MLU is not the best means for determining utterance length because of the degree of arbitrariness in determining morphemes and doubts about the productivity of morphemes. High correlations have been found between MLU and average utterance length measured in words (.98 to 99) and in syllables (.91 to .99) (Arlman-Rupp et al., 1976; Hickey, 1991), leading researchers to suggest that these might be better units to use for calculating utterance length. Crystal suggested the syllable as the measurement unit because of its independence from linguistic analysis. However, Hickey noted that it is harder to count syllables, and that the count could be inflated by reduplication and use of diminutives. She concluded that MLU counted in words "was found to be a measure which best balanced effectiveness and ease of application" (p. 568). Similarly, Arlman-Rupp et al. suggested that counting words is easier, faster, more reliable, and theoretically more sound because no ad hoc decisions need to be made.

Content coverage. Content coverage is concerned with the degree of representativeness with which the test samples the domain of interest (Lieberman & Michael, 1986). For MLU, we want to know whether the full range of utterances that the child is capable of has been produced, and whether the utterances have been sampled in the appropriate proportions. Retherford (1993) suggested conducting a distributional analysis of the language sample to check for representativeness. She suggested that a larger than expected number of single-word utterances in response to questions might indicate that a sample is nonrepresentative of the child's language. The difficulty here is that we do not have data on the expected utterance length distribution for children at varying MLU levels. Griffiths (1974) observed that the utterance distribution is highly skewed, with a large proportion of shorter utterances and relatively few longer utterances. She suggested that median utterance length might, therefore, be a better index of linguistic development than the mean. Davis (1937), however, observed that individual children show considerable consistency in the pattern of their utterance lengths. She reported that children with a higher mean utterance length tend to use many long sentences, whereas children with a lower mean utterance length produce few, if any, long sentences.

There are a number of rules that eliminate utterances from the MLU calculation (see Table 3). McCarthy (1930) found low levels of comprehensibility for young children (only 26% for 18-month-olds and 67% for 24-month-olds) and questioned the representativeness of utterance samples from which a large proportion of utterances was excluded. In light of this issue, it would be a good idea to set a criterion for using MLU (e.g., see the guidelines for using Developmental Sentence Scoring; Lee, 1974).

Another issue to consider in evaluating content relevance is the nature of the child's responses to test stimuli. What we want to ensure is that the child's performance reflects the domain of interest rather than the method for measuring that trait. As noted above, Stalnaker and Creaghead (1982) found MLU to be lower in a questioning during play condition than in a free play condition, possibly due to the larger proportion of sentence fragments and elliptical responses to the questions. This is an important finding because Johnston, Miller, Curtiss, and Tallal (1993) found that, when matched on MLU, LI children produced more elliptical responses to questions than did NL children, and that these elliptical responses were half the length of other sentences. Because LI children tend to be less talkative and more difficult to understand, there may be a tendency to ask them more questions in order to elicit more utterances and check understanding (Yoder, Davies, & Bishop, 1992). This may increase the number of utterances but may also result in reducing MLU.

Utility

One type of evidence for test validity is its usefulness in predicting performance now and in the future. This has been called *criterion-related validity.* This source of evidence determines an empirical relationship between the test and some criterion measure. A relationship between the test and some current measure is called *concurrent validity* or *diagnostic utility.* A relationship between the test and some future measure is called *predictive validity* or *predictive utility.* The aim here is to show that the test can be substituted as a measurement of the trait.

Several authors have investigated the relationship between MLU and other measures of language production. Scarborough, Rescorla, Tager-Flusberg, Fowler, and Sudhalter (1991) looked at the correlation between MLU and the Index of Productive Syntax (IPSyn) for both NL children and LI children. For the NL group, they found a high correlation (.93) for MLUs less than 3.0, but a much lower correlation (.58) for MLUs greater than 3.0. The correlations for the LI group showed the same trend, but were lower. MLU overestimated grammatical complexity as measured by IPSyn more frequently for the LI children than for the NL children. This means that MLU was less efficient in predicting lower IPSyn scores for the LI group. Klee and Fitzgerald (1985) looked at the relationship between MLU and syntactic development as determined by the Language Assessment, Remediation, and Screening Protocol (LARSP). The frequency of three-element clauses increased with MLU, but there was no correlation of MLU with two- and four-element clauses or with complex sentences. Although all phrase levels increased with MLU, this was not significant.

