CICSD

A Psychometric Analysis of Childhood Vocabulary Tests

Ellen L. Bogue
Laura S. DeThorne
University of Illinois at Urbana—Champaign

Barbara A. Schaefer
Pennsylvania State University, University Park

Standardized tests, defined as measures that provide normative data through the use of standardized administration materials and procedures, are commonly used by both clinicians and investigators within speech-language pathology. Standardized tests are employed across ages, settings, and disorders for a variety of purposes. Within speech-language pathology, such normed measures are often used to screen for deficits, to identify specific areas of strengths and weaknesses, to plan intervention, and to monitor language progress (Merrell & Plante, 1997; Plante & Vance, 1995).

A child's performance on a standardized test will have different implications depending on the purpose for which the test is being used, as DeThorne and Schaefer (2004) noted in their discussion of high- versus low-stakes testing. A *high-stakes* testing environment is one in which the outcome will determine an individual's diagnosis, educational placement, or treatment eligibility, whereas screening or research studies would be considered *low-stakes* testing environments, at least for the individual being tested.

Because standardized vocabulary tests are commonly used for a variety of high-stakes purposes, it is important for clinicians to understand how effective a vocabulary test is for its intended purpose. Understanding the effectiveness of standardized tests is contingent in part on their psychometric properties.

ABSTRACT: **Purpose**: The psychometric properties of 10 standardized vocabulary tests were reviewed in order to help empower practicing clinicians and investigators to become more critical consumers of such products. **Method:** Ten standardized vocabulary tests were selected for this study through a systematic review of the University of Illinois' test collections, online databases, and ancestral searches. Each measure was evaluated based on established criteria related to sample standardization, reliability, and validity. **Results:** Of the 10 tests, 2 met all of the preset criteria for the standardization sample, but no test met all of the criteria in the area of reliability or validity. The Boehm Test of Basic Concepts, Third Edition (Boehm, 2000) and the Montgomery Assessment of Vocabulary Acquisition (MAVA; Montgomery, 2008) met the highest number of criteria, meeting only 7 and 6, respectively, of the total 11 criteria specified. Only 1 test, the MAVA, provided direct evidence of sensitivity and specificity. **Conclusion:** Clinicians and investigators need to be aware of the psychometric properties of the specific measures they use and be sure to integrate multiple forms of information when deciding which test to use for vocabulary assessment.

KEY WORDS: vocabulary, assessment, psychometrics, reliability, validity

Three aspects of psychometrics are considered here: those related to the constituency of the standardization or normative sample, evidence of reliability, and evidence of validity. Each of these areas will be reviewed, followed by a general discussion of language assessment procedures, with specific focus on vocabulary in particular.

Psychometric Properties of Test Construction

Standardization sample. The first area of psychometric analysis relates to the quality of the standardization sample, which must be adequate in size, representativeness, and recency. In order for a standardization sample to be representative, it must be relatively large in size in order to encompass "the full range of the distribution" (Andersson, 2005, p. 208). A common sample size criterion is 100 or more individuals for each normative group in the sample, which is most often divided by age and occasionally by grade level (Andersson, 2005; DeThorne & Schaefer, 2004; McCauley & Swisher, 1984), although other researchers suggest a more stringent criterion, particularly for very young children (Alfonso & Flanagan, 2009).

In the present study, we initially considered defining the standardization sample based on how the norms were presented for scoring in the conversion tables (Bogue & DeThorne, 2012), but this was highly variable and often defined a subgroup as small as a month, which seemed too stringent, or as large as 2+ years, which seemed too broad. Consequently, for the purposes of this analysis, we compromised and defined each subgroup by a 1-year interval, as this is consistent with how children are often grouped in the educational system.

A representative sample is also one that includes individuals who are characteristic of the population for whom the test was developed, particularly with regard to cultural background and language variation. Current guidelines for test development recommend that the standardization sample contain groups that are in proportion with the overall population with regard to race/ethnicity, geographic region, parent education level/socioeconomic status, and gender (Alfonso & Flanagan, 2009). With regard to including individuals with impairments in the normative sample, professional views are mixed. Some professionals suggest that children with impairments should be included in the sample so that the full range of language abilities is accurately represented (Andersson, 2005; DeThorne & Schaefer, 2004). This position argues that failing to include children with language impairment (LI) in the standardization sample would serve to inflate the normative data and potentially increase false negatives in the assessment process. However, including children with LI in standardization samples may decrease a test's sensitivity (Pena, Spaulding, & Plante, 2006). In sum, the disagreement surrounding the inclusion of individuals with impairments within test construction highlights the frequent trade-offs inherent in maximizing test sensitivity and specificity.

The final aspect of the standardization sample to be addressed here is recency. Because characteristics of a population can change over time, it is also important for the standardization sample to be fairly recent: A test will not be effective if a child is being measured against an outdated sample. A prime example of recency from the realm of cognitive testing is the Flynn effect, which is the tendency of IQ scores to increase by three points per decade (Flynn, 1999). We could argue that vocabulary is particularly sensitive to changes over time, as new words are coined and meanings easily change within the span of a generation. Also, the objects often portrayed within vocabulary test picture plates, such as telephones and computers, change in appearance. For example, telephones today look very different from how they did 15 years ago.

Reliability. Another important criterion for a test in psychometric evaluation is reliability, which is a measure of a test's consistency across examiners (interrater reliability), across test items (internal consistency), and over time (test–retest reliability). *Interrater* reliability measures the consistency with which separate administrators score a test, ensuring that scores do not vary considerably between different raters. *Internal* consistency compares a child's scores on one subset of test items (e.g., odd-numbered items) with scores on another subset (e.g., even-numbered items; the average of all possible ways to split the test using Cronbach's coefficient alpha is preferable; Suen & Ary, 1989), thus measuring the consistency of the construct being evaluated across items. *Test–retest* reliability reflects the correlation between an individual's scores on the same test administered after a certain time period. This evaluates how reliable a child's scores would be on subsequent administrations of the same test. Across most types of reliability, a common criterion for subtests and composite scores is a coefficient of greater than or equal to .90 within each normative group (Alfonso & Flanagan, 2009; Andersson, 2005; DeThorne & Schaefer, 2004; McCauley & Swisher, 1984).

Related to reliability is the concept of standard error of measurement (*SEM*), which is derived from a test's internal reliability and allows the calculation of a confidence interval (CI) for an individual child's score based on the inherent error of test construction and administration (Sattler, 2001). CIs typically

correspond to standard deviations of the normal curve forming either a 68% or 95% CI. For example, if a child receives a standard score of 94 and the *SEM* is ±4 standardized points, then the 68% CI for the child's score would be 90–98. The statistical interpretation of such a CI is that if the child were administered the same test on 100 occasions, the true score would likely fall between 90 and 98 68% of the time. A smaller *SEM* results in a tighter CI, which corresponds to greater confidence in the child's score. Although *SEM* can be calculated by the test administrator from the reliability estimate and standard deviation of the distribution, provision of the *SEM* in the test's manual for each normative group makes it more accessible and likely to be considered by test administrators.

