LSHSS

Eligibility Criteria for Language Impairment: Is the Low End of Normal Always Appropriate?

Tammie J. Spaulding
Elena Plante
Kimberly A. Farinella
The University of Arizona, Tucson

The problem of who should be identified as having a language impairment is of central importance to both clinical practice and the research enterprise within the field of speech-language pathology. Standardized testing, although not the only criterion, is one factor that is used in the diagnosis of child language impairment (Wilson, Blackmon, Hall, & Elcholtz, 1991). It is commonly assumed that children with language impairments can be identified because they will obtain low scores on tests of language. Indeed, school systems support this practice, frequently requiring children to score at the low end of a test's normative distribution to qualify for services. District eligibility criteria in multiple states include that the child will obtain a low test score on one or more language tests (e.g., –1.5 *SD* in Missouri and South Dakota, –1.75 *SD* in Wisconsin, –1.5 to –2.0 *SD* in New York and Arizona, –2.0 *SD* in Kentucky). However, applying such low cutoff scores for diagnosing the presence or absence of language impairment assumes that children with language impairments routinely obtain low scores on any of the many available language tests, whereas their typically developing peers will typically obtain higher scores. The variability of score cutoffs across districts clearly illustrates the arbitrary nature of these criteria for diagnosing child language impairments.

ABSTRACT: **Purpose:** The assumption that children with language impairment will receive low scores on standardized tests, and therefore that low scores will accurately identify these children, is examined through a review of data in the manuals of tests that are intended for use in identifying such children.
Method: Data from 43 commercially available tests of child language were compiled to identify whether evidence exists to support their use in identifying language impairment in children.
Results: A review of data concerning the performance of children with impaired language failed to support the assumption that these children will routinely score at the low end of a test's normative distribution. A majority of tests reported that such children scored above 1.5 *SD* below the mean, and scores were within 1 *SD* of the mean for more than a quarter (27%) of the tests. The primary evidence needed to support the purpose of identification, test sensitivity and specificity, was available for 9 of the 43 tests, and acceptable accuracy (80% or better) was reported for 5 of these tests.
Implications: Specific data supporting the application of ''low score'' criteria for the identification of language impairment is not supported by the majority of current commercially available tests. However, alternate sources of data (sensitivity and specificity rates) that support accurate identification are available for a subset of the available tests.

KEY WORDS: language disorder, assessment, evidence-based practice, school-age children

The assumption that language impairment will be identified by a low test score is also found in the subject selection criteria for research on specific language impairment (SLI). However, researchers appear to use a somewhat more relaxed criteria for cutoff scores. A review of recent articles published (August, 2003–April, 2004) in journals by the American Speech-Language-Hearing Association (ASHA)[1] suggests that a majority of researchers select participants with SLI based on a language score set anywhere between 1 *SD* below the normative mean (Flax et al., 2003; Ford & Milosky, 2003; Paradis, Crago, Genesee, & Rice, 2003) and 1.5 *SD* below the mean (Dollaghan, 2004; Gray, 2003; Maillart, Schelstraete, & Hupet, 2004; Leonard et al., 2003; Wells

[1] The review included all articles that specified the selection of participants with SLI (of any age) that were published in the *Journal of Speech, Language, and Hearing Research, American Journal of Speech-Language Pathology,* and *Language, Speech, and Hearing Services in Schools.*

0161-1461/06/3701-0061

& Peppé, 2003) on one or more tests of language. This low score assumption, which underlies such applications of low cutoff scores, is further evident in theoretical positions that have been expressed concerning the nature of SLI as the low end of the normal continuum (e.g., Leonard, 1991). This position has also been applied to practical concerns. For example, some researchers have advocated including children with impaired language in the normative groups of standardized tests so that the low end of the normative range is not truncated (McFadden, 1996). This clearly presupposes that children with language impairments will score at the lower end of the normal range.

