Utility of the MacArthur–Bates Communicative Development Inventory in Identifying Language Abilities of Late-Talking and Typically Developing Toddlers

John Heilmann
Susan Ellis Weismer
Julia Evans
Christine Hollar
University of Wisconsin–Madison

The present study investigated the validity of the MacArthur–Bates Communicative Development Inventory (CDI) for a group of toddlers 30 months of age. Study 1 examined the concurrent validity of the CDI for a group of 38 late talkers. Significant correlations were found between the CDI and direct measures of language abilities. Study 2 used likelihood ratio analysis to determine how well the CDI sorted 100 toddlers (38 late talkers and 62 children with a history of normal language development) according to language status based on direct assessment measures. The analyses showed that the CDI was effective in identifying children with low language skills up to the 11th percentile and in identifying children with normal language skills above the 49th percentile.

Key Words: language assessment, late talkers, Communicative Development Inventory, concurrent validity, language expression

The MacArthur–Bates Communicative Development Inventory (CDI) long form has been widely used for both clinical and research purposes. Various studies have found that this parent report measure is effective in characterizing children's early language skills (Dale, 1991; Dale, Bates, Reznick, & Morisset, 1989; Miller, Sedey, & Miolo, 1995; Thal, O'Hanlon, Clemmons, & Fralin, 1999). It has been suggested that the CDI is a useful tool for clinicians to assess early language skills, as it is not affected by some of the performance difficulties potentially encountered in this population. Such difficulties include toddlers' low rate of communicative behavior resulting in insufficient samples and noncompliance with adults' requests (Paul, 2001).

The CDI has been used extensively in studies of early language skills of toddlers. Several studies have used the CDI as the primary dependant variable to investigate early language development, which has led to the development of models for both normal language processes (e.g., Bates et al., 1994; Bauer, Goldfield, & Reznick, 2002; Marchman & Bates, 1994) and language processes of populations with special needs (e.g., Caselli et al., 1998; Harris, Bellugi, Bates, Jones, & Rossen, 1997; Hick, Joseph, Conti Ramsden, Serratrice, & Faragher, 2002). The CDI has also been used to assess children's language levels in both typical populations (e.g., Dale, Dionne, Eley, & Plomin, 2000; Farrar & Maag, 2002; Feldman et al., 2003) and clinical populations (e.g., Chilosi, Cipriani, Bertuccelli, Pfanner, & Cioni, 2001; Lyytinen, Eklund, & Lyytinen, 2003; Yoshinaga-Itano, Snyder, & Day, 1998).

Numerous studies have used the CDI to identify and study toddlers who are significantly behind their peers in language development at approximately 2 years of age. These children, referred to as late talkers (LTs), have been the focus of studies investigating specific theoretical claims about normal language development (Thal, Bates, Goodman, & Jahn-Samilo, 1997), studies of toddlers at risk for later specific language impairments (Ellis Weismer & Evans, 2002), and studies examining treatment effects in LTs (Girolametto, Pearce, & Weitzman, 1996; Girolametto, Wiigs, Smyth, Weitzman, & Pearce, 2001; Robertson & Ellis Weismer, 1999). Late talkers are generally identified around 24 months, though there is some variability in specific ages of participants in the

1058-0360/05/1401-0040

literature (the range of ages for LTs in the studies in this review was 18 to 30 months). The most common criterion for classifying children as LTs has been total productive vocabulary at or below the 10th percentile on the Words & Sentences form of the CDI (CDI–WS; e.g., Ellis Weismer & Evans, 2002; Gershkoff-Stowe, Thal, Smith, & Namy, 1997; Robertson & Ellis Weismer, 1999). For example, Ellis Weismer and Evans (2002) used this criterion to identify their LT group in an investigation of novel word learning in LTs. In an intervention study, Robertson and Ellis Weismer (1999) employed the same criterion to identify a group of LTs in an investigation of changes in linguistic and social skills in response to an interactive language treatment program. In other intervention studies examining both the short-term and long-term effects of treatment for LTs, a more stringent criterion of less than or equal to the 5th percentile on the total productive vocabulary section of the CDI–WS has been used (e.g., Girolametto et al., 1996, 2001).

The measurement properties of the CDI have been examined across a number of studies. One way the validity of the CDI has been evaluated is by establishing the concurrent validity of this measure with respect to the degree of correspondence between reported language abilities and direct assessment of language functioning. Moderate to strong correlations have been found between the CDI–WS and direct language measures for typical populations, children with developmental disabilities, and children with specific language delay (see Appendix A for a list of these studies and a summary of the reported correlations). Dale et al. (1989) completed the first published study examining the concurrent validity of the Early Language Inventory (ELI), an early version of the CDI. Both the ELI and a direct measure of language skills, adapted from specific items from the Bayley Scales of Infant Development, were administered to a group of 2-year-old children. They examined several groups, including typically developing, high social risk, preterm, and precocious children. While the different groups had associated factors that might influence language development, the investigators did not state that any of the children had explicit language disorders. Significant correlations were found between the ELI and the Expressive Language score derived from the Bayley Scales across all children in the study. In a subsequent study, Dale (1991) conducted a more thorough examination of the concurrent validity of the CDI–WS for typically developing 2-year-olds. The CDI–WS measures of total productive vocabulary, three longest utterances, and sentence complexity were significantly correlated (ranging from .47 to .79) with direct measures of vocabulary and syntax obtained from language sample analysis and standardized tests.

The validity of the CDI–WS has also been examined for children with language delay. Miller et al. (1995) assessed the validity of the CDI for children with developmental disabilities. Participants included 44 children with Down syndrome and 46 typically developing children, matched for mental age. Results from the total productive vocabulary section of the CDI–WS were compared to a direct measure of vocabulary. Significant correlations were found across all measures, and no significant difference was found in the strength of correlations between children with Down syndrome and typically developing children. The concurrent validity of the CDI–WS has also been examined for children with specific language delay who were above the age range for which the CDI was normed. Thal et al. (1999) examined children with language delay who were between the ages of 39 and 49 months; all children were chronologically above the 30-month ceiling for which the CDI–WS form was normed. Thal and colleagues found significant correlations across reported measures of vocabulary and syntax from the CDI and direct measures of vocabulary and syntax (ranging from .52 to .86 for the toddler form). They concluded that the CDI is effective in assessing the language skills of children above the age limit for which the CDI was intended.