There have been no studies that have looked specifically at the predictive utility of MLU. However, there is some indication that a low MLU at one age may not predict a low MLU at a later age. A study by Scarborough et al. (1991) included 5 children with early expressive language delay who were followed longitudinally at 6-month intervals, 4 from 30 to 48 months and 1 from 36 to 48 months. All of the children scored more than 1.5 *SD* below the mean for MLU at the first sampling session. Four of these children reached an MLU within 1 *SD* of the mean by 42 months, and the remaining child did so by 48 months. Additional studies are needed with independent and validated criterion measures.

Meaningfulness

Messick (1980) suggested that construct validity be thought of as interpretative meaningfulness. This means that construct validation is "the process of marshaling evidence to support the inference that an observed consistency in test performance has a particular meaning" (p. 1015).

Relationship of MLU to age. One source of evidence of meaningfulness would be to demonstrate a relationship between MLU and age, because utterance length is thought to be measuring an aspect of development. Several studies have reported a correlation between age and MLU for NL children (deVilliers & deVilliers, 1973; Klee et al, 1989a; Miller & Chapman, 1981; Scarborough, Wyckoff, & Davidson, 1986) and for LI children (Klee et al, 1989a; Scarborough et al ., 1986).

In a longitudinal study of 6 children between the ages of 24 and 60 months, Scarborough et al. (1986) reported a linear pattern of MLU change with age only below 42 months. Above that age, the rate of change in MLU decreased. Scarborough et al. noted that the MC (1981) data also show this curvilinear pattern, but to a lesser extent. Other researchers have not found a correlation between age and MLU. In a study of 18 children between the ages of 25 and 47 months, Klee and Fitzgerald (1985) found no correlation between age and MLU. Conant (1987), however, reanalyzed the Klee and Fitzgerald data and found a correlation for the 3-year-olds but not for the 2-year-olds. Conant therefore concluded that there is a relationship between age and MLU for 3-year-olds but not for younger children.

Data from longitudinal studies suggest that the relationship between age and MLU is not a simple one. Consider the 3 children reported by Brown (1973). All 3 were initially reported at an MLU of 1.75, but Eve was 18 months old whereas Adam and Sarah were 27 months old. Eve reached Stage V (MLU greater than 4.0) by 26 months, with one slight drop in MLU at 24 months. Adam and Sarah achieved this MLU at 42–43 months, with a number of drops in MLU along the way. Klee (1992) reported a similar non-monotonic pattern for another child who was followed longitudinally, and further noted that the drop in MLU was accompanied by increases in grammatical development that were manifested by more complex sentences, more morpheme use, and phrasal elaboration.

Scarborough et al. (1991) followed 15 NL children longitudinally at 6-month intervals from 24 to 48 months. Six of the children showed drops in MLU at some point after MLU had reached 2.98, although none of these drops in MLU was sufficient to have altered the diagnostic conclusion of NL because the children all achieved an MLU no lower than 1 *SD* below the mean. Another group of 5 children with early expressive language delay was followed longitudinally from 30 to 48 months. These children showed no drops in their MLU scores. However, for one child, the MLU at 48 months dropped below –1 *SD*, although he had achieved an MLU within 1 *SD* at the prior sampling.

Group differentiation. Another source of evidence for meaningfulness would be to look at the accuracy with which children are categorized as either LI or NL based on MLU. Table 10 defines the terms for this analysis.

TABLE 10. Terms for group differentiation.