A further problem with the application of an arbitrary cutoff score is that it is applied to any test that the researcher or clinician should choose, without reference to how children actually score on the tests selected for use. For example, one might question, when applying a cutoff score of –1.5 *SD* to identify children with language impairments, whether the test being used for this purpose provides evidence that such children are likely to obtain scores this low or lower. A mismatch between the cutoff score criterion and the actual typical scores associated with impairment on the test chosen for use could lead to systematic under- or overidentification (Plante & Vance, 1994). For example, this method of identification is typically employed regardless of language domain assessed. Although multiple language areas may be affected in children who exhibit language impairments, these children typically are not uniformly impaired across the language modalities. For example, a child with SLI is likely to exhibit deficits in the areas of morphology and syntax (Leonard, 1998; Rice, 1994). In contrast, tests of single-word vocabulary do not yield strong identification of children with SLI who are broadly selected, without particular reference to their vocabulary skills (Gray, Plante, Vance, & Henrichsen, 1999), despite evidence of broader semantic deficits (Brackenbury & Pye, 2005).

This pattern of differential impairment of children with SLI strongly suggests that tests designed to reflect models of typical language skills (e.g., lexicon size, English morphology) are less likely to be effective identifiers than are tests that target language skills known to be impaired in children with SLI. For example, tests that assess morphology typically contain items that represent the breadth of English morphology rather than concentrate on those morphological items that are associated with errors in SLI. Therefore, such tests may contain a few items that children with SLI routinely fail intermixed with many that they routinely pass. This problem has led to the suggestion that a more effective means of identifying such children would be to test the particular language features that are difficult for children with SLI (Rice & Wexler, 2001). For example, Rice and Wexler (1996) suggested that difficulty with certain aspects of verb morphology may constitute a *clinical marker* for a language impairment and that evaluating performance on these specific language targets can improve the identification of SLI (Rice & Wexler, 2001).

We can conclude that the practice of identifying language impairment by application of a low but arbitrary cutoff score to any of a number of commercially available tests is relatively common in both clinical practice and research. What is unknown is the extent to which existing evidence can provide empirical support for this practice. The absence of any data-based evaluation for the use of a low cutoff score across tests is contradictory to the standards for evidence-based practice. Briefly, an evidence-based practice framework mandates that clinicians evaluate the presence and strength of data relevant to the clinical procedures and measures that they intend to use. Within an evidence-based practice frame of reference, a clinician should evaluate the extent to which a particular cutoff score is actually effective for distinguishing normal from impaired language on the particular test that he or she has selected for use.

Application of a low, arbitrary cutoff score for diagnosing the presence or absence of language impairment is not the only method available for identifying children with impaired language. Researchers have begun to advocate abandoning the use of arbitrary cutoff score criteria applied across multiple tests in favor of adopting evidence-based cutoff scores derived for each particular test (e.g., Dollaghan, 2004; Gray et al., 1999; Merrell & Plante, 1997; Perona, Plante, & Vance, in press; Plante & Vance, 1994, 1995; Rescorla, 1989; Rescorla & Alley, 2001; Rice & Wexler, 2001). The empirically derived cutoff score for a particular test would reflect the highest levels of sensitivity (i.e., percentage of children with language impairment who are diagnosed as impaired) and specificity (i.e., percentage of children with typical language skills who are diagnosed as typical). A test's sensitivity and specificity data provide the clinician with direct evidence of its ability to differentiate children with language impairment from those with typically developing language skills.

The purpose of this study was two-fold. First, we were interested in whether data are provided within currently available norm-referenced test manuals to support the idea that children with language impairments are likely to obtain low scores relative to their typically developing peers. This assumption underlies the practice of applying a low cutoff score for the identification of language impairments across a variety of tests. To determine this, we report the relative magnitude of the group differences (i.e., the effect size δ) between the language impaired and typically developing samples from data reported within the test manuals. If children with language impairments consistently score at the low end of the normal distribution, then the difference between the sample with language impairment and the matched typically developing sample or normative sample should reflect this. If this is the case, a low cutoff score criterion applied across tests will be sufficient for identifying language impairment. However, if score differences are frequently small, this would be evidence against applying a low cutoff score for diagnosing language impairments. This, in turn, would have important implications for how children are qualified for services or selected for research when using the currently available commercial tests.

Ideally, however, identification accuracy is not measured by mean group differences but rather by sensitivity and specificity data. Under an evidence-based practice framework, these two metrics are the primary evidence needed by clinicians to support the use of a test for the identification of language impairments. Although this information was virtually absent in 1994 when Plante and Vance first recommended reliance on sensitivity and specificity for the identification of language impairment, many tests have since been introduced or updated. Therefore, our second purpose was to determine how many of the tests selected provide information on sensitivity and specificity within their test manuals. We will report the sensitivity and specificity rates along with the cutoff scores that the test authors recommend for determining the presence or absence of language impairment.