While the validity of the CDI has been documented across various studies, the measurement properties of the CDI have been criticized, and its usefulness for clinical purposes such as identifying language delay has been questioned (Feldman et al., 2000). Feldman and colleagues administered the CDI to a large sample (N = 2,156) of infants and toddlers that was considerably more ethnically and socioeconomically diverse than the sample used to norm the CDI. Specifically, they administered the Words & Gestures form of the CDI (CDI–WG) to toddlers 10–13 months of age and the CDI–WS to toddlers 22–25 months of age. Findings indicated that the majority of the scales on the CDI were developmentally sensitive, with scores increasing monotonically across these age ranges. However, there was considerable variability in performance, with standard deviations approximating or exceeding mean values on both vocabulary scales of the CDI–WG and three grammatical scales of the CDI–WS (word forms-irregulars, word forms-overregularized, and sentence complexity). Significant sociodemographic effects were found, but the directionality of differences in scores was not consistent across the two CDI inventories. Finally, only modest correlations were found between the CDI–WG at 1 year and the CDI–WS at 2 years. Based on these results, Feldman and colleagues suggested caution in using the CDI to identify language delay in individual children, to compare performance for children with differing sociodemographic backgrounds, or to assess treatment effects.

Fenson et al. (2000) responded to Feldman et al.'s (2000) cautions regarding the use of the CDI, claiming that their interpretations were overly pessimistic. In particular, Fenson et al. argued that the extent of variability reflects the very nature of early language development, rather than being a reflection of psychometric weakness on the part of the CDI. They acknowledged that 1 year of age is likely to be too young to identify individual children at risk for language delay, but they asserted that the low predictive power of the CDI at 12 months does not mean that it has limited utility as a screening tool at later developmental levels. Given the marked increase in stability of language abilities in children 12–24 months old, the predictive power of the CDI to identify children at risk for language delay should increase with age. As pointed out by Fenson and colleagues, the same trends regarding variability and

low predictive power at very early stages of development are observed with direct measures; thus, these assessment issues are not specific to the CDI.

Considering the change in the stability of language skills in the early stages of language development, we were interested in investigating the ability of the CDI–WS to accurately assess the language skills of LTs at 30 months of age. As noted earlier, one way to demonstrate the validity of the CDI is to establish the concurrent validity of this measure with respect to the degree of correspondence between reported language abilities and direct assessment of language functioning. Another way to assess the integrity of a test is to determine its ability to sort children as having either low language abilities (language delay) or normal language abilities. One method for evaluating this type of distinction is through the use of likelihood ratio analysis, which compares results of a given test to a gold standard (Sackett, 1991). In the case of the CDI, one could use likelihood ratios to compare results from the CDI parent report to a gold standard acquired from direct measures of language ability, such as scores from standardized tests and language sample analysis. This technique is often used in clinical medicine but has only been used in a handful of studies within the area of language disorders (Dollaghan & Campbell, 1998; Ellis Weismer et al., 2000).

The purpose of the present investigation, therefore, was to determine the validity of the CDI–WS in characterizing language skills of 30-month-old toddlers who were initially identified as LTs at 24 months. Two studies were conducted. In Study 1, concurrent validity of the CDI was evaluated by examining performance on this parent report measure compared to direct assessments of language skills derived from a standardized test and spontaneous language samples for toddlers at 30 months of age. In Study 2, likelihood ratio analysis was used to assess the accuracy of the CDI in classifying typically developing and late-talking toddlers into language ability groups.

Study 1

Participants

Participants in these studies were part of a larger longitudinal project investigating specific language delay. The sample of LTs and typical talkers was recruited via a birth registry maintained by the Research Participation Core at the Waisman Center at the University of Wisconsin–Madison, flyers posted throughout the community, advertisements in local newsletters, posters at health fairs, and referrals from Birth to Three providers. Study 1 examined 38 toddlers identified as LTs who were initially classified as LTs by scoring at or below the 10th percentile for total productive vocabulary on the CDI–WS at 24 months of age. Gender-based norms were used in establishing the 10th percentile cutoff, allowing for developmental differences in boys and girls. The mean number of words reported for the LTs on the CDI at 24 months was 46.11 (*SD* = 30.13), with a mean percentile of 3.47 (*SD* = 3.82). Boys produced a mean of 43.08 words (*SD* = 21.14) and had a mean percentile rank of 3.65 (*SD* = 3.78), whereas girls produced a mean of 52.67 words (*SD* = 44.37), with a mean percentile of 3.08 (*SD* = 4.03).

All participants in the study met specific inclusionary criteria. Each of the participants' parents initially completed a background questionnaire that provided information regarding developmental milestones, medical history, identified or suspected areas of difficulty (e.g., hearing impairment, cognitive disability, motor impairment, problems with social interaction), family characteristics, and the language spoken in the home. Based on this questionnaire, all children were reported to be typically developing at 24 months in all areas other than language and were from a monolingual English-speaking home. Additional inclusionary criteria were based on direct assessments of the children's performance at 30 months. All participants were required to (a) score within the normal range on the Denver II (Frankenburg et al., 1992), a general developmental measure; (b) exhibit normal hearing as screened by distortion product otoacoustic emissions using a Biologic OAE screener (2000, 3000, 4000, and 5000 Hz in at least one ear); and (c) demonstrate normal oral and speech motor abilities as evaluated by a pediatric clinical assessment tool developed by Robbins and Klee (1987).

Twelve (32%) of the LTs were girls, and 26 (68%) were boys. An uneven distribution of gender is common in this population, as late talking is more prevalent in boys. Gender ratios of boys to girls in previous studies range from 3:1 or 4:1 (Ellis Weismer, Murray-Branch, & Miller, 1994; Paul & Smith, 1993; Thal, Tobias, & Morrison, 1991) to 19:1 (Rescorla & Goossens, 1992). Maternal education (number of years of schooling) was used as a measure of socioeconomic status. The mothers of the LTs in this study had a mean of 15.50 years of education (*SD* = 2.20). The sample primarily consisted of White children from middle-class backgrounds. Thirty-six of the 38 participants were White, 1 was African American, and 1 was biracial (African American-White). Five of the LTs were receiving speech-language intervention at 30 months according to parental report.

Procedure

All parents received and completed the CDI–WS when their child was 24 months of age. Parents completed the inventory at home around their child's second birthday (*M* = 23.84 months, *SD* = 1.41). Direct assessment tasks were completed in the laboratory at 30 months (*M* = 29.63, *SD* = 0.54); parents accompanied their children to the Waisman Center for two 1-hr sessions. Each session was completed in a quiet room, with both the examiner and parent in the room. The data were recorded via audio- and videotaping, and were later reviewed for scoring. The majority of the parents who attended the sessions were mothers. At the time of the evaluation, parents completed a second CDI–WS, while their children completed the assessment tasks. There were only a few occasions where the parent was unable to complete the inventory over the two sessions. In these cases, the parents mailed the CDI to the examiners shortly after their last visit.