Validity. The third area of psychometric evaluation is validity. According to Messick (2000), validity is a property of test scores and is a singular concept that formerly was referred to as construct validity; however, "several complementary forms of evidence need to be integrated in construct validation, evidence bearing on test content, score structure, substantive processes, generalizability, external relationships, and testing consequences" (p. 4). Validity is a measure of how well a test score assesses the construct it claims to test, which is the most important and relevant measure of a test's effectiveness. A test's score can be reliable without being valid, but it cannot be valid without being reliable.

Like reliability, evidence for validity takes many forms, but unlike reliability, established criteria are difficult to find. Evidence of validity can be reflected in developmental trends, correlation with similar tests, factor analyses, group comparisons, and predictive validity (DeThorne & Schaefer, 2004; McCauley & Swisher, 1984; Sattler, 2001). Regarding developmental trends, a skill like language development is expected to improve with age. Consequently, raw scores on a language test are expected to increase with age. As such, evidence of a positive association between age and raw scores provides a basic form of validity evidence. Although language raw scores should improve with age, a number of other factors also develop with age, so this particular form of validity is far from sufficient in documenting the validity of a test's construction (DeThorne & Schaefer, 2004).

A second form of support for convergent validity evidence comes from a test's correlation with other tests that are designed to assess a similar construct. For example, a newly developed vocabulary test would be expected to correlate highly with other commonly used vocabulary measures. This facet of validity only has strength if the test scores being used as a comparison are themselves valid and reliable: A high correlation with a poorly constructed test only means that the new test is similarly flawed.

Also significant to the evaluation of validity are factor analyses, group comparisons, and predictive validity. Factor analyses assist in determining the interrelationships between variables or items so as to investigate how different items are related to each other, and results can determine whether they are measuring the same or contrasting construct or skill (Gorsuch, 1983, 1997; Sattler, 2001). Although factor analyses can be applied to unidimensional measures such as standardized vocabulary tests to confirm item contributions and inclusion, factor analysis is more commonly applied to multidimensional assessments, which are not the focus of this study; thus, factor analysis will not be reviewed here. However, particularly germane to the validity of standardized vocabulary tests is the concept of group comparisons. As the name implies, group comparisons involve administering a test to relevant subgroups of a population. Relevant in this case would refer to children with vocabulary impairments compared to peers with typically developing (TD) language. Because these two groups, by definition, differ in their vocabulary abilities, a difference between these two groups that favored those with TD language would provide evidence of a language test's validity. Despite the strength of this approach, group differences can still mask a substantial amount of individual variation. Said another way, subgroup distributions can overlap substantially even when the means differ. Ultimately, the extent of overlap between groups governs a test's diagnostic accuracy.

Related most directly to diagnostic accuracy is a test's evidence of sensitivity and specificity. *Sensitivity* measures how well a test score identifies individuals who possess a certain trait; *specificity* measures how well the score classifies those without the trait. In both instances, this is based on a predetermined gold standard. Because one of the most common uses of standardized vocabulary tests is to assist in the diagnosis of LI, a test needs to be strong in both sensitivity and specificity. Unfortunately, high sensitivity (i.e., identifying all children with true vocabulary impairment) often increases the likelihood of overidentifying TD children as impaired (i.e., false positives), thereby leading to lower specificity. Similarly, high specificity often increases the likelihood of missing children with true impairment (i.e., false negatives), thereby leading to lower sensitivity.

Both sensitivity and specificity are likely to vary based on the cutoff criterion used to categorize a child's language as impaired as well as the prevalence of LI in the standardization sample. Consequently, a

test manual should ideally report both sensitivity and specificity for each normative group based on the most commonly employed cutoff criteria (i.e., –1.0 to –1.9 *SD*s below the mean; Eickhoff, Betz, & Ristow, 2010). The test's sensitivity and specificity values should also meet specified criteria. For example, Hargrove (2006) suggested 80% sensitivity and 90% specificity, whereas Plante and Vance (1995) linked acceptability to the purpose of testing. For diagnostic purposes, Plante and Vance suggest that 90% accuracy is *good* and 80% is *fair*; assuming these values apply to both sensitivity and specificity. For screening, Plante and Vance recommend 90%–100% sensitivity, 80% specificity for a *good* rating, and 70% for *fair*.

Unlike sensitivity and specificity values, which relate to present conditions, predictive validity attempts to determine a test score's ability to predict an individual's performance over time as well as in related areas, such as success in school and reading ability. Although an important form of validity evidence, such information is rarely provided in test manuals, perhaps due to the required longitudinal nature of such data. Due to this reason, and to the fact that criteria for adequate predictive validity have not been established, we did not focus on predictive validity in our test review.

Strengths and Weaknesses

Standardized measures are a key component of most diagnostic batteries but are also inherently limited. The strength of standardized measures lies in their ability to provide both a quantitative depiction of a child's abilities as well as normative data against which an individual's score can be compared. Assuming that a test is psychometrically sound, having a quantitative score obtained under controlled circumstances makes it easier to compare a child to his or her peers and to determine whether or not the child's performance can be considered within *average* range.

Despite such strengths, clinicians should not rely on standardized tests as the sole means of assessment (McCauley & Swisher, 1984). Even with sound psychometric properties, standardized tests only provide information about a child's intrinsic abilities during a single point in time within a relatively contrived context. As such, test scores are influenced by factors other than vocabulary skill, such as attention, frustration tolerance, test anxiety, and familiarity with the examiner or testing situation (Fleege, Charlesworth, Burts, & Hart, 1992; Fuchs & Fuchs, 1986; Speltz, DeKlyen, Calderon, Greenberg, & Fisher, 1999). Consequently, standardized tests need to be balanced with other forms of assessment (Watkins & DeThorne, 2000), particularly those with stronger social validity, meaning that they are more representative of language use in everyday contexts. As an example that often resonates within the speech-language pathology community, consider applications to graduate school. Few individuals, students or professors, would advocate for using GRE scores as the sole assessment of graduate school qualification. Many would advocate for the inclusion of more functional and situated assessments, such as letters of recommendation and interviews. Similarly, the validity of child vocabulary tests is supported through information from observational measures and caregiver report, particularly for children from nonmainstream populations (Washington, 1996; Watkins & DeThorne, 2000).

Parent Report Measures

Due in part to limitations of standardized tests, parent report measures are an important additional form of evidence in any assessment of a child's language. Parent report measures are those measures that are intended to be completed by the caregiver of the child under consideration. These measures often take the form of checklists or Likert-scale items that are given to parents to complete. Like other methods of assessment, parent report measures have both advantages and disadvantages. In terms of advantages, parent report measures are cost and time effective for the clinician and less stressful for the child than standardized measures. According to Watkins and DeThorne (2000), such measures can also be useful in identifying areas of strength and weakness and offer inherent social validity, meaning that they represent how well a child is functioning in his or her community. Parent report measures are also valuable because parents are able to observe their child's language skills in a variety of settings and across time.