METHOD

Material

This review includes the latest edition of 43 commercially available norm-referenced standardized tests. These tests were identified by reviewing current vendor catalogs for norm-referenced language tests for use with children ages 3 to 18 years. Test manuals not already owned by the authors were purchased for review. Those selected for study included ones that, as advertised, claimed to test English language skills for the purpose of identifying childhood language impairments. Tests selected for review were not restricted by language domain but did not include those targeting primarily academic skills. Tests that used interview methods or observations of spontaneous behavior in favor of those that scored elicited responses from the child were also excluded. Tests that indicated that they should be used primarily as screening measures or criterion-referenced measures were generally excluded from review. However, the Diagnostic Evaluation of Language Variation (DELV; Seymour, Roeper, & deVilliers, 2003), a criterion-referenced measure, as well as the Structured Photographic Expressive Language Test—Preschool (SPELT–P; Werner & Kresheck, 1983) and The Renfrew Bus Story (Cowley & Glasgow, 1994), two screening measures, were included. These were included specifically because their manuals indicated that they could be used for the purpose of identifying language impairment in children. Technical information about each test's construct validity relevant to the purposes of this study was collected from the technical data provided in each test manual. Although this information is sometimes available in the peer-reviewed literature as well, we confined our analysis to information available in the test manual because this is the primary source available to clinicians.

Procedures

Three clinically certified speech-language pathologists (SLPs) reviewed the tests. All reviewers had clinical experience with test administration and had taken advanced statistical coursework during their graduate training.[2] Each author examined a subset of the 43 tests. Data collection was performed independently by one of the three reviewers for most tests, with two or more of the reviewers examining sections of manuals that contained potential ambiguities. Data from six of the 43 tests were collected by two of the SLPs for reliability purposes. Data from these six tests (14% of the total number of tests reviewed) were collected and calculated by two of the authors to obtain an estimate of interexaminer reliability. Comparison of the data revealed 100% interexaminer agreement calculated as follows: percentage of agreement/(agreement + disagreement). This is not surprising given that the manuals either contained the information and it was recorded properly or they did not.

Consistent with our first purpose, data relevant to determining the magnitude of differences between language-impaired and matched, typically developing groups (or normative samples) were recorded. These data included the mean differences in subtest scores, test composite scores, and/or total test scores for children with language impairment and typical samples. To calculate the differences in group performance, we subtracted the mean of the language-impaired group from the mean of the control group and divided this by the larger of the two group standard deviations. Using the larger of the standard deviations is considered a more conservative approach for determining the magnitude of the mean difference, and it limited the impact of ceiling effects that occurred for the control or normative sample with some tests. The resulting metric reflects the mean group difference in units of standard deviation (i.e., the effect size δ).

In many cases, a typical language or control sample was provided for these comparisons. In these cases, the means and standard deviations of this group were recorded. When a typical sample was not provided, we assumed the normative mean (100 or 10) and standard deviation (15 or 3) as the basis against which the language-impaired sample was compared. If alternate forms of the same test were published, we averaged the results for both versions so that the resulting distribution would reflect a single value for each test. When total test score means and standard deviations were provided for more than one sample (e.g., different age groups), we calculated the average mean and standard deviation weighted for the number of subjects in each age group.

Included in our review of language-impaired samples were groups of children who had been identified as having language delay, language impairment, language disorder, or SLI. Information on how these children were originally identified was largely absent or extremely general (e.g., previously identified by school systems). However, given epidemiologic data that indicate that the rate of SLI in the population is relatively low (7.4%), and that most cases (71%) of SLI go unidentified clinically (Tomblin et al., 1997), it is not likely that these language-impaired samples contained large numbers of cases of either mild language problems or typical language misidentified as an impairment.