Direct measures of language skills were administered by one of four ASHA-certified examiners. The children received several speech, language, and cognitive measures during the two sessions. The first session was composed of the following: hearing screening, Bayley Scales of Infant Development—Second Edition (Bayley, 1993), Arizona Articulation Proficiency Scale—Third Revision (Fudala, 2000), oral motor exam, and parent–child language sample. The second session included administration of the Denver II (Frankenburg et al., 1992), Preschool Language Scale—Third Edition (PLS–3; Zimmerman, Steiner, & Pond, 1992), and the examiner–child language sample.

The concurrent validity of the CDI was examined by comparing the CDI completed at 30 months with the direct language measures obtained at 30 months. For the purposes of this study, three subsections of the CDI were examined: total productive vocabulary (in which the parent is presented with a large word list from which he or she identifies the words the child produces), mean of three longest utterances (index of utterance length/syntax), and sentence complexity (measure of morphosyntax). Each subsection of the CDI was compared with each of the direct measures. Direct language assessment measures included the Expressive portion of the PLS–3 and measures derived from the examiner–child and parent–child language samples. Trained graduate students transcribed and analyzed the samples using Systematic Analysis of Language Transcripts (SALT; Miller & Chapman, 2002). Utterances were segmented using T-units. The first 10 min of the parent–child language samples and the first 12 min of the examiner–child language samples were analyzed. The additional 2 min for the examiner–child interactions were included to allow for the extra time needed for the children to become comfortable with the examiner. SALT provides analysis of mean length of utterance (MLU) in morphemes and other summary analyses for either complete and intelligible utterances or total utterances. Because the toddlers were at an early stage of language development, many of the utterances were not complete and intelligible. To achieve an acceptable amount of data for the analyses, all utterances were used in calculating the language sample measures. The MLU for all utterances and mean of the three longest intelligible utterances (M3L) were calculated. Number of different words (NDW) was also calculated from the first 50 utterances in each sample. One toddler did not produce a total of 50 utterances in the parent–child language sample, and 1 other child did not produce a total of 50 utterances in the examiner–child language sample.

Richards and colleagues (Richards & Malvern, 1997) have expressed concern regarding the use of NDW as a measure of lexical diversity, arguing that it is affected by differences in length across language samples. Specifically, they argued that as children's utterances increase in length, they produce a larger number of total words (NTW). When comparing language samples matched on the number of utterances, children with higher MLU values will have higher NTW values, resulting in biased NDW values. To address this issue, the differences in length across language samples were controlled by holding NTW constant. Partial correlation coefficients were calculated between NDW and the three CDI measures, with NTW as the covariate. By removing the covariance due to the length of the child's language sample (i.e., NTW), an estimate of lexical diversity that is not affected by sample size was achieved.

Agreement

Point-to-point agreement was calculated for 13% ($n = 5$) of the participants in Study 1. All language samples were transcribed by two independent judges. Interrater agreement was calculated for both the adult and child utterances, with parent–child language samples showing the following agreement: morpheme-by-morpheme: 4,603/5,053 (91.1%), utterance segmentation: 1,541/1,587 (97.1%). Interrater agreement for the examiner–child language samples was as follows: morpheme-by-morpheme: 5,370/5,669 (94.7%), utterance segmentation: 1,674/1,725 (97.0%).

Results

Descriptive statistics. Descriptive statistics for measures from the CDI at 30 months are summarized in Table 1, and descriptive statistics for the direct measures of language status at 30 months are presented in Table 2. While all of the LTs scored below the 10th percentile on the total productive vocabulary section of the CDI at 24 months, the average performance on the total productive vocabulary section of the CDI increased to the 15th percentile at 30 months. Several of the participants demonstrated notable increases in their language skills between 24 and 30 months. These late bloomers accounted for the overall increase in the total vocabulary percentile rank. The measures obtained from the examiner–child and parent–child 30-month language samples were similar.

Correlations of direct measures and the CDI. A one-tailed Pearson correlation coefficient was computed for each measure from the CDI and each direct language measure. Partial correlation coefficients were calculated for NDW and each of the CDI measures, controlling for NTW. Correlations between the CDI and direct measures are presented in Table 3. Due to the large number of comparisons, the familywise Type I error rate was controlled using the false detection rate method (Benjamini & Hochberg, 1995). This method accounted for the multiple comparisons while controlling for the most relevant Type I errors,

TABLE 1. Descriptive data for Study 1 for the MacArthur–Bates Communicative Development Inventory (CDI) at 30 months.

	M	*SD*
Total productive vocabulary	264.50	142.62
Vocabulary percentile	15.00	14.21
M3L	3.81	2.08
Complexity score	5.77	6.13
Complexity percentile	14.42	8.37

Note. M3L = mean length of the three longest intelligible utterances.

TABLE 2. Descriptive data for Study 1 for direct assessment measures at 30 months.

	M	*SD*
PLS–3: Expressive	92.15	13.00
MLU		
Examiner–child	1.67	0.44
Parent–child	1.52	0.46
M3L		
Examiner–child	3.97	0.99
Parent–child	3.78	1.46
NDW		
Examiner–child	34.42	10.57
Parent–child	33.25	10.87

Note. PLS–3: Expressive = Preschool Language Scale—Third Edition, Expressive Communication score; MLU = mean length of utterance in morphemes based on total utterances; NDW = number of different words based on first 50 utterances.

resulting in a reduction in alpha from .05 to .01. A significant correlation was found between each section of the CDI and all but one of the direct measures. The correlation between the CDI:M3L and M3L for the examiner–child language sample was not significant ($p = .05$). The significant correlations were moderate in strength, ranging from .38 to .67. The three strongest correlations were the following: CDI complexity/examiner–child MLU (.67), CDI total productive vocabulary/PLS–3: Expressive (.63), and CDI:M3L/PLS–3: Expressive (.60). The weakest correlations were from the measure of M3L collected from the examiner–child language sample and the total productive vocabulary and M3L sections of the CDI.

Discussion

Results of Study 1 suggest that the CDI–WS is a valid tool to assess the language skills of 30-month-old LTs. Overall, these results are consistent with previous studies examining the concurrent validity of the CDI in typically developing children (Dale, 1991; Dale et al., 1989) and children with language delay (Miller et al., 1995; Thal et al., 1999). However, the correlations observed in this study were not as strong as those reported in previous studies. All prior studies, with the exception of Dale et al. (1989), had correlations that spanned through the .70s and .80s. A restricted distribution of language skills could account for the weaker correlations in the present study. Because all of the children in Study 1 were LTs and all children were within a narrow age range (28–32 months) at the time of testing, the language scores for children in the present study may not have been as widely distributed as those in previous studies. Such a restricted distribution makes it more difficult for the participants' scores to line up on compared measures, resulting in correlations with decreased strength.