Including parents in the process of assessing children's language skills is critical in maximizing the accuracy of professionals' assessments and the effectiveness of clinicians' interventions (cf., American Academy of Pediatrics, 2003; McCollum & Yates, 1994; Prelock, 2006; Thies & McAllister, 2001). Clinically, parent report measures provide one relatively easy and valid means to incorporate caregivers in the assessment process (Dinnebeil & Rule, 1994; Oliver et al., 2002; Saudino et al., 1998). Dinnebeil and Rule's (1994) study demonstrated the validity of parents' estimations of their children's skills through a review of the literature concerning the congruence of parents' and professionals' judgments. Results of their review of 23 studies demonstrated a strong positive correlation between the two, with a mean correlation coefficient of .73.

Despite such evidence, parent report, like any single measure, is limited in its perspective. One limitation of parent report is that normative data may not be available, thereby making it difficult to compare a child to his or her peer group. One parent report measure of vocabulary that does provide normative data is the MacArthur-Bates Communicative Development Inventory, Third Edition (MCDI–III**;** Fenson et al., 2007). Caregivers complete a vocabulary checklist on a standard form and then the examiner is able to compare those results to norms collected in comparable circumstances. Given the normative data, paired with consistent scoring and administration procedures, the MCDI–III is included in the present review. In sum, the most valid assessment results are likely to emerge from a combination of assessment methods, integrating results from standardized testing, parent report, and behavioral observation (Watkins & DeThorne, 2000).

Previous Research

Previous research has demonstrated that many standardized tests may fall short of psychometric expectations. McCauley and Swisher (1984), for example, reviewed 30 preschool language and articulation tests and found that three tests of vocabulary—the Expressive One-Word Picture Vocabulary Test (EOWPVT; Gardner, 1979), the Peabody Picture Vocabulary Test (PPVT; Dunn, 1965), and the Peabody Picture Vocabulary Test—Revised (PPVT–R; Dunn & Dunn, 1982)—failed to meet psychometric criteria. The 30 tests were evaluated on the basis of 10 criteria related to standardization sample, reliability, and validity; half of the criteria were met by fewer than six tests. Results of a review of 21 tests of child language, two of which were vocabulary tests, published 10 years later (Plante & Vance, 1994, p. 16) suggested that there was "little improvement in overall quality" of tests since McCauley and Swisher's study. The Plante and Vance (1994) review used the same 10 criteria as McCauley and Swisher's study and similarly found that of the 21 tests reviewed, no test met more than seven of the psychometric criteria.

Studies regarding the diagnostic accuracy of various global language measures have also been performed, suggesting that although some tests might be fairly accurate in identifying an impairment, others may not be as precise. For example, Spaulding, Plante, and Farinella (2006) demonstrated that for a majority of the 43 tests in their review of child language measures, the scores of children with a previously identified LI were not consistently at the low end of the distribution. Ten of the tests included in their study were standardized vocabulary tests: the Boehm Test of Basic Concepts, Third Edition (Boehm–3; Boehm, 2000); Boehm Test of Basic Concepts—Preschool (Boehm–P3; Boehm, 2001); Comprehensive Receptive and Expressive Vocabulary Test, Second Edition (CREVT–2; Wallace & Hammil, 2002); Expressive One-Word Picture Vocabulary Test—Revised (EOWPVT–R; Gardner, 1990); Expressive Vocabulary Test (EVT; Williams, 1997); Peabody Picture Vocabulary Test, Third Edition (PPVT–III; Dunn & Dunn, 1997); Receptive One-Word Picture Vocabulary Test (ROWPVT; Gardner, 1985); Test of Word Knowledge (TWK; Wiig & Secord, 1992); Word Test—Adolescent (WORD: A; Zachman, Huisingh, Barrett, Orman, & Blagden, 1989); and Word Test, Elementary—Revised (WORD: R; Huisingh, Barrett, Zachman, Blagden, & Orman, 1990). For four of the vocabulary tests involved in the study (CREVT–2, EVT, PPVT–III, and Boehm–3), mean differences between the group with LI and the normative or control groups were less than 1.5 *SD*s. This highlights the substantial overlap that may occur between the scores of the two groups.

Similar to Spaulding et al. (2006), the results of a study of four vocabulary tests conducted by Gray, Plante, Vance, and Henrichsen (1999) suggested that none of the tests in their study was a strong indicator of specific language impairment (SLI). In their study, the PPVT–III, ROWPVT, EOWPVT, and EOWPVT–R (all of which are reviewed, based on their most recent editions, in the present study) were administered to preschool-age children with SLI and to TD preschool-age children. Although the children with SLI did score lower than the TD children, they still scored within the typical range.

Although vocabulary tests have been included in reviews of standardized language measures, we were unable to find a comprehensive review focused on unidimensional vocabulary tests. We considered unidimensional tests to be those that focus on vocabulary only and no other aspects of language development and those that are not subtests of a more comprehensive assessment, such as the Clinical Evaluation of Language Fundamentals, Fourth Edition (CELF–4; Semel, Wiig, & Secord, 2003) or the Preschool Language Scale, Fifth Edition (PLS–5; Zimmerman, Steiner, & Pond, 2011). Consequently, the aim of the present study was to review common unidimensional child vocabulary tests on the basis of specific psychometric criteria in terms of the standardization sample, reliability, and validity. As vocabulary tests are often key components in high-stakes assessments, speech-language pathologists must be fully aware of each test's psychometric properties. The ultimate goal of this review, then, is to aid clinicians and researchers in making informed decisions regarding

the administration and interpretation of standardized vocabulary tests.

METHOD AND RESULTS

Test Selection

Tests were included in this review if they met three inclusionary criteria derived from the literature on test development (e.g., Alfonso & Flanagan, 2009; DeThorne & Schaefer, 2004; McCauley & Swisher, 1984; Plante & Vance, 1994). First, each test had to be standardized in the sense that it employed prescribed materials and procedures and provided normative data. Second, the test had to be a unidimensional test of vocabulary in children under 18 years of age. Multidimensional tests that included a vocabulary subtest as part of a larger, more comprehensive assessment were not included (e.g., CELF–4, PLS–5, and Test of Language Development—Intermediate, Fourth Edition [Hammill & Newcomer, 2008]). Several of the tests evaluated include normative data past 18 years of age, but only the data provided for 18 years and younger were reviewed. Finally, tests also had to have been developed or revised within the past 20 years.