Term	Definition	Formula
Efficiency	Overall categorization accuracy for both impairment and non-impairment	(A + C) ÷ (A + B + C + D)
Sensitivity	Percentage of LI children who are correctly identified	A ÷ (A + B)
Specificity	Percentage of non-LI children who are correctly identified	C ÷ (C + D)
Positive predictive value	Percentage of children identified as LI who are LI	A ÷ (A + D)
Negative predictive value	Percentage of children identified as non-LI who are non-LI	C ÷ (B + D)

Note. A = LI children correctly identified as LI by MLU; B = LI children incorrectly identified as non-LI by MLU; C = non-LI children correctly identified as non-LI by MLU; D = non-LI children incorrectly identified as LI by MLU.

One type of measure is categorization accuracy. Klee et al. (1989b) compared a diagnostic decision based on MLU to the outcome of a diagnostic evaluation by a speech-language pathologist when a –1 *SD* cutoff based on the MC data was applied. Both sensitivity and specificity rate, and thus overall efficiency, were 83%, meaning that 40 out of 48 children were correctly classified based on MLU. These accuracy rates are less than the 90% rate suggested as adequate by Plante and Vance (1994) in their review of standardized language tests.

A second type of measure is predictive values that look at how well high or low MLU predicts language status. Predictive values are affected by the prevalence of language impairment within the general population (Dunn, Flax, Sliwinski, & Aram, 1996; Klee et al., 1989b). Assuming that MLU might be used for a general screening, such as that implemented in the schools, the 8% prevalence rate reported by Tomblin (1996) could be applied. At –1 *SD*, there would be a positive predictive value of only 28% (66 LI children out of 232 with low MLU) and a negative predictive value of 98% (764 NL children out of 778 with MLU above the cutoff). Because a screening aims for few false negatives and can tolerate some false positives, these data don't support using MLU for a screening purpose.

It is unlikely, however, that MLU would be used for such a widespread screening. The more likely scenario is that MLU would be used as a diagnostic tool for a referred population. The commonly used clinical cutoff for this aim is –1.5 to –2 *SD* below the mean. Using the MC data and applying a –1.5 *SD* cutoff yields an efficiency of 79%, with 96% specificity and 54% sensitivity. Using this cutoff, only one of the NL children was misclassified and 11 out of 24 LI children were misclassified (Klee et al., 1989b). Thus, this cutoff gives a good level of specificity for identification. Although it may seem that the low sensitivity precludes using MLU for diagnostic purposes, this may not be the case. Predictive values based on the 8% prevalence of language impairment in the general population rate yields a positive predictive value of 96% and a negative predictive value of 54%. Note that because the prevalence of language impairment in a referred population will be higher than the prevalence in the general population, the actual negative predictive value is likely to be somewhat higher than this. Thus, although we may not be able to conclude that MLUs above a certain level mean that language is normal, we may be able to use a low MLU as evidence of language impairment.

Discussion

In evaluating an assessment instrument, we have to know its purpose and whether there is evidence that it can be interpreted relative to that purpose (Plante, 1996). MLU should not be viewed as a measure of syntactic development but as one way of measuring utterance length. The evidence concerning group differentiation suggests that MLU will identify some, although not all, preschool children with language impairment. It also appears that a cutoff can be set such that we can identify the majority of children who are not impaired. We can, therefore, interpret a low MLU as supporting a diagnosis of language impairment. However, an MLU above the cutoff cannot be interpreted to mean that a child does not have an impairment.