TABLE 3. Correlations between the CDI and the direct assessment measures.

	Total productive vocabulary	M3L	Complexity
MLU			
Examiner–child	.46*	.44*	.67*
Parent–child	.58*	.56*	.52*
M3L			
Examiner–child	.38*	.34	.56*
Parent–child	.57*	.57*	.52*
NDW[a]			
Examiner–child	.58*	.53*	.57*
Parent–child	.51*	.40*	.43*
PLS–3: Expressive	.63*	.60*	.47*

[a]Partial correlation for number of different words based on 100-word sample, controlling for number of total words.
*Significant at $p = .01$.

Study 2

Participants

Study 2 consisted of 100 toddlers who were part of the same longitudinal study examining specific language delay. The 38 participants from Study 1 were included in Study 2. In addition, 62 children who were identified as having normal language at 24 months were included in Study 2. Specifically, these children scored above the 10th percentile on the total productive vocabulary section of the CDI–WS at 24 months (mean total productive vocabulary at 24 months = 328.4, $SD = 165.4$, for the normal language group compared to $M = 46.1$, $SD = 30.1$, for the LT group). All participants met the same inclusionary/exclusionary criteria as described in Study 1. Again, maternal education was used as a measure of socioeconomic status, with a mean of 15.85 years ($SD = 2.09$). Children in Study 2 also primarily consisted of White children from middle-class backgrounds. Ninety-three of the toddlers were White, 2 were African American, 1 was of Asian decent, and 4 were biracial (African American-White).

Procedure

The same protocol described in Study 1 was used in Study 2. In order to determine the CDI's ability to classify children, likelihood ratios were computed for several percentile cutoffs for the total productive vocabulary section of the CDI. Sackett (1991) described the calculation of likelihood ratios for a positive test result (LR+) as the proportion of true positives and false positives (true positives/false positives, or sensitivity/[1 – specificity]). A higher LR+ indicates that test results were more likely to come from children who exhibit language delay than a child with typical language development, as defined by a gold standard. For example, if a diagnostic cutoff (e.g., scoring at or below the 15th percentile on a language test) produces an LR+ of 20, then children performing at that level are 20 times more likely to have a true language delay than no language delay.

Likelihood ratios can also be calculated for a negative test result (LR–), which is the proportion of false negatives and true negatives (false negatives/true negatives, or [1–sensitivity]/specificity). The lower the LR– is for a

negative test, the greater the likelihood that the test result came from typically developing children. Therefore, an LR– close to zero is informative. For example, if a diagnostic cutoff (e.g., scoring at or above the 40th percentile on a language test) produces a likelihood ratio of 0.04, then children performing at that level are less than 1/20 times as likely to have a true language delay than no language delay. One of the advantages of likelihood ratios is that they can be calculated for several levels of a test result (Sackett, 1991). Likelihood ratios in intermediate ranges are calculated in the same way as likelihood ratios for positive test results (true positives/false positives, or sensitivity/[1 – specificity]). See Appendixes B and C for a review of calculating likelihood ratios.

Clinically, calculation of likelihood ratios can aid in better understanding the measurement properties of an assessment tool, leading to more informed use of the tool. By completing a series of LR+ for each level of the assessment (e.g., percentiles or standard scores), the clinician can determine how likely it is that a client has a true impairment based on her or his performance on the assessment. For example, if the client achieves a standard score of 78 on a language assessment, and the LR+ at 78 was 25, the client would be 25 times as likely to have true language impairment than not to have language impairment. Thus, the clinician could be relatively certain that the client has true language impairment. Calculating a series of LR– can be clinically useful in ruling out impairment in a client. By comparing the client's performance on a test to the LR– values, the clinician can determine the probability of the client truly not having a disorder. For instance, if a client obtained a standard score of 98 on a language assessment, and the LR– value was .06, the likelihood would only be 1/17 that the client had a language impairment. Thus, the clinician can be relatively certain that the client does not have language impairment.

In order to determine the likelihood ratio for the present study, a gold standard had to be set. While it is difficult to have a definitive measure of language delay at 30 months of age, several instruments are available and commonly used to classify children. For the present study, a standardized test (PLS–3) and measures from language samples were used as the basis for the gold standard. Local norms were established for each of these measures, which were acquired from typically developing children in the Madison area (PLS–3 Expressive, $N = 105$: $M = 116.10$, $SD = 15.52$; parent–child language sample, $N = 66$: mean MLU = 2.65, $SD = 0.58$; examiner–child language sample, $N = 72$: mean MLU = 2.79, $SD = 0.61$).

Children were classified as being in the low language group if they scored more than 1 *SD* below the mean on the PLS–3 Expressive Communication section *and* more than 1 *SD* below the mean MLU, based on local norms. To meet the gold standard, the child had to score greater than 1 *SD* below the mean on either a parent–child language sample or examiner–child language sample. Several other criteria for low language were examined (e.g., –1.25 *SD*, –1.5 *SD*). However, these criteria resulted in weaker likelihood ratios, which did not improve the ability to classify children. It is important to note that the local norms on the PLS–3 Expressive are considerably higher than the national norms. Therefore, low language skills do not necessarily indicate clinical language delay. However, the low language group had lower language skills than the rest of this particular population, placing them in the lowest 16 percent of the sample population. Children who did not meet the criteria of the low language gold standard were classified as having normal language.

Agreement

Point-to-point agreement was calculated for 10% ($n = 10$) of the participants' language samples in Study 2. A second transcriber examined the original transcript and recorded the number of morphemes and utterances that were judged to be different. The transcription agreement for the parent–child language samples was: morpheme-by-morpheme: 9,522/10,605 (89.8%), utterance segmentation: 2,901/3,032 (95.7%). Agreement for the examiner–child language samples was as follows: morpheme-by-morpheme: 11,656/12,442 (93.7%), utterance segmentation: 3,214/3,341 (96.2%).

Results

Descriptive statistics. Descriptive statistics for performance on the CDI and direct language measures for all participants are presented in Table 4. Thirty-seven children met the gold standard criteria for low language at 30 months; 30 of those participants had been classified as LTs, and 7 had been classified as having normal language at 24 months based on the CDI. Sixty-three of the participants did not meet the gold standard for delay and were considered to have normal language at 30 months of age.

TABLE 4. Descriptive data (means and standard deviations) from Study 2 for the CDI and direct assessment measures at 30 months.

	Low language at 30 months		Normal language at 30 months	
	M	*SD*	*M*	*SD*
CDI: Total vocabulary	247.90	141.53	521.37	121.49
PLS–3: Expressive	88.41	8.17	116.10	12.68
MLU				
Examiner–child	1.67	0.51	2.63	0.77
Parent–child	1.50	0.40	2.51	0.62

Eight of the children with normal language had been classified as LTs, and 55 had been classified as having normal language at 24 months.