Based on these inclusionary criteria, relevant tests were first identified through test inventories from the applied health sciences library and the speech-language pathology clinic of the University of Illinois, literature review via online databases (e.g., PsycInfo, ERIC, PubMed), and ancestral searches. These search procedures identified 10 standardized vocabulary tests, summarized in Table 1, that served as the focus of this review. These tests are all assessments of semantic knowledge—three targeting receptive knowledge only, four targeting expressive knowledge only, and three tapping both receptive and expressive knowledge. In terms of required tasks, all but the MCDI–III and the WORD Tests (The WORD Test 2: Adolescent [WORD–2:A; Bowers, Huisingh, LoGiudice, & Orman, 2005]; The WORD Test 2: Elementary [WORD–2:E; Bowers, Huisingh, LoGiudice, & Orman, 2004]) include a picture-labeling component, for example, "Show me X" or "What is this?" In contrast, the WORD tests involve subtests that are focused on associations, synonyms, semantic absurdities, antonyms, definitions, and flexible word use. Unlike all of the other measures, the MCDI–III is a parent report measure that includes a checklist of common early vocabulary. Caregivers are asked to fill in bubbles next to the words their children say and/or understand. Although different in terms of administration procedures from the other tests, the MCDI–III is included here given its standardized administration, normative data, and heavy usage in the assessment of vocabulary in early intervention.

Review Process

We evaluated each test on the basis of its psychometric properties, including the makeup of the standardization sample as well as evidence of score reliability and validity, largely following the criteria set forth by DeThorne and Schaefer (2004) and informed by the criteria identified in Alfonso and Flanagan (2009) for preschool tests. The evaluation was based exclusively on the information that was provided in the test manuals, which were individually reviewed by the first author through multiple passes. The specification of evaluation criteria is summarized below according to the three primary areas of standardization sample, reliability, and validity.

Standardization sample. The standardization sample was considered adequate based on three criteria taken from DeThorne and Schaefer (2004): adequacy of size, comparison to census data, and recency. A summary of performance of each individual criterion is provided in the following paragraphs. Information regarding the standardization sample of the individual tests is summarized in columns 2–4 of Table 2.

Size. First, in terms of sample size, at least 100 individuals in each normed subgroup were needed, meaning that, for the purposes of this study, each 1-year interval (whether by age or grade level) had to include 100 children or more. Although the specific number may seem arbitrary, it is consistent with criteria detailed by Alfonso and Flanagan (2009). The rationale is that each normative group needs to be large enough to capture the inherent variability associated with any trait (Sattler, 2001).

Of the 10 tests reviewed, the Boehm–3 and the MCDI–III met the size criterion, with the remaining eight tests failing. The Expressive One-Word Picture Vocabulary Test, Fourth Edition (EOWPVT–4; Martin & Brownell, 2011a) and Receptive One-Word Picture Vocabulary Test, Fourth Edition (ROWPVT–4; Martin & Brownell, 2011b), which were normed using the same sample, failed to include 100 individuals for three age groups, ages 2 (n = 82), 5 (n = 86), and 11 (n = 96), and collapsed groups for ages 13 and 14, 15 and 16, and 17 and 18. The Expressive Vocabulary Test, Second Edition (EVT–2; Williams, 2007) and Peabody Picture Vocabulary Test, Fourth Edition (PPVT–4; Dunn & Dunn, 2007), also normed on the same sample, failed to provide a sample size per year, collapsing ages 15 and 16 as well as ages 17 and 18. The WORD–2:E missed the cutoff by only two individuals (n = 98) for the 11;6–11;11

Table 1. Summary of the 10 childhood vocabulary measures reviewed in this study.

Test	*Age range (years;months)*	*Testing time*	*Subtests*	*Picture plate description*	*Price*
Boehm Test of Basic Concepts, Third Edition (Boehm–3)[a]	Kindergarten–Second Grade	1 session, 45 min 2 sessions, 30 min each		Full color drawings	$154 for a complete kit of one form
Comprehensive Receptive and Expressive Vocabulary Test, Second Edition (CREVT–2)	Expressive: 4;0–89;11 Receptive: 5;0–89;11	Both subtests, 20–30 min One subtest, 10–15 min	Receptive & Expressive	Color photographs, six pictures per plate	$279
Expressive One-Word Picture Vocabulary Test (EOWPVT–4)[b]	2;0–80+	15–20 min		One picture per plate	$175
Expressive Vocabulary Test (EVT–2)[c]	2;6–81+	10–20 min		Full color drawings, one picture per plate	$419 for Forms A & B, $227 for one form
MacArthur-Bates Communicative Development Inventory, Third Edition (MCDI–III)	CDI: Words and Gestures: 8–18 months CDI: Words and Sentences: 16–30 months	N/A		N/A	$121.95 (Words and Gestures & Words and Sentences)
Montgomery Assessment of Vocabulary Acquisition (MAVA)	3;0–12;11	30–40 min for both tests	Receptive & Expressive	Full color drawings Receptive: four pictures per plate Expressive: one picture per plate	$199
Peabody Picture Vocabulary Test, Fourth Edition (PPVT–4)[c]	2;6–81+	10–15 min		Full color drawings, one picture per plate	$419 for Forms A & B, $227 for one form
Receptive One-Word Picture Vocabulary Test, Fourth Edition (ROWPVT–4)[b]	2;0–80+	15–20 min		Full color line drawings; four pictures per plate	$175
The WORD Test 2: Adolescent (WORD–2:A)	12;0–17;11	30 min	Tasks: Associations, Synonyms, Semantic Absurdities, Antonyms, Definitions, Flexible Word Use	N/A	$160
The WORD Test 2: Elementary (WORD–2:E)	6;0–11;11	30 min	Tasks: Associations, Synonyms, Semantic Absurdities, Antonyms, Definitions, Flexible Word Use	N/A	$160

[a]Spanish version also available; Available in forms E & F. [b]Spanish version available. [c]Available in Forms A & B.

Table 2. Evaluation of each of the 10 measures reviewed based on specific psychometric criteria.

	Standardization sample			*Reliability*				*Validity*			
Vocabulary test	*Sizeable*	*Census data*	*Recent*	*Internal*	*Test–retest*	SEM	*Inter-examiner*	*Develop-mental*	*Test comparison*	*Group comparisons*	*Sensitivity & specificity*
Boehm–3	+	+	+	–	+	+	0	+	+	0	0
CREVT–2	–	+	+	–	–	–	–	–	+	+	0
EOWPVT–4	–	+	+	–	–	–	0	–	+	+	0
EVT–2	–	+	+	–	–	–	0	–	+	+	0
MCDI–III	+	+	+	–	–	–	0	+	+	–	0
MAVA	–	+	+	+	–	0	–	+	+	–	+
PPVT–4	–	+	+	–	–	–	0	–	+	+	0
ROWPVT–4	–	+	+	–	–	–	0	–	+	+	0
WORD–2:A	–	+	+	–	–	0	–	–	0	+	0
WORD–2:E	–	+	+	–	+	0	–	+	0	+	0

Note. + = specified criteria met; – = specified criteria not met; 0 = no evidence provided in the test manual. SEM = standard error of measurement.