We also need to evaluate the degree of confidence that we can have in making this interpretation (Plante, 1996). Our review of MLU suggests that we can only be moderately confident, at best, in interpreting MLU in this way. We do not have real norms for MLU, both because of the limitations of the available reference data and because it is not possible to completely standardize the collection procedures for language sampling. Because LI children

tend to be less talkative than NL children, it is not uncommon to base MLU on a small number of utterances. However, test-retest reliability is insufficient except for large sample sizes that are considerably more than the 50- to 100-utterance sample sizes that are typically collected. The low reliability results in a high SEM and, consequently, large confidence intervals. This means that only very low MLUs should be interpreted as evidence of a language impairment. It may even be the case that any child whose MLU is low enough to fall below the criterion when using confidence intervals would be identifiable as having language limitations without actually determining MLU. Thus, it may not be a worthwhile time investment to collect and transcribe a language sample just for the purpose of calculating MLU. This possibility needs to be considered because, in order to compare MLU to the available reference data, the language sample must be collected in a way that is consistent with how the reference data were collected. However, this may not be the best way to gather language samples for descriptive analysis and goal selection (see, for instance, Eisenberg, 1994, for a discussion of limitations of language sampling).

It may be the case that utterance length can be used as evidence for language impairment, but that MLU is not the best measure of utterance length for this purpose. Several authors (Arlman-Rupp et al., 1976; Crystal, 1974; Hickey, 1991) have noted the difficulty of counting morphemes and have suggested that words would be a better unit for measuring utterance length. MLU does not show a steady increase with age for all children when followed longitudinally (Brown, 1973; Klee, 1992; Scarborough et al., 1991), but shows periods in which there is a drop in MLU. Use of the mean as the measure of central tendency for utterance length may be inappropriate because utterance length is not symmetrically distributed about the mean (Griffiths, 1974). Median or modal utterance length may, therefore, be more appropriate measures. The median and mode are also less affected by the presence of nonrepresentative segments of the sample than would be the mean. Another possibility for analyzing utterance length would be to look at the length of the longer utterances above the mode. A distribution pattern of utterance lengths that cluster close to the mode versus a more dispersed distribution with longer utterances may be diagnostic of language impairment (see, for instance, discussions of utterance length distribution in Davis, 1937, and Retherford, 1993). At present, however, we do not have data on these alternative measures of utterance length.

Ultimately, valid use of any assessment tool is up to the user. Language samples must be gathered and scored in a way that is consistent with whatever reference database will be used, controlling for setting, participants, activity, type of material, and sample size, and following the specific guidelines for determining utterances and morphemes. Adequate time must be allotted to collect a sufficient number of usable utterances. MLU should not be used for sample sizes that are smaller than the minimum for those reference data, preferably at least 100 utterances and never fewer than 50. MLU should also not be used if more than half of a child's utterances must be excluded. Low MLU may be used as one piece of evidence supporting a diagnosis of language impairment in preschool children, but should never be used alone for this purpose (Lahey, 1988; Leadholm & Miller, 1992; Miller, 1981; Nelson, 1998; Owens, 1999).

Addendum

In the February 2001 issue of *JSLHR*, Judith Johnston discussed an alternative calculation of mean utterance length that excluded certain discourse-sensitive utterances. These included self-repetitions, imitations, single-word yes/no responses, and elliptical responses to *wh*-questions. Excluding these utterances resulted in eliminating an average of 60 utterances from each child's sample (approximately 20% of the total number of utterances). As a result, MLU increased for both LI and NL children by anywhere from 3% to 49%. The magnitude of the effect related to the relative proportion of questions asked by the adult participant and, for the LI children, also to language level and proficiency. This study underscores the importance of following the scoring conventions that are used for developing the reference data set. In addition, this study raises issues about the validity of MLU, specifically concerning what is actually being measured given the influence on this index of discourse variables.

Author Note

Portions of this paper were compiled while the authors were at Columbia University Teachers College in New York. We would like to thank Thomas Klee and Hollis Scarborough for sharing their MLU data with us. We would also like to thank the anonymous reviewers for their comments and suggestions on earlier versions of this paper.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985).

Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Arlman-Rupp, A. J. L., van Niekirk-de Hahn, D., & van de sandt-Koenderman, M. (1976). Brown's early stages: Some evidence from Dutch. *Journal of Child Language, 3*, 267–274.