Consistent with our second purpose, we determined the frequency with which test manuals presented information on sensitivity and specificity, the sensitivity and specificity rates provided, and the cutoff score used to derive the sensitivity and specificity data. If this information was not explicitly stated, but sufficient data to easily calculate it were presented, we determined the sensitivity and specificity and counted this information as present. For example, if a test provided a table of correctly and incorrectly identified children but did not calculate sensitivity and specificity from these data, we calculated sensitivity and specificity by hand (see Figure 1 for how this is calculated) and counted this information as present. If the test reported sensitivity and specificity data in association with more than one cutoff score, we selected the cutoff score that represented the best balance between sensitivity and specificity (i.e., minimized the difference between these two values). Two tests presented information in a way that required judgment by the examiners in terms of estimating this information. The DELV (Seymour et al., 2003) presented identification accuracy for four qualitatively labeled levels of performance. To calculate sensitivity and specificity for this test, the groups categorized as showing "weakness" and "low average" skills were used to indicate impairment, and the "average" and "strength" categories were used to indicate typically developing abilities. This was done because it maximized

[2] Although analysis of the statistical data related to calculating mean group differences involved basic skills (subtraction, division), identification of sensitivity and specificity data can require knowledge of statistics to recognize these data and how they were derived.

Figure 1. Calculations for test sensitivity and specificity.

a) Children with language impairment correctly identified as impaired	c) Normal children incorrectly identified as language impaired
b) Children with impaired language incorrectly identified as normal	d) Normal children correctly identified as normal

Sensitivity= a/(a+b)
Specificity= d/(c+d)

correct identification of typically developing children and children with impairment. The Patterned Elicitation Syntax Test (PEST; Young & Perachio, 1993) recommended use of a cutoff of the lower 10th percentile (–1.28 *SD*) to identify impairment, and sensitivity was calculated based on scores reported for individual children with language impairment provided in the manual. No information was available to calculate specificity in a similar manner, but it was assumed to be 90% based on a 10th percentile cutoff. Finally, the overall sensitivity and specificity of the Test of Early Grammatical Impairment (TEGI; Rice & Wexler, 2001) was calculated using a weighted average across ages for purposes of display.

Additional information collected included the gender representation, the percentage of minorities, and the socioeconomic status of children included in these group studies in order to assess whether clinicians were provided information that would allow them to judge how representative the test groups were to their local population.

RESULTS

Mean Group Differences

The mean group difference reflects the average difference between the scores of children with language impairment and those of their typically developing peers. Therefore, it is a useful way to reflect how discrepant scores for children with language impairments are likely to be. Note, however, that only 50% of the children with language impairment will fall below the mean group difference.

None of the test manuals indicated mean group differences between the language-impaired and typically developing or normative sample. However, information was available within 33 of the 43 test manuals to calculate the mean group difference. Specifically, means and standard deviations for the performance of groups identified as having impaired language were available for 33 tests. The results for these tests are listed in Table 1.

Magnitude of mean difference for total language scores. Figure 2 presents the relative performance difference of the language-impaired sample as compared to typical or normative samples based on total test scores. The distribution of mean group differences for the 33 tests that reported this information was approximately normally distributed[3] around a mean of 1.34 and a standard deviation of .47. Nine of the 33 tests reported mean score differences that were within 1 *SD* of each other. Another 14 tests reported average group differences of between 1.0 and 1.5 *SD*. Only 10 of the 33 tests reported score differences greater than 1.5 *SD*. Note again that because 1.5 *SD* is the mean group difference, only 50% of all children with language impairment would fall below a typical clinical cutoff score of –1.5 *SD* on these 10 tests. The list of tests corresponding to each of these categories is provided in Table 2.

Note that measurement error, which affects score stability at the level of the individual child, is not likely to have affected the distribution we obtained here or subsequent group analyses. This is because measurement error distributes normally. This means that we could expect that the instances for which scores were over- or underestimated across children would balance out within a sample, leaving the mean for each group (and therefore, the mean difference) unchanged.

Magnitude of mean differences by language modality and domain. We were interested in determining whether the magnitude of mean score differences varied as a function of the type of language skills assessed. The distribution of mean group differences for total scores for expressive and receptive tests is provided in Figure 3. Tests that did not fit cleanly in either category were excluded from this analysis. Of the two domains, expressive scores (M = 1.37, SD = .51) tended to be only slightly higher than receptive scores (M = 1.24, SD =.36), with substantial overlap between the distributions. Likewise, we broke out test score differences by language domain. Figure 4 presents the mean group difference for tests of single-word vocabulary (receptive and expressive), tests of broader semantic abilities (i.e., tests targeting any aspect of semantic knowledge rather than lexical labels), and tests measuring skills in the morphosyntactic domain. Again, mean group differences were similar and tended to be only slightly higher for expressive (M = .90, SD = .26) than receptive (M = .81, SD = .26) single-word vocabulary tests. Tests and subtests that tapped more general semantic abilities showed somewhat more robust score differences (M = 1.21, SD = .36) than those seen for the single-word vocabulary tests. Finally, tests and subtests for morphology and syntax showed the largest mean differences (M = 1.28, SD = .42).