Likelihood ratio. Likelihood ratios were calculated for a variety of percentile cut-points using total productive vocabulary from the CDI–WS collected at 30 months (Table 5). The strongest likelihood ratios for a positive test result were found at the 11th percentile. Children scoring at or below the 11th percentile had a likelihood ratio of 42.5. This means that a child scoring in the 11th percentile on the CDI–WS has about 42 times the likelihood of being in the low language group versus having normal language. Such a strong likelihood ratio is due to the high true negative (specificity) rate at this cut-point. Of the 26 children who scored at or below the 11th percentile, only 1 child had normal language according to the gold standard.

At the other end, likelihood ratios for a negative test result were calculated. In the current sample, scores on the CDI at or above the 49th percentile produced likelihood ratios of zero. This means if a child scored above the 49th percentile on the CDI, it is zero times as likely that the child is in the low language group rather than the normal language group. In other words, the true positive (sensitivity) rate at the 49th percentile was outstanding, accounting for all children who met the gold standard.

While the CDI effectively classified children according to language status at the low end and upper half of the distribution (i.e., ≤ 11th percentile and ≥ 49th percentile), it was less accurate in identifying children between these two cut-points. Above the 11th percentile, likelihood ratios for positive tests began to decline rapidly. Several cut-points were tried in this middle region, resulting in likelihood ratios of modest strength. In this middle range, no clear trend was noted between percentiles on the CDI and meeting the gold standard. Between the 12th and 20th percentiles, the likelihood ratio dropped to 0.75. The likelihood ratio between the 21st and 48th percentiles was 0.59. Both ratios are not large enough to predict positive test results and are not small enough to predict negative test results. The data summarized in Table 5 illustrate the strength of identification at the low and upper half percentiles, and variability in the intermediate percentile ranges.

Receiver operating characteristic curve. Another way to assess a test's ability to appropriately identify a group of participants is through the use of a receiver operating characteristic (ROC) curve. ROC curves plot the true positive rate on the *y*-axis and the false positive rate on the *x*-axis. Sackett (1991) states that a perfect test would have a curve that goes straight up the *y*-axis (indicating a perfect true positive rate), meets in the uppermost left-hand corner, then proceeds directly across the *x*-axis (indicating a perfect false positive rate). In discussing the interpretation of ROC curves, Tape (2003) presented several hypothetical curves, which were rated as "excellent," "good," or "worthless." As the amount of area under a curve increased, the rating of the hypothetical curve increased. Tape's hypothetical curves are presented in Figure 1, with the ROC curve from the present study overlaid in a bolded black line.

Visual inspection shows that the ROC curve from the present study is well within the "good" to "excellent" range. The curve is consistent with the data from the likelihood ratio analysis. The CDI shows "excellent" ability to appropriately identify children who have low language skills, as demonstrated by the true positive line that maintains close proximity to the *y*-axis. This is related to the strong LR+ value observed at the 11th percentile. The curve also shows the "excellent" ability to identify children with average to high language skills, as the curve traverses closely to the *x*-axis and demonstrates an excellent false positive rate. This is consistent with the low LR– value at the 49th percentile. While the ROC curve from the present study is not as strong in the intermediate ranges, the CDI is still considered a "good" test according to Tape's (2003) hypothetical criteria. This reduction in strength is consistent with the decrease in likelihood ratios for scores between the 11th and 49th percentiles. It is important to note that these ROC analyses are rough estimates and should be interpreted cautiously.

Discussion

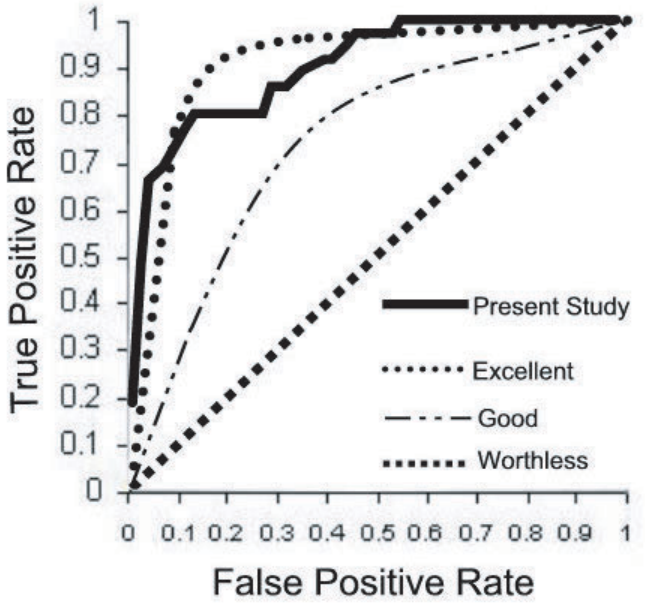
Study 2 demonstrates that the CDI is an effective tool to sort toddlers into lower and higher language level groups. Children were effectively classified through the 11th percentile due to the excellent specificity of the CDI at this level, and they were successfully classified beyond the 49th percentile due to the outstanding sensitivity at the upper end of the test. The strength of the likelihood ratios is quite telling. In fact, ratios for a positive test result were

TABLE 5. The likelihood ratio values, distribution, and proportion for children with low language (LL) and normal language (NL) at each of the percentile cuts.

Percentile	Likelihood ratio	LL		NL	
		#	Proportion	#	Proportion
≤11th	42.5[a]	25	.68	1	.02
12th–20th	0.75[b]	5	.14	12	.19
21st–48th	0.59[b]	6	.16	21	.33
≥49th	0[c]	1	.03	29	.62

[a]Likelihood ratio for a positive test result.
[b]Likelihood ratio for intermediate test results.
[c]Likelihood ratio for a negative test result.

FIGURE 1. Receiver operating characteristic (ROC) curve for data from the present study compared to hypothetical ROC curves created by Tape (2003).



stronger than other studies using likelihood ratios to examine language assessments. Using language treatment status as a gold standard, Dollaghan and Campbell (1998) calculated likelihood ratios of 25.15 for a positive result on a nonword repetition task. Values of likelihood ratios dropped to 3.73 when assessing *z* scores from the spoken language quotient of the Test of Language Development—Second Edition, with the same gold standard. Also, these likelihood ratios are much higher than Ellis Weismer et al. (2000) found for the nonword repetition task using either treatment status or standardized test scores as the gold standard. In sum, the likelihood ratios calculated for the CDI–WS are in the upper end of the range, as compared to the other studies examining language disorders that have used the same technique. It is possible that the relatively stronger likelihood ratios in the present study can, at least partially, be attributed to the fact that the assessment measure (CDI) and gold standard both focused exclusively on productive language abilities.