Bain, B. A., Olswang, L. B., & Johnson, G. A. (1992). Language sampling for repeated measures with language-impaired preschoolers: Comparison of two procedures. *Topics in Language Disorders*, 12, 12–27.

Bernstein, D. K., & Tiegerman-Farber, E. (1997). *Language and communication disorders in children* (4th ed.). Boston, MA: Allyn & Bacon.

Bloom, L., & Lahey, M. (1978). *Language development and language disorders*. New York: Wiley.

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

Bountress, M. G., Bountress, N. G., & Tonelson, S. W. (1988). The influence of racial experimenter effects upon mean length of utterance. *Clinical Linguistics and Phonetics, 2*, 47–53.

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

Casby, M. W. (1984). Application of intermittent time-sampling to calculations of mean length of utterance. *Perceptual & Motor Skills*, 58, 715–718.

Cole, K. N., Mills, P. E., & Dale, P. S. (1989). Examination of test-retest and split-half reliability for measures derived from language samples of young handicapped children. *Language, Speech, and Hearing Services in Schools, 20*, 259–268.

Conant, S. (1987). The relationship between age and MLU in young children: A second look at Klee and Fitzgerald's data. *Journal of Child Language, 14*, 169–173.

Cordes, A. K., & Ingham, R. J. (1994). The reliability of observational data: Issues in the identification and measurement of stuttering events. *Journal of Speech and Hearing Research, 37*, 279–294.

Crystal, D. (1974). Review of R. Brown, *A First Language: The Early Stages. Journal of Child Language, 1*, 289–307.

Darley, F. L., & Moll, K. L. (1960). Reliability of language measures and size of language samples. *Journal of Speech and Hearing Research, 3*, 166–173.

Davis, E. A. (1937). Mean sentence length compared with long and short sentences as a reliable measure of language development. *Child Development, 8*, 69–79.

deVilliers, J. G., & deVilliers, P. A. (1973). Development of the use of word order in comprehension. *Journal of Psycholinguistic Research, 2*, 331–341.

Dollaghan, C. A., Campbell, T. F., Paradise, J. L., Feldman, H. M., Janosky, J. E., Pitcairn, D. N., & Kurs-Lasky, M. (1999). Maternal education and measures of early speech and language. *Journal of Speech, Language, and Hearing Research, 42*, 1432–1443.

Dromi, E., & Berman, R. A. (1982). A morphemic measure of early language development: Data from modern Hebrew. *Journal of Child Language, 9*, 403–424.

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. (1994). Investigating children's language: A comparison of conversational sampling and elicited production. *Journal of Psycholinguistic Research, 26*, 519–538.

Fey, M. E. (1986). *Language intervention with young children*. Needham Heights, MA: Allyn & Bacon.

Fuchs, D., Fuchs, L., Benowitz, S., & Barringer, K. (1987). Norm-referenced tests: Are they valid for use with handicapped students? *Exceptional Children, 54*, 263–271.

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

Gray, S., Plante, E., Vance, R., & Hendrichsen, M. (1999). The diagnostic accuracy of four vocabulary tests administered to preschool-age children. *Language, Speech, and Hearing Services in Schools, 30*, 196–206.

Griffiths, P. (1974). Review of M. Bowerman, *Early Syntactic Development. Journal of Child Language, 1*, 123.

Haynes, W., Purcell, E., & Haynes, M. (1979). A pragmatic aspect of language sampling. *Language, Speech, and Hearing Services in Schools, 10*, 104–110.

Haynes, W. O., & Pindzola, R. H. (1998). *Diagnosis and evaluation in speech pathology*. (5th ed.). Boston, MA: Allyn & Bacon.

Hickey, T. (1991). Mean length of utterance and the acquisition of Irish. *Journal of Child Language, 3*, 553–569.

Hutchinson, T. A. (1996). What to look for in the technical manual: Twenty questions for users. *Language, Speech, and Hearing Services in Schools, 27*, 109–121.