Magnitude of mean difference by age. Twelve tests provided additional information, including means and standard deviations across non-overlapping age groups. The magnitude of mean differences for tests broken down by age is reported in Table 3. These data suggest that there is notable variability in terms of mean score differences by age within tests of child language. However, there is no predictable trend (e.g., increasing or decreasing differences with age) that applies across tests.

Sensitivity and Specificity

Sensitivity and specificity information was provided in the manuals of nine out of a total of 43 tests examined. These data are presented in Figure 5. Compared against the 80% criterion suggested by Plante and Vance (1994), the Clinical Evaluation of

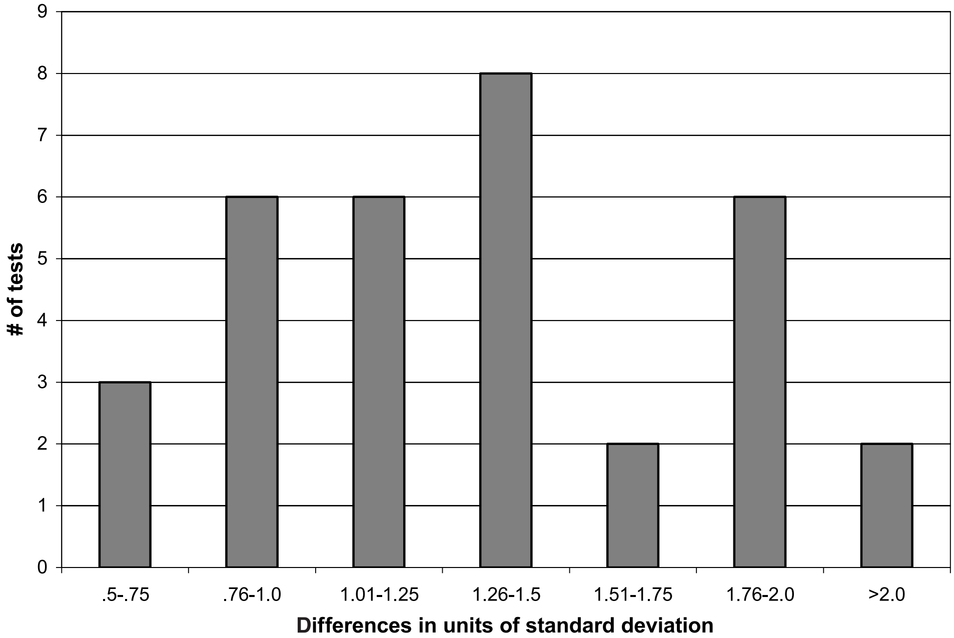
[3] Note that this distribution of means corresponds to the distribution predicted by the theorem of central limits. This statistical principle predicts that the means of samples that are drawn from a single larger population will distribute normally if the population from which they are drawn is also distributed normally.

Table 1. Mean group differences for normal and language-impaired groups.