While the likelihood ratios for the lower and upper cut-points on the CDI for the present study were informative, the intermediate cuts were not able to effectively classify the toddlers into language groups. This same trend was noted in the Dollaghan and Campbell study (1998). Likelihood ratios dropped to 3.11 and 0.62 for two intermediate cuts on the nonword repetition task. The decreased ability to identify language levels at intermediate test points may be a general trend in language assessments. It is important to note that there are only a few studies using likelihood ratio analysis for language assessment, so comparisons and conclusions should be made with caution.

To better understand the appropriate cutoffs for various uses of the CDI, further examination of sensitivity and specificity is required. Table 6 summarizes sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) for the 11th, 19th, and 49th percentiles. PPV is the percentage of participants who screen positive and are true positives according to the gold standard. NPV is the percentage of participants who screen negative and are true negatives according to the gold standard.

As can be seen in Table 6, the 11th percentile affords excellent specificity and PPV, which is consistent with the strong likelihood ratio for a positive test result. However, sensitivity and the NPV are relatively modest at the 11th percentile. Therefore, using such a percentile cut would ensure that the majority of the children performing under this point have low language, though several children with low language would not be identified. Using this percentile cut would be useful in situations where the goal is to selectively identify children with language delay. For example, using a percentile cut with high specificity may be valuable in identifying a clinical population (e.g., LTs) for research purposes, where the goal is to identify children at the lowest performance levels while excluding children with normal language skills.

Cutting at the 49th percentile results in excellent sensitivity and NPV. However, the 49th percentile cut results in low specificity and PPV. It may initially seem desirable to use such a cutoff as a screening criterion, as all children with true language delay would potentially be identified. However, this would result in a very large number of false positives given the number of children with normal language skills below the 49th percentile.

Inspection of the data shows that the 19th percentile provides the best compromise between sensitivity and specificity values. Furthermore, PPV and NPV are more evenly distributed in comparison to the lower and higher percentile cuts. Using the 19th percentile as a criterion identifies the majority of the children with low language. At the same time, the number of children who are falsely identified is relatively modest. Such a percentile cut would be appropriate for screening children for language delay in order to determine those needing further clinical evaluation. While there would be several false positives, the majority of the children with true language delay would be identified.

General Discussion

Concurrent validity has been widely used to establish the utility of assessment batteries. Several previous studies, along with the present study, provide evidence that the CDI

TABLE 6. Additional measurement properties at three CDI cutoffs.

Percentile	Sensitivity	Specificity	PPV	NPV
11th	0.68	0.98	0.96	0.81
19th	0.81	0.79	0.70	0.89
49th	1.00	0.44	0.51	0.97

Note. PPV = positive predictive value; NPV = negative predictive value.

is significantly correlated to established measures of language development across different ages and impairment levels. There was, however, noticeable variation in the strength of correlations across studies. The relatively weaker correlations in the present investigation may have been due to the restricted range of language abilities represented in the LTs.

While concurrent validity studies have shown that the CDI is generally correlated with other language measures, such analysis does not provide insight to the test's ability to identify language levels of individual children. For such an analysis, we examined the usefulness of each performance level on the total productive vocabulary section of the CDI using likelihood ratio analysis. Scores acquired from assessments are linear in nature, with stronger language skills represented by higher percentile ranks, and lower language skills represented by lower percentile ranks. Different assessment batteries may have greater ability to identify language delay through a greater range of percentiles. Examining the diagnostic power of each percentile level shows how well a test can appropriately identify children. Likelihood ratio analysis for the CDI revealed that this measure does an excellent job of classifying 30-month-old toddlers at lower levels and average to higher levels of language performance. It is less effective at classification in the intermediate range (mid to low levels of language performance).

It is difficult to fully know how well a given test appropriately identifies children according to their language status given the lack of a definitive gold standard for early language delay. This is not a problem specific to the present study but common across the behavioral sciences. In clinical medicine, gold standards are often very distinct, definitive measures of pathology, such as a biopsy. Less invasive diagnostic markers can be compared to the definitive gold standard, providing insights regarding efficient assessment practices. Gold standards for behavioral sciences are not as discrete as most medical models and are often continuous in nature (e.g., language delay) and more difficult to define. Without a definitive diagnostic marker for comparison, an accurate gold standard can be quite elusive.

Obtaining adequate breadth across comparison measures improves the accuracy of a gold standard. Fey and Gillam (2003) cautioned against validating individual assessment measures with another similar assessment measure. They suggested the use of ecological validation measures that encompass a greater depth of language abilities, including language sample analysis. The gold standard used in this study was attained from scores that were representative of children's language skills across domains (semantics and syntax) and contexts (parent–child and examiner–child interactions). By examining children's skills across domains and contexts, we achieved some breadth and ecological validity. Furthermore, the gold standard was acquired from direct measures, including standardized assessment and language sample analysis. Such measures are sufficiently different from parent report, yet measure the same underlying skills.

It is important to note some caveats regarding the present study. To begin with, it should be emphasized that the sample was not a clinical sample. The children in the low language group had lower language skills than the rest of the participants but did not necessarily have clinical language impairment. Clinical language delay criteria vary greatly across studies and are typically greater than the –1 *SD* criterion used in the present study. While some children received speech and language services, this was not used as a criterion in the present study. Treatment status has been used as a gold standard in other studies examining validation of a language assessment measure. However, using treatment status as a gold standard for LTs would not be appropriate given the disagreement regarding appropriate intervention for LTs. Some argue that given the high proportion of LTs who catch up, either a "wait and see" (Whitehurst et al., 1991) or a "watch and see" (Paul, 1996) approach should be taken; others argue that LTs are at risk for future language impairments and should receive intervention (see discussion of this issue by Ellis Weismer, 2000). To acquire a better gold standard, future studies may try to quantify some of the variables used more often in clinical practice, including parent interview and levels of parent/teacher concern.

Another limitation of the current study is the lack of racial, ethnic, and socioeconomic diversity in the present sample. The percentile cut-points discussed in Study 2 were obtained from predominantly White children from a middle-class background and are not likely to apply across all children from diverse sociodemographic backgrounds (Feldman et al., 2000). Additional studies examining concurrent validity and using likelihood ratio analysis with a more heterogeneous group should examine the CDI's utility across cultural contexts. Given the relative simplicity of calculating likelihood ratios, such analyses can be completed within clinical practice. It is quite realistic for a clinic or school district to develop a gold standard and then to compare an assessment measure's ability to identify children who meet the gold standard. This type of analysis can provide information concerning the classification ability of a test such as the CDI for a local population and can aid in interpretation and development of assessment protocols.