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

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

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

Jones, M., Weismer, S. E., & Schumacher, K. (2000, June). *Grammatical morphology in school-age children with and without language impairment: Discriminant function analysis.* Poster presented at the Symposium on Research in Child Language Disorders, University of Wisconsin-Madison.

Kayser, H. (1989). Speech and language assessment of Spanish-English speaking children. *Language, Speech, and Hearing Services in Schools*, 20, 226–244.

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

Klee, T. (1992). Measuring children's conversational language. In S. F. Warren & J. Reichle (Eds.), *Causes and effects in communication and language intervention* (pp. 315–330). Baltimore, MD: Brookes.

Klee, T., & Fitzgerald, M. D. (1985). The relation between grammatical development and mean length of utterance in morphemes. *Journal of Child Language*, *12*, 251–269.

Klee, T., Schaffer, M., May, S., Membrino, I., & Mougey, K. (1989a). *Predictive value of MLU in normal and language impaired preschool children.* Unpublished manuscript.

Klee, T., Schaffer, M., May, S., Membrino, I., & Mougey, K. (1989b). A comparison of the age-MLU relation in normal and specifically language-impaired preschool children. *Journal of Speech and Hearing Disorders*, *54*, 226–233.

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.

Lahey, M. (1988). *Language disorders and language development.* New York: MacMillan.

Lahey, M., Liebergott, J., Chesnick, M., Menyuk, P., & Adams, J. (1992). Variability in children's use of grammatical morphemes. *Applied Psycholinguistics*, *13*, 373–398.

Leadholm, B. J., & Miller, J. F. (1992). *Language Sample Analysis: The Wisconsin guide.* Madison, WI: Wisconsin Department of Public Instruction.

Lee, L. L. (1974). *Developmental Sentence Analysis.* Evanston, IL: Northwestern University Press.

Lieberman, R. J., & Michael, A. (1986). Content relevance and content coverage in tests of grammatical ability. *Journal of Speech and Hearing Disorders*, *51*, 71–81.

Linares-Orama, N., & Sanders, L. J. (1977). Evaluation of syntax in three-year-old Spanish-speaking Puerto Rican children. *Journal of Speech, Language, and Hearing Research*, *20*, 350–357.

Loeb, D. F., Kinsler, K., & Bookbinder, L. (2000, November). *Current language sampling practices in preschools.* Poster presented at the Annual Convention of the American Speech-Language-Hearing Association, Washington, D.C.

Long, S. (1991). Integrating microcomputer applications into speech and language assessment. *Topics in Language Disorder*, *11*, 1–17.

Lund, N. J., & Duchan, J. F. (1993). *Assessing children's language in naturalistic contexts* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

McCarthy, D. A. (1930). *The language development of the preschool child* (Institute of Child Welfare Monograph No. 4). Minneapolis, MN: University of Minnesota Press.

McCauley, R. J. (1996). Familiar strangers: Criterion-referenced measures in communication disorders. *Language, Speech, and Hearing Services in Schools*, *27*, 122–131.

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

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

Meline, T. J., & Meline, N. C. (1981). Normal variation and prediction of mean length of utterance from chronological age. *Perceptual and Motor Skills*, *53*, 376–378.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*, 1012–1027.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*, 5–11.

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

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.

Miller, J. F., & Chapman, R. S. (1984–2000). *Systematic Analysis of Language Transcripts* (*SALT*). Madison, WI: University of Wisconsin-Madison Waisman Center, Language Analysis Laboratory.

Miller, J. F., Frieberg, C., Rolland, M. B., & Reves, M. A. (1992). Implementing computerized language sample analysis in the public school. *Topics in Language Disorder*, *12*, 69–82.

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

Nelson, N. W. (1998). *Childhood language disorders in context: Infancy through adolescence* (2nd ed.). Boston, MA: Allyn & Bacon.

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.

Owens, R. E. (1999). *Language disorders: A functional approach to assessment and intervention* (3rd ed.). Boston, MA: Allyn & Bacon.