		Mean difference for normal and language impaired samples					
		Scale scores		*Subtest types*			
Test	*Total score*	*Expressive*	*Receptive*	*Single-Word Vocabulary: Expressive*	*Single-Word Vocabulary: Receptive*	*Semantic*	*Morphosyntax*
ALL	.759						
BBCS–R	n/a						
BLT–2	1.360	1.360				1.033^{s}	1.433^{s}
BOEHM–3	n/a						
BOEHM–P3	1.168					1.168	
CASL	.881	0.579^{c}	0.741^{c}			1.250^{c}	1.377^{c}
CELF–4	2.487	2.430^{c}	1.890^{c}	1.212^{s}		1.787^{c}	2.318^{c}
CELF–P	1.448	1.651^{c}	1.000^{c}	0.966^{s}		0.824^{s}	1.286^{ms}
CREVT–2	0.600			0.533^{s}	0.433^{s}		
DELV	0.951					0.931^{c}	0.798^{c}
ELT	1.822	1.822				1.096^{ms}	1.403^{s}
EOWPVT	n/a						
EVT	0.524			0.524			
FLT–A2	0.837					0.874^{ms}	0.892^{ms}
LPT–R	1.457					1.457	
OWLS–OC	1.234	1.277^{c}	0.998^{c}				
OWLS–WE	1.699						
PEST	n/a						
PLAI–2	1.267	1.000^{c}	1.000^{c}			0.667^{ms}	
PLS–4	1.916	1.728^{c}	1.807^{c}				
PPVT–III	0.549				0.549		
ROWPVT	n/a						
SPELT–3	n/a						
SPELT–P	n/a						
TACL–3	1.200		1.200		0.667^{s}		1.000^{ms}
TEEM	n/a						
TELD–3	1.433	1.133^{c}	1.233^{c}				
TEGI	1.768						1.768
THT	1.791					1.384^{ms}	1.112^{ms}
TLC–E	1.562						
TLT–R	1.453		1.453			1.196^{ms}	
TNL	2.000						
TOLD–I3	0.933	1.000^{c}	1.000^{c}		1.000^{s}	1.067^{c}	0.933^{c}
TOLD–P3	1.400	1.133^{c}	1.267^{c}	1.000^{s}	1.000^{s}	1.267^{c}	1.400^{c}
TOPS–R	1.254						
TOSS–P	1.296			1.022^{s}	0.948^{s}	1.296	
TOWK	1.159			1.011^{s}	1.190^{s}	1.159	
TOWL–3	1.133						
TRBS	n/a						
TTC	n/a						
TWT–A	2.023					2.023	
TWT–R	1.772					1.772	
UTLD–4	0.933				0.667^{s}	0.800^{c}	0.933

Note. ALL = Analysis of the Language of Learning (Blodgett & Cooper, 1987), BBCS–R = Bracken Basic Concept Scale—Revised (Bracken, 1998), BLT–2 = Bankson Language Test—Second Edition (Bankson, 1990), BOEHM–3 = Boehm Test of Basic Concepts—Third Edition (Boehm, 2000), BOEHM–P3 = Boehm Test of Basic Concepts—Preschool (Boehm, 2001), CASL = Comprehensive Assessment of Spoken Language (Carrow-Woolfolk, 1999a), CELF–4 = Clinical Evaluation of Language Fundamentals—Fourth Edition (Semel, Wiig, & Secord, 2003), CELF–P = Clinical Evaluation of Language Fundamentals—Preschool (Wiig, Secord, & Semel, 1992), CREVT–2 = Comprehensive Receptive and Expressive Vocabulary Test—Second Edition (Wallace & Hammill, 2002), DELV = Diagnostic Evaluation of Language Variation (Seymour, Roeper, & de Villiers, 2003), ELT = The Expressive Language Test (Huisingh, Bowers, LoGuidice, & Orman, 1998), EOWPVT = Expressive One-Word Picture Vocabulary Test—Revised (Gardener, 1990), EVT = Expressive Vocabulary Test (Williams, 1997), FLT–AT = The Fullerton Language Test for Adolescents—Second Edition (Thorum, 1986), LPT–R = The Language Processing Test—Revised (Richard & Hanner, 1995), OWLS = OWLS Listening Comprehension and Oral Expression Scale (Carrow-Woolfolk, 1995), OWLS-WE = OWLS Written Expression Scale (Carrow-Woolfolk, 1996), PEST = Patterned Elicitation Syntax Test (Young & Perachio, 1993), PLAI–2 = Preschool Language Assessment Instrument—Second Edition (Blank, Rose, & Berlin, 2003), PLS–4 = Preschool Language Scales—Fourth Edition (Zimmerman, Steiner, & Pond, 2002), PPVT–III = Peabody Picture Vocabulary Test—Third Edition (Dunn & Dunn, 1997), TEGI = Test of Early Grammatical Impairment (Rice & Wexler, 2001), ROWPVT = Receptive One-Word Picture Vocabulary Test (Gardener, 1985), SPELT–3 = The Structured Photographic Expressive Language Test—Third Edition (Dawson, Stout, & Eyer, 2003), SPELT–P = Structured Photographic Language Test—Preschool (Werner & Kresheck, 1983), TACL–3 = Test of Auditory Comprehension of Language—Third Edition (Carrow-Woolfolk, 1999b), TEEM = Test for Examining Expressive Morphology (Shipley, Stone, & Sue, 1983). TELD–3 = Test of Early Language Development—Third Edition (Hresko, Reid, & Hammill, 1999), THT = The Help Test (Lazzari, 1996), TLC–E = Test of Language Competence—Expanded Edition (Wiig & Secord, 1989), TLT–R = The Listening Test—Revised (Barrett, Huisingh, Zachman, Blagden, & Orman, 1992), TNL = Test of Narrative Language (Gillam & Pearson, 2004), TOLD–I3 = Test of Language Development—Intermediate—Third Edition (Hammill & Newcomer, 1997), TOLD–P3 = Test of Language Development—Primary, Third Edition (Newcomer & Hammill, 1997), TOPS–R = Test of Pragmatic Skills—Revised (Shulman, 1986), TOSS–P = Test of Semantic Skills—Primary (Bowers, Huisingh, LoGiudice, & Orman, 2002), TOWK = Test of Word Knowledge (Wiig & Secord, 1992), TOWL–3 = Test of Written Language—Third Edition (Hammill & Larsen, 1996), TRBS = The Renfrew Bus Story (Cowley & Glasglow, 1994), TTC = Token Test for Children (DiSimoni, 1978), TWT–A = The Word Test—Adolescent (Zachman, Huisingh, Barrett, Orman, & Blagden, 1989), TWT–R = The Word Test—Elementary—Revised (Huisingh, Barrett, Zachman, Blagden, & Orman, 1990), UTLD–4 = Utah Test of Language Development—Fourth Edition (Mecham, 2003).
^{c}composite score; ^{s}subtest score; ^{ms}multiple subtest score.