Despite the concerns raised by Feldman et al. (2000) regarding the usefulness of the CDI–WS for identifying language delay at 2 years, we have documented that the measure is significantly correlated with direct assessment measures and can accurately identify children's language level at the lower end and upper half of the distribution at 30 months of age. The CDI was less effective at sorting children according to language status when they obtained midrange scores (above the 11th percentile but below the 49th percentile). Feldman et al. (2000) conclude that in-depth clinical evaluation should be used to reach conclusions regarding evaluations and management decisions. We agree that such in-depth analysis allows for the greatest amount of breadth in assessment and is the best technique to determine the level of children's language skills. The CDI appears to be a valid measure that has strong utility within such an in-depth analysis. In some situations, including large research studies and screening of large groups of children, in-depth evaluation of each child is not

realistic. For these situations, the CDI appears to be a viable measure to use by itself given the relative ease of administration and validity of the measure, particularly when using this measure to identify upper and lower ends of linguistic functioning.

Acknowledgments

Funding for this research was provided by National Institute on Deafness and Other Communication Disorders (NIDCD) Grant 5 R01 DC03731, "Linguistic Processing in Specific Language Delay," and by Core Grant P30 HD03352 to the Waisman Center from the National Institute of Child Health and Human Development. Support for John Heilmann's participation in the project was provided by NIDCD Grant 5 T32 DC005459, "Interdisciplinary Research Training in Speech-Language Disorders." We would like to extend thanks to all the children and families who made this research possible.

References

Bates, E., Marchman, V. A., Thal, D., Fenson, L., Dale, P., Reznick, J. S., et al. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language, 21,* 85–123.

Bauer, D. J., Goldfield, B. A., & Reznick, J. S. (2002). Alternative approaches to analyzing individual differences in the rate of early vocabulary development. *Applied Psycholinguistics, 23,* 313–336.

Bayley, N. (1993). *Bayley Scales of Infant Development—Second Edition.* New York: The Psychological Corporation.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57,* 289–300.

Caselli, M. C., Vicari, S., Longobardi, E., Lami, L., Pizzoli, C., & Stella, G. (1998). Gestures and words in early development of children with Down syndrome. *Journal of Speech, Language, and Hearing Research, 41,* 1125–1135.

Chilosi, A. M., Cipriani, P., Bertuccelli, B., Pfanner, L., & Cioni, G. (2001). Early cognitive and communication development in children with focal brain lesions. *Journal of Child Neurology, 16,* 309–316.

Dale, P. S. (1991). The validity of a parent report measure on vocabulary and syntax at 24 months. *Journal of Speech and Hearing Research, 34,* 565–571.

Dale, P. S., Bates, E., Reznick, J. S., & Morisset, C. (1989). The validity of a parent report instrument of child language at twenty months. *Journal of Child Language, 16,* 239–249.

Dale, P. S., Dionne, G., Eley, T. C., & Plomin, R. (2000). Lexical and grammatical development: A behavioural genetic perspective. *Journal of Child Language, 27,* 619–642.

Dollaghan, C., & Campbell, T. F. (1998). Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research, 41,* 1136–1146.

Ellis Weismer, S. (2000). Intervention for children with developmental language delay. In D. Bishop & L. Laurence (Eds.), *Speech and language impairments in children: Causes, characteristics, intervention and outcome* (pp. 157–176). East Sussex, England: Psychology Press.

Ellis Weismer, S., & Evans, J. (2002). The role of processing limitations in early identification of specific language impairment. *Topics in Language Disorders, 22,* 15–29.

Ellis Weismer, S., Murray-Branch, J., & Miller, J. (1994). A prospective longitudinal study of language development in late talkers. *Journal of Speech and Hearing Research, 37,* 852–867.

Ellis Weismer, S., Tomblin, J. B., Zhang, X., Buckwalter, P., Chynoweth, J. G., & Jones, M. (2000). Nonword repetition performance in school-age children with and without language impairment. *Journal of Speech, Language, and Hearing Research, 43,* 865–878.

Farrar, M. J., & Maag, L. (2002). Early language development and the emergence of a theory of mind. *First Language, 22,* 197–213.

Feldman, H. M., Dollaghan, C. A., Campbell, T. F., Colborn, D., Janosky, J. E., Kurs-Lasky, M., et al. (2003). Parent-reported language skills in relation to otitis media during the first 3 years of life. *Journal of Speech, Language, and Hearing Research, 46,* 273–287.

Feldman, H. M., Dollaghan, C. A., Campbell, T. F., Kurs-Lasky, M., Janosky, J. E., & Paradise, J. L. (2000). Measurement properties of the MacArthur Communicative Development Inventories at ages one and two years. *Child Development, 71,* 310–322.

Fenson, L., Bates, E., Dale, P., Goodman, J., Reznick, J. S., & Thal, D. (2000). Measuring variability in early child language: Don't shoot the messenger. *Child Development, 71,* 323–328.

Fey, M., & Gillam, R. (2003, June). *Measuring language development and language disorders: Documenting treatment outcomes.* Paper presented at the Symposium for Research in Child Language Disorders, Madison, WI.

Frankenburg, W. K., Dodds, J., Archer, P., Bresnick, B., Maschka, P., Edelman, N., et al. (1992). *Denver II.* Denver, CO: Denver Developmental Materials.

Fudala, J. B. (2000). *Arizona Articulation Proficiency Scale—Third Revision.* Los Angeles: Western Psychological Services.

Gershkoff-Stowe, L., Thal, D., Smith, L., & Namy, L. (1997). Categorization and its developmental relation to early language. *Child Development, 68,* 843–859.

Girolametto, L., Pearce, P., & Weitzman, E. (1996). Interactive focused stimulation for toddlers with expressive vocabulary delays. *Journal of Speech and Hearing Research, 39,* 1274–1283.

Girolametto, L., Wiigs, M., Smyth, R., Weitzman, E., & Pearce, P. (2001). Children with a history of expressive vocabulary delay: Outcomes at 5 years of age. *American Journal of Speech-Language Pathology, 10,* 358–369.

Harris, N. G. S., Bellugi, U., Bates, E., Jones, W., & Rossen, M. (1997). Contrasting profiles of language development in children with Williams and Down syndromes. *Developmental Neuropsychology, 13,* 345–370.

Hick, R. F., Joseph, K. L., Conti Ramsden, G., Serratrice, L., & Faragher, B. (2002). Vocabulary profiles of children with specific language impairment. *Child Language Teaching and Therapy, 18,* 165–180.

Lyytinen, P., Eklund, K., & Lyytinen, H. (2003). The play and language behavior of mothers with and without dyslexia and its association to their toddlers' language development. *Journal of Learning Disabilities, 36,* 74–86.