(years;months) group, and the WORD–2:A failed due to collapsing the 16;0–17;11 group rather than dividing it per year. The Montgomery Assessment of Vocabulary Acquisition (MAVA; Montgomery, 2008) failed to meet the size criterion for three groups: 12;0–12;11 for the Receptive test (n = 78), and 11;0–11;11 (n = 92) and 12;0–12;11 (n = 84) for the Expressive test. The CREVT–2 failed to meet the size criterion because it did not present numbers in 1-year intervals, instead dividing the sample into broad age categories (e.g., "elementary school," "secondary school," "young adults"). Finally, it seems important to mention that the MCDI–III normative data separate girls and boys and delineate 1-month subgroups. Although the younger age group of the MCDI–III, compared to other tests, may have been reason to define narrower subgroups for our size criterion, we kept the 1-year standard for consistency across tests.

Census data. The second criterion, also concerned with representativeness, was that data from the standardization sample had to be provided in conjunction with the most recent U.S. census data available at the time of the test's development in order to assist comparisons in terms of race/ethnicity, geographic region, parent education level/socioeconomic status, and gender. For most tests reviewed here, that census data would have come from 2000. Results are summarized in column 3 of Table 2. All 10 tests presented their standardization sample data in comparison to census data.

This review did not evaluate how adequately the standardization sample matched the census data due to limited precedent in this area and to the complexity in determining what constitutes adequate group representation. To illustrate, the PPVT–4 presents the representativeness of its standardization sample in relation to 2004 census data for race/ethnicity, parent or examinee education level, and geographic region, each at various age ranges. For example, the PPVT–4 standardized sample for 2- to 5-year-olds was 16.3% African American, 18.8% Hispanic, 60.1% White, and 4.9% Other; whereas the comparable U.S. census population data at that age range was 16.6% African American, 18.2% Hispanic, 59.0% White, and 6.3% Other. Such information does not specify what it means for a test to be representative, but it does allow examiners to make better informed decisions regarding whether or not a test is appropriate for a specific child.

Recency. The third and final criterion for the standardization sample is that it had to have been collected within the last 15 years, consistent with DeThorne and Schaefer's (2004) criterion (see column 4 of Table 2). This criterion was important given the shift in vocabulary knowledge over time and the tendency for images to become dated. All 10 measures had standardization sample data that were collected within the past 15 years. Although the current version of the MCDI–III was published in 2007 and states that the standardization sample has been updated since past editions (pp. 52–53), no year is given for when the updated sample was collected. Thus, because the last edition was published in 2007, we are interpolating that the sample of the 2007 edition meets the recency criterion. Similarly, the MAVA does not explicitly state when the standardization sample was collected, but because it was published in 2008, it is unlikely that the sample was collected more than 15 years ago.

To summarize the results of the tests' standardization samples evaluation, two of the 10 tests fully met all three criteria: the Boehm–3 and the MCDI–III.

Reliability. Turning now from the standardization sample characteristics to test consistency, each test's reliability was evaluated in terms of internal consistency, test–retest reliability, *SEM,* and interexaminer reliability. Based on prior review criteria (Alfonso & Flanagan, 2009; DeThorne & Schaefer, 2004; McCauley & Swisher, 1984), interexaminer reliability and internal consistency were considered acceptable if the correlation coefficients were at or above .90 for each 1-year age group. Although the .90 criterion is often considered the standard for test–retest reliability as well, this standard pertains to contrasts that are not expected to change quickly over time. In contrast, vocabulary, especially in children, changes rapidly. Consequently, we would not expect test–retest reliability values to be as high as other forms of reliability. Additionally, we did not control for time between test administrations; thus, a considerable amount of variability might be expected. Therefore, test–retest reliability coefficients were considered acceptable if they were greater than or equal to .70.

Unlike the other forms of reliability, no clear precedent exists for what the *SEM* should be, given that it is in part derived from the internal consistency estimate. Consequently, an additional cutoff was not set for this value; instead, it was expected that test manuals provided the *SEM* for each normed subgroup so that examiners could easily calculate CIs for the resulting standard scores.

Internal consistency. The first aspect of reliability reviewed was internal consistency, which is a reflection of how reliable a test's scores are associated across items. The MAVA passed the .90 criterion, but the remaining nine tests failed for one of two different reasons: either values fell below the .90 criterion, or data were not presented for each normed subgroup. The Boehm–3 and WORD–2:E fell into the first

category, presenting internal consistency data for each normative group, but the values reported were lower than .90 for at least one subgroup. The WORD–2: E had no value greater than .84, with a range of .69–.80, and the Boehm–3 values ranged from .80 to .91. The WORD–2:A failed both because it did not present data for each 1-year interval, collapsing 16;0–17;11, and because of low values, with no individual subtest value greater than .84, and the lowest value at .72. Similarly, the CREVT–2 did not present data for each 1-year interval (data were presented per year until age 17, but age 18 was included in the collapsed 18–29 age group) and because of low values: 76% (65/86) of the values were above .90, with the lower coefficients ranging from .78 to .89 for our target age range of 18 years or younger. The PPVT–4 and EVT–2 presented high values (all ≥ .90) but failed due to collapsed ages 15 and 16 and 17 and 18. The EOWPVT–4 and ROWPVT–4 also failed because they collapsed ages 13 and 14, 15 and 16, and 17 and 18, although values were all equal to or greater than .90. The MCDI–III provided strong values of .95 through .96 across the different vocabulary scales (i.e., Words and Gestures—Words Understood, Words and Gestures—Words Produced, and Words and Sentences—Words Produced); however, these values were collapsed across normative subgroups.

Test–retest reliability. Only two of the tests reviewed met the .70 criterion for test–retest reliability (the Boehm–3 and WORD–2:E), although all of the tests did present some test–retest data. Seven of the tests failed to meet criterion because they reported reliability coefficients based on collapsed subgroups, which can mask substantial variability: EOWPVT–4, ROWPVT–4, CREVT–2, EVT–2, PPVT–4, MAVA, and WORD–2:A. Though each of these tests presented collapsed data, all presented values that were greater than .70. The MCDI–III provided a description of its test–retest reliability, which suggested that its normative subgroups had been collapsed and that not all values met the .70 standard. Specifically, the MCDI–III manual stated that for the Words and Gestures portion of the measure, vocabulary comprehension correlations were "in the upper .80s" except for the 12-month-old group, for which the correlation was .61 (p. 101). Similarly, correlations for vocabulary production were reportedly "low" in the 8- to 10-month group, and were "in the mid-.80s" for later months. For CDI: Words and Sentences, test–retest correlations were reported "above .90" at each age (p. 101).