Paul, R. (2000). *Language disorders from infancy through adolescents* (2nd ed.). Saint Louis, MO: Mosby-Year Book.

Plante, E. (1996). Observing and interpreting behaviors: An introduction to the clinical forum. *Language, Speech, and Hearing Services in Schools*, *27*, 99–101.

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

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

Rice, K. J., Rice, M. L., & Redmond, S. M. (2000, June). *MLU outcomes for children with and without SLI: Support for MLU as a matching variable.* Poster session at the Symposium on Research in Child Language Disorders, University of Wisconsin-Madison.

Rollins, P. R. (1995, November). *MLU as a matching variable: Understanding its limitations.* Poster presented at the Annual Convention of the American Speech-Language-Hearing Association, Orlando, FL.

Rollins, P. R., Snow, C. E., & Willett, J. B. (1996). Predictors of MLU: Semantic and morphological developments. *First Language*, 16, 243–259.

Rondal, J. A., Ghiotto, M., Bredart, S., & Bachelet, J. (1987). Age-relation, reliability, and grammatical validity of measures of utterance length. *Journal of Child Language*, *14*, 433–446.

Sabers, D. L. (1996). By their tests we will know them. *Language, Speech, and Hearing Services in Schools*, *27*, 102–109.

Scarborough, H. S., Rescorla, L., Tager-Flusberg, H., Fowler, A. E., & Sudhalter, V. (1991). The relation of utterance length to grammatical complexity in normal and language-disordered groups. *Applied Psycholinguistics*, *12*, 23–45

Scarborough, H. S., 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.

Scott, C. M., & Taylor, A. E. (1978). A comparison of home and clinic gathered language samples. *Journal of Speech and Hearing Disorders*, *43*, 482–495.

Stalnaker, L. D., & Creaghead, N. A. (1982). An examination of language samples obtained under three experimental conditions. *Language, Speech, and Hearing Services in Schools*, *13*, 121–128.

Thal, D. J., O'Hanlon, L., Clemmons, M., & Fralin, L. (1999). Validity of a parent report measure of vocabulary and syntax for preschool children with language impairment. *Journal of Speech, Language, and Hearing Research*, *42*, 482–498.

Thordardottir, E. T., & Weismer, S. E. (1998). Mean length of utterance and other language sample measures in early Icelandic. *First Language*, *18*, 1–32.

Tomblin, B. (1996, June). *The big picture of SLI: Results of an epidemiological study of SLI among kindergarten children.* Paper presented at the Symposium on Research in Child Language Disorders, University of Wisconsin-Madison.

Vaughn-Cooke, F. B. (1986). The challenge of assessing the language of non-mainstream speakers. In O. L. Taylor (Ed.), *Treatment of communication disorders in linguistically diverse populations* (pp. 23–48). San Diego, CA: College-Hill.

Wagner, C. R., Nettelbladt, U., Sahlen, B., & Nilholm, C. (2000). Conversation versus narration in pre-school children with language impairment. *International Journal of Language and Communication Disorders*, *35*, 83–93.

Westby, C. E. (2000). Multicultural issues in speech and language assessment. In J. B. Tomblin, H. L. Morris, & D. C. Spriestersbach (Eds.), *Diagnosis in speech-language pathology* (pp. 35–62). San Diego, CA: Singular.

Yoder, P. J., Davies, B., & Bishop, K. (1992). In S. F. Warren & J. Reichle (Eds.), *Causes and effects in communication and language intervention* (pp. 255–275). Baltimore, MD: Brookes.

Received August 18, 2000
Accepted March 19, 2001
DOI: 10.1044/1058-0360(2001/028)

Contact author: Sarita Eisenberg, PhD, Department of Communication Sciences and Disorders, Montclair State University, 1 Normal Avenue, Upper Montclair, NJ 07043.
E-mail: eisenbergs@mail.montclair.edu

Key Words: language assessment, mean length of utterance, language impairment, preschool language, psychometrics