Marchman, V. A., & Bates, E. (1994). Continuity in lexical and morphological development: A test of the critical mass hypothesis. *Journal of Child Language, 21,* 339–366.

Miller, J. F., & Chapman, R. S. (2002). Systematic Analysis of Language Transcripts (Research Version 7.0) [Computer software]. Madison, WI: Language Analysis Laboratory.

Miller, J. F., Sedey, A. L., & Miolo, G. (1995). Validity of parent report measures of vocabulary development for children with Down syndrome. *Journal of Speech and Hearing Research, 38,* 1037–1044.

Paul, R. (1996). Clinical implications of the natural history of slow expressive language development. *American Journal of Speech-Language Pathology, 5,* 5–21.
Paul, R. (2001). *Language disorders from infancy through adolescence: Assessment & intervention* (2nd ed.). St. Louis, MO: Mosby.
Paul, R., & Smith, R. (1993). Narrative skills in 4-year-olds with normal, impaired, and late-developing language. *Journal of Speech and Hearing Research, 36,* 592–598.
Rescorla, L., & Goossens, M. (1992). Symbolic play development in toddlers with expressive specific language impairment (SLI-E). *Journal of Speech and Hearing Research, 35,* 1290–1302.
Richards, B. J., & Malvern, D. D. (1997). *Quantifying lexical diversity in the study of language development.* Reading, England: The University of Reading, New Bulmershe Papers.
Robbins, J., & Klee, T. (1987). Clinical assessment of oropharyngeal motor development in young children. *Journal of Speech and Hearing Disorders, 52,* 271–277.
Robertson, S., & Ellis Weismer, S. (1999). Effects of treatment on linguistic and social skills in toddlers with delayed language development. *Journal of Speech, Language, and Hearing Research, 42,* 1234–1248.
Sackett, D. L. (1991). *Clinical epidemiology: A basic science for clinical medicine* (2nd ed.). Boston: Little Brown.
Tape, T. (2003). *The area under an ROC curve.* Retrieved July 9, 2003, from University of Nebraska Medical Center, Department of General Internal Medicine Web site: http://gim.unmc.edu/dxtests/roc3.htm
Thal, D., Bates, E., Goodman, J., & Jahn-Samilo, J. (1997). Continuity of language abilities: An exploratory study of late- and early-talking toddlers. *Developmental Neuropsychology, 13,* 239–273.
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,* 482–496.
Thal, D., Tobias, S., & Morrison, D. (1991). Language and gesture in late talkers: A 1-year follow-up. *Journal of Speech and Hearing Research, 34,* 604–612.
Thorndike, R., Hagen, E., & Sattler, J. (1986). *The Stanford–Binet Intelligence Scale, Fourth Edition.* Chicago: Riverside.
Whitehurst, G. J., Fischel, J. E., Lonigan, C. J., Valdez-Menchaca, M. C., Arnold, D. S., & Smith, M. (1991). Treatment of early expressive language delay: If, when, and how. *Topics in Language Disorders, 11,* 55–68.
Yoshinaga-Itano, C., Snyder, L. S., & Day, D. (1998). The relationship of language and symbolic play in children with hearing loss. *Volta Review, 100,* 135–164.
Zimmerman, I., Steiner, V., & Pond, R. (1992). *Preschool Language Scale—Third Edition.* Chicago: The Psychological Corporation.

Received October 30, 2003
Revision received June 7, 2004
Accepted December 13, 2004
DOI: 10.1044/1058-0360(2005/006)

Correspondence concerning this article should be addressed to John Heilmann, 1975 Willow Drive, Madison, WI 53706. E-mail: jjheilmann@wisc.edu

Appendix A

Summary of Concurrent Validity Studies

	CDI: Total productive vocabulary	CDI: Three longest utterances	CDI: Complexity
MLU (language sample)	.68[a]**	.74**	.76**
	.84[b]**	.63**	.69**
Number of different words (language sample)	.74[a]**	.71**	.77**
	.78[b]**	.56**	.62**
	.75[c]**		
	.82[d]**		
Type-token ratio	.53[a]**	.38	.47*
	–.18[b]	–.27	–.04
IPSyn	.78[a]**	.78**	.79**
	.83[b]**	.58**	.67**
EOWPVT	.73[a]**	.54**	.54**
	.86[b]**	.61**	.77**
Stanford–Binet memory for sentences	.75[a]**	.48*	.66**
	.67[b]**	.59**	.52**
Bayley—Expressive Language items	.70[c]**		
	.77[d]**		
	.54[e]**		

Note. CDI = Communicative Development Inventory; MLU = mean length of utterance; IPSyn = Index of Productive Syntax; EOWPVT = Expressive One-Word Picture Vocabulary Test; Stanford–Binet = Stanford–Binet Intelligence Scale, Fourth Edition (Thorndike et al., 1986).

[a]Data for row are from Dale (1991): 24 children, 2 years old, typically developing.
[b]Data for row are from Thal et. al (1999), experiment 1: 20 children, ages 39–49 months, with specific language impairment.
[c]From Miller et al. (1995): 46 children, ages 11–26 months, typically developing.
[d]From Miller et al. (1995): 44 children, ages 16–68 months, with Down syndrome.
[e]From Dale et al. (1989): 32 children, 20 months of age, typically developing.

*$p < .05$. **$p < .01$, two-tailed.

Appendix B

Calculation of Likelihood Ratios (Adapted from Sackett, 1991)

Sensitivity, specificity, and likelihood ratios for positive and negative test results are calculated for each level of a test result (e.g., percentiles and standard score values). The first step is to complete the following table for each test result level. Then, calculate sensitivity and specificity: sensitivity = true positive/(true positive + false negative); specificity = true negative/(false positive + true negative). Finally, calculate likelihood ratio for a positive test result (LR+) and a negative test result (LR–): LR+ = sensitivity/(1 – specificity); LR– = (1 – sensitivity)/specificity.

	Gold standard	
Experimental measure	Positive	Negative
Positive	True positive	False positive
Negative	False negative	True negative

Note. True positive = number of participants who scored positive on both the experimental measure and gold standard; false negative = number of participants who did not meet the experimental measure criterion but did meet the gold standard criterion; false positive = number of participants who met the experimental measure criterion but did not meet the gold standard criterion; true negative = number of participants who did not meet both the experimental measure and the gold standard criteria.

Appendix C

Example: Calculation of LR+ and LR– for the 11th Percentile on the CDI (Experimental Measure)

	Gold standard	
Experimental measure	Positive	Negative
Positive	25	1
Negative	12	62

Note. Sensitivity = 25/(25 + 12) = 0.6757; specificity = 62/(1 + 62) = 0.9841; LR+ = 0.6757/(1 – 0.9841) = 42.5; LR– = (1 – 0.6757)/0.9841 = 0.33.