SEM. Of the 10 reviewed tests, only one passed the criterion for the presence of *SEM* for each normed subgroup: the Boehm–3. The remaining nine tests failed, either due to a failure to provide *SEM* data at all or due to reporting it in a way that prevented a meaningful comparison to our criterion. With regard to the former, the MAVA did not present *SEM* data in the test manual, though according to the manual (p. 28), it does include software that provides 90% CIs for each administration of the test. Although useful for interpretation of an individual score, an explicit list of *SEM* values by normed subgroup is needed to make a priori decisions about a test's use. Similarly, the MCDI–III manual failed to report *SEM* values, although it did provide standard deviations for each normative group and an explanation of how *SEM* is calculated. The CREVT–2, EOWPVT–4, ROWPVT–4, EVT–2, and PPVT–4 failed to report data for each normative group. The WORD–2:A and WORD–2:E tests reported *SEM* in relation to test–retest reliability values rather than in relation to internal consistency, thereby making the values difficult to compare to our criterion. Consequently, we scored these tests as not meeting the specified criterion.

Interexaminer reliability. No measure met the criterion of equal to or greater than .90 for interexaminer reliability. Of the 10 measures, six, the Boehm–3, EOWPVT–4, ROWPVT–4, EVT–2, PPVT–4, and MCDI–III, did not report interexaminer reliability at all. The CREVT–2 failed to provide values for each normative group, but for the groups reported (data were presented for each subtest for both Form A and Form B), all values were greater than .90. Similarly, the MAVA reported values collapsed across subgroups that exceeded .90; however, it was different in that the values were derived from multiple examiners (three for the Receptive portion and four for the Expressive portion) rather than pairs. The WORD–2:A and WORD–2:E reported interexaminer reliability as percent identical and percent different comparisons, and thus their data were not able to be compared with this study's criterion. However, the percent identical comparisons were high, ranging from 96.4% to 99.8% for the WORD–2:A and from 96.3% to 99.6% for the WORD–2:E.

The results related to the review of all of the reliability evidence are presented in columns 5 through 8 in Table 2. In sum, none of the 10 tests included in this study fully met all of the criteria for reliability. However, one test, the Boehm–3, met two out of the four reliability criteria (test–retest and *SEM*).

Validity. Reliability provides a necessary but insufficient indication of a test's validity. Consequently, additional indices of test score validity are required, although clear quantitative criteria have not been established. Accordingly, we reviewed the test scores for evidence of developmental trends, correlation with similar tests, and group differences, stressing that data must be present to allow test users to make their

own decisions about the adequacy of evidence for individual purposes. Because of the aforementioned trade-off regarding the inclusion of children with LI (Andersson, 2005; DeThorne & Schaefer, 2004; Pena et al., 2006), we did not specify whether or not children with LI should be included in the standardization sample, but instead focused on the need to include key information on sensitivity and specificity as validity evidence. In order for tests to meet criterion, evidence of each of these forms of validity simply had to be provided in the test manuals, with the following additional specifications: For developmental trends, there had to be a discernable increase between the raw scores of each age group (i.e., for each year), no matter how small. For group differences, the tests had to present data on children with an LI compared to typical peers or normative data. Last, for correlations with similar tests, evidence of a moderate to large correlation (≥0.3; Cohen, 1988) with at least one other standardized vocabulary test or vocabulary subtest of a global language measure was required. Finally, sensitivity and specificity were taken into account. Following the criteria set forth by Hargrove (2006), 80% sensitivity and 90% specificity were required in order for the tests in this study to be considered sufficiently sensitive and specific. The results related to a review of all of the tests' validity are presented in columns 9 through 12 in Table 2.

Developmental trends. When looking at the mean scores across age groups, four of the tests (Boehm–3, MCDI–III, MAVA, and WORD–2:E) demonstrated evidence of an increase in raw scores present across all age groups. Only the EVT–2, PPVT–4, EOWPVT–4, ROWPVT–4, and CREVT–2 explicitly discussed the developmental trends as a form of validity evidence. The EOWPVT–4, EVT–2, PPVT–4, ROWPVT–4, and WORD–2:A failed due to collapsed subgroups (for specific information regarding which subgroups were collapsed, see above discussion of size). Our requirement was for tests to demonstrate an increase in raw scores between each year, and with collapsed groups, this could not be determined. The CREVT–2 failed to meet criterion because its scores remained the same between ages 15 to 16 years for the Receptive portion of both Forms A and B, between 13 to 14 years for the Expressive form of Form A, and between 11 to 12 years and 15 to 16 years for Form B.

Test comparison. Although eight tests presented evidence of correlation with other tests purported to measure similar abilities (the WORD–2:A and WORD–2:E did not present test comparison data), only seven met the specified criteria. Specifically, the Boehm–3 failed to meet this criterion because it was only compared to (a) an earlier version of the Boehm, which did not provide external validation evidence, and (b) global achievement tests, with no direct correlations between the Boehm–3 and language test portions reported. Although a large variety of other tests were reported for this form of validity, including measures of IQ, literacy, and academic achievement, all reviewed tests (other than the Boehm–3) included at least one other language-based measure, such as other measures included in this study, as well as global language measures (e.g., CELF–4, PLS–4). However, the types of measures to which the tests were compared, as well as the strengths of their correlations, varied widely.

Group comparisons. Although the tests reviewed provided evidence of group comparisons on a wide variety of populations, including racial and socioeconomic status comparisons, the current study required at least one mean comparison between a TD group and a group with LI. Seven of the tests reviewed passed this criterion: CREVT–2, EOWPVT–4, EVT–2, PPVT–4, ROWPVT–4, WORD–2:A, and WORD–2:E. The Boehm–3 failed because it did not provide any evidence for group comparison; the MAVA failed because although it discussed a field study with students receiving special education services, it did not present the values of the comparison or specify whether the special education group had LI; and the MCDI–III failed because it provided the results of comparisons of groups of differing maternal education and birth order, but not a group with LI.

Sensitivity and specificity. Although it may well be considered the most important information to have when determining the validity of a test's scores, sensitivity and specificity evidence was only presented by one test, the MAVA, with the remaining nine tests failing to pass the specified criterion. The present study follows the criteria of 80% sensitivity and 90% specificity set forth by Hargrove (2006). The MAVA presented extremely high sensitivity and specificity for both –1 and –1.5 *SD* cutoffs for both the Receptive and Expressive subtests, passing these criteria. For the Receptive portion, sensitivity values were 97% and 100% for –1 *SD* and –1.5 *SD* cut-offs, respectively, and specificity was 100% and 85%. Expressive values for sensitivity and specificity were all 100% except for sensitivity at the –1.5 *SD* cutoff, which was 83%. However, these values were derived from collapsed age groups, which is a limitation.

To summarize, none of the 10 tests analyzed in this study passed all of the validity criteria. However, the MAVA did emerge as the strongest test in the realm of validity evidence, passing three of the four validity criteria.

DISCUSSION

We evaluated 10 commonly used standardized vocabulary tests on the basis of their standardization sample, reliability, and validity evidence. In terms of the standardization sample, most of the tests came reasonably close to meeting the criteria. Specifically, all 10 tests passed in terms of representativeness and recency, suggesting that current test developers are, at the very least, attempting to have their standardization samples in proportion with the current population. Eight tests failed to meet the criteria of at least 100 individuals per normative subgroup. However, most either failed only at one or two subgroups and were usually very close to having an adequate number or failed due to collapsed age groups, usually combining 2 years at one or two ages. Evidence of reliability and validity were less encouraging. Specifically, none of the 10 tests passed all of the reliability criteria, although one, the Boehm–3, passed two of the four criteria. Evidence of test–retest and interexaminer reliability were particular areas of need. With regard to validity, one test, the MAVA, met three of the four designated criteria. Additionally, only the MAVA reported sensitivity and specificity data, which is arguably one of the most informative pieces of validity evidence that a test can provide, at least for the purpose of diagnostic accuracy. Given these results, the remainder of the discussion will highlight recommendations, both for clinical practice and for test development, as well as limitations of our review.

Recommended Measures

In terms of the sheer number of criteria met, the Boehm–3 and the MAVA met the highest number of criteria, meeting seven and six, respectively, out of 11 total. That said, the selection of individual measures should always be determined based on the complex needs of the case. For example, the MCDI–III provides a particularly useful option for gaining input from caregivers. Also, the WORD–2:A and WORD–2:E are unique in offering a multidimensional assessment of vocabulary, which could provide a more comprehensive view of a child's vocabulary. Whatever test is chosen, the examiner should always have a clear sense of the psychometric strengths and weaknesses of the tests they are using (American Psychological Association, 2010; American Speech-Language-Hearing Association, 2010).

Suggestions for Test Development

The results of this study suggest several considerations for test development. Based on the criteria used in this review, it is clear that stronger reliability values (≥.90, or ≥.70 for test–retest), as well more consistent methods of measuring and reporting reliability data, particularly in test–retest and interexaminer reliability, are common areas in need of improvement. Another clear area of improvement would be including data regarding diagnostic sensitivity and specificity; such data were presented in only one of the 10 measures reviewed. As these are important and useful measures of how well a test score can discriminate between TD individuals and those with LI, sensitivity and specificity data provide valuable information. However, this form of evidence is also limited by circularity, as the original designation of TD or LI used to derive sensitivity and specificity values was likely determined by other standardized assessments.

Finally, it seems worth noting that focusing on a narrower age range might lead to stronger psychometric properties. For example, the Boehm–3, which met the highest number of criteria, was developed for children in kindergarten through second grade. Tests that were developed for the extensive age range of 2;6 to 81+ years, such as the PPVT–4 and EVT–2, did not fare as well. Logistically, it stands to reason that it would be easier to develop and standardize a test focused on a 3-year age range rather than one that aims to be appropriate across the life span. Although tests developed for use with a wide age range are economical for clinicians and are widely marketable, it may not be realistic to expect one test to be able to reliably and validly assess individuals of a vast array of age.

Limitations in Our Criteria

Although this evaluation did discriminate between tests, limitations are inherent in the criteria used. First, cutoff values, such as those employed for reliability criteria, inherently create an arbitrary dichotomy out of a continuous variable. For example, the difference between a reliability value of .89 and .90 is negligible. However, the dichotomous pass/fail distinction was consistent with prior literature (Andersson, 2005; DeThorne & Schaefer, 2004; McCauley & Swisher, 1984) and was considered a useful way to simplify a large amount of complex data. That said, we have incorporated information within the text regarding how far values fell from the cutoff value so that readers can make informed judgments for their individual purposes.

A second limitation that should be mentioned here relates to the criteria for validity, which were qualitative rather than quantitative in nature. In other words, the criteria specified what information should

be provided but provided less specification on what the data should look like. For example, data regarding group differences were required without specification of how much of a group difference between children with and without LI was an adequate indication of validity. Though limited, our review approach was generally consistent with prior standards for validity evidence (Andersson, 2005; DeThorne & Schaefer, 2004; McCauley & Swisher, 1984) as specifics for validity are contingent in part on the purpose of the assessment and the characteristics of the child being considered. However, the issue of validity in standardized testing remains an important topic in need of examination and critique and thus, the currently used forms of validity evidence may be limited (Sattler, 2001).

Conclusion

Although no single test reviewed in this study met all of the psychometric criteria, most showed positive evidence in multiple areas. In particular, the Boehm–3 and MAVA met the highest number of criteria. Although tests were evaluated in this study for general purposes, in truth, the appropriateness of any given measure should be considered on a case-by-case basis. The strength of individual tests will vary with a number of factors, including child age, specific form of potential impairment, and purpose of the assessment. Regardless of which standardized tests are employed, best practice in assessment is contingent on integrating multiple forms of assessment, incorporating both parent report and observational measures (Watkins & DeThorne, 2000). Single test scores should never be used in isolation to make high-stakes decisions. Tests can only measure performance on 1 day, under specific circumstances, and may not represent an individual's strengths and weaknesses accurately. On a related note, standardized tests are intended to reflect an individual's intrinsic skills and abilities, whereas other forms of assessment allow us to get a more functional view of how successfully vocabulary skills are being deployed and taken up in everyday contexts.

REFERENCES

Alfonso, V. C., & Flanagan, D. P. (2009). Assessment of preschool children. In B. A. Mowder, F. Rubinson, & A. E. Yasik (Eds.), *Evidence-based practice in infant and early childhood psychology* (pp. 129–166). Hoboken, NJ: Wiley & Sons.

American Academy of Pediatrics. (2003). Family-centered care and the pediatrician's role. *Pediatrics, 112*, 691–696.

American Psychological Association. (2010). *Ethical principles of psychologists and code of conduct*. Available from www.apa.org/ethics/code/

American Speech-Language-Hearing Association. (2010). *Code of ethics*. Available from www.asha.org/policy-ET2010-00309/

Andersson, L. (2005). Determining the adequacy of tests of children's language. *Communication Disorders Quarterly, 26*(4), 207–225.

Boehm, A. E. (2000). *Boehm Test of Basic Concepts, Third Edition*. San Antonio, TX: The Psychological Corporation.

Boehm, A. E. (2001). *Boehm Test of Basic Concepts—Preschool*. San Antonio, TX: The Psychological Corporation.

Bogue, E., & DeThorne, L. (2012, February). *A psychometric analysis of childhood vocabulary tests*. Poster presented at the 52nd annual Illinois Speech-Language-Hearing Association, Rosemont, IL.

Bowers, L., Huisingh, R., LoGuidice, C., & Orman. J. (2004). *The WORD Test—2, Elementary*. East Moline, IL: LinguiSystems.

Bowers, L., Huisingh, R., LoGuidice, C., & Orman. J. (2005). *The WORD Test—2, Adolescent*. East Moline, IL: LinguiSystems.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

DeThorne, L. S., & Schaefer, B. A. (2004). A guide to child nonverbal IQ measures. *American Journal of Speech-Language Pathology, 13*, 275–290.

Dinnebeil, L. A., & Rule, S. (1994). Congruence between parents' and professionals' judgments about the development of young children with disabilities: A review of the literature. *Topics in Early Childhood Special Education, 14*, 1–25.

Dunn, L. M. (1965). *Peabody Picture Vocabulary Test*. Circle Pines, MN: AGS.

Dunn, L. M., & Dunn, L. M. (1982). *Peabody Picture Vocabulary Test—Revised*. Circle Pines, MN: AGS.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test, Third Edition*. Circle Pines, MN: AGS.

Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test, Fourth Edition*. San Antonio, TX: Pearson.

Eickhoff, J., Betz, S. K., & Ristow, J. (2010, June). *Clinical procedures used by speech language pathologists to diagnose SLI*. Poster session presented at the Symposium on Research in Child Language Disorders, Madison, WI.

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *MacArthur-Bates Communicative Development Inventory, Third Edition*. Baltimore, MD: Brookes.

Fleege, P. O., Charlesworth, R., Burts, D. C., & Hart, C. H. (1992). Stress begins in kindergarten: A look at behavior during standardized testing. *Journal of Research in Childhood Education, 7*, 20–26.

Flynn, J. R. (1999). Searching for justice: The discovery of IQ gains over time. *American Psychologist, 54*, 5–20.

Fuchs, D., & Fuchs, L. S. (1986). Test procedure bias: A meta-analysis of examiner familiarity effects. *Review of Educational Research, 56*, 243–262.

Gardner, M. F. (1979). *Expressive One-Word Picture Vocabulary Test*. East Aurora, NY: Slosson Educational Publications.

Gardner, M. F. (1985). *Receptive One-Word Picture Vocabulary Test*. Novato, CA: Academic Therapy Publications.

Gardner, M. F. (1990). *Expressive One-Word Picture Vocabulary Test—Revised*. Novato, CA: Academic Therapy Publications.

Gorsuch, R. L. (1983). *Factor analysis*. Hillsdale, NJ: Erlbaum.

Gorsuch, R. L. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of Personality Assessment, 68*(3), 532–560.

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

Hammill, D. D., & Newcomer, P. L. (2008). *Test of Language Development—Intermediate, Fourth Edition*. San Antonio, TX: Pearson.

Hargrove, P. (2006). EBP tutorial #10: EBP metrics for assessment. *Language Learning and Education, 13*, 23–24.

Huisingh, R., Barrett, M., Zachman, L., Blagden, C., & Orman, J. (1990). *The Word Test, Elementary—Revised*. East Moline, IL: LinguiSystems.

Martin, N. A., & Brownell, R. (2011a). *Expressive One-Word Picture Vocabulary Test, Fourth Edition*. Novato, CA: Academic Therapy Publications.

Martin, N. A., & Brownell, R. (2011b). *Receptive One-Word Picture Vocabulary Test, Fourth Edition*. Novato, CA: Academic Therapy Publications.

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

McCollum, J., & Yates, T. J. (1994). Dyad as focus, triad as means: A family-centered approach to supporting parent–child interactions. *Infants and Young Children, 6*(4), 54–63.

Merrell, A. W., & Plante, E. (1997). Norm-referenced test interpretation in the diagnostic process. *Language, Speech, and Hearing Services in Schools, 28*, 50–58.

Messick, S. (2000). Consequences of test interpretation and use: The fusion of validity and values in psychological assessment. In R. D. Goffin & E. Helmes (Eds.), *Problems and solution in human assessment* (pp. 3–19). Boston, MA: Kluwer Academic.

Montgomery, J. K. (2008). *Montgomery Assessment of Vocabulary Acquisition*. Greenville, SC: Super Duper Publications.

Oliver, B., Dale, P. S., Saudino, K., Petrill, S. A., Pike, A., & Plomin, R. (2002). The validity of parent-based assessment of non-verbal cognitive abilities of three-year olds. *Early Child Development and Care, 172*, 337–348.

Pena, E. D., Spaulding, T. J., & Plante, E. (2006). The composition of normative groups and diagnostic decision making: Shooting ourselves in the foot. *American Journal of Speech-Language Pathology, 15*, 247–254.

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

Plante, E., & Vance, R. (1995). Diagnostic accuracy of two tests of preschool language. *American Journal of Speech-Language Pathology, 4*(2), 70–76.

Prelock, P. A. (2006). Working with families and teams to address the needs of children with mental retardation and developmental disabilities. *Perspectives on Language Learning and Education, 13*, 7–11.

Sattler, J. M. (2001). *Assessment of children: Cognitive applications* (4th ed). San Diego, CA: Author.

Saudino, K. J., Dale, P. S., Oliver, B., Petrill, S. A., Richardson, V., Rutter, M., … Plomin, R. (1998). The validity of parent-based assessment of the cognitive abilities of 2-year-olds. *British Journal of Developmental Psychology, 16,* 349–363.

Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical Evaluation of Language Fundamentals, Fourth Edition*. San Antonio, TX: Pearson.

Spaulding, T. J., Plante, E., & Farinella, K. A. (2006). Eligibility criteria for language impairment: Is the low end of normal always appropriate? *Language, Speech, and Hearing Services in Schools, 37*, 61–72.

Speltz, M. L., DeKlyen, M., Calderon, R., Greenberg, M. T., & Fisher, P. A. (1999). Neuropsychological characteristics and test behavior in boys with early onset conduct problems. *Journal of Abnormal Psychology, 108*, 315–325.

Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data*. Hillsdale, NJ: Erlbaum.

Thies, K. M., & McAllister, J. W. (2001). The health and education leadership project: A school initiative for children and adolescents with chronic health conditions. *Journal of School Health, 71*, 167–172.

Wallace, G., & Hammil, D. D. (2002). *Comprehensive Receptive and Expressive Vocabulary Test, Second Edition*. Austin, TX: Pro-Ed.

Washington, J. A. (1996). Issues in assessing the language abilities of African American children. In A. G. Kamhi, K. E. Pollock, & J. L. Harris (Eds.), *Communication development and disorders in African American children: Research, assessment, and intervention* (pp. 35–54). Baltimore, MD: Brookes

Watkins, R. V., & DeThorne, L. S. (2000). Assessing children's vocabulary skills: From word knowledge to word-learning potential. *Seminars in Speech and Language, 21*(3), 235–245.

Wiig, E. H., & Secord, W. (1992). *Test of Word Knowledge*. San Antonio, TX: The Psychological Corporation.

Williams, K. T. (1997). *Expressive Vocabulary Test*. Circle Pines, MN: AGS.

Williams, K. T. (2007). *Expressive Vocabulary Test, Second Edition*. San Antonio, TX: Pearson.

Zachman, L., Huisingh, R., Barrett, M., Orman, J., & Blagden, C. (1989). *The Word Test—Adolescent*. East Moline, IL: LinguiSystems.

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2011). *Preschool Language Scale, Fifth Edition*. San Antonio, TX: Pearson.

Contact author: Ellen L. Bogue, Blank Children's Hospital, Pediatric Therapy, 1200 Pleasant Court, Des Moines, IA 50309. E-mail: ellen.bogue@unitypoint.org.