Figure 2. Distribution of mean group differences (in units of standard deviation) for total test scores obtained for children with and without impaired language.



Language Fundamentals—Fourth Edition (CELF–4; Semel, Wiig & Secord, 2003), Preschool Language Scales—Fourth Edition (PLS–4; Zimmerman, Steiner, & Pond, 2002), TEGI (Rice & Wexler, 2001), Test of Language Competence—Expanded Edition (TLC–E; Wiig & Secord, 1989), and Test of Narrative Language (TNL; Gillam & Pearson, 2004) were the only tests of the 43 reviewed that reported acceptable identification accuracy in the test manual. The cutoff scores associated with the reported sensitivity and specificity data ranged from less than –1.0 *SD* to less than –2.0 *SD* when cutoff scores were based on total or composite scores (see Table 4).

Table 2. Summary of score differences for language-impaired and normative or control groups.

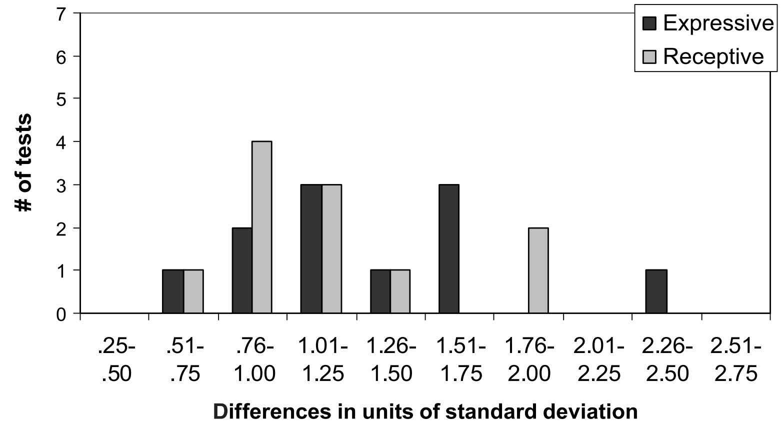
<1 SD *difference*	*Between 1 and 1.5* SD *difference*	*>1.5* SD *difference*
1. ALL	1. BLT–2	1. CELF–4
2. CASL	2. BOEHM–3	2. ELT
3. CREVT–2	3. CELF–P	3. OWLS–WE
4. DELV	4. LPT–R	4. PLS–4
5. EVT	5. OWLS	5. TEGI
6. FLT–AT	6. PLAI–2	6. THT
7. PPVT–III	7. TACL–3	7. TLC–E
8. TOLD–I3	8. TELD–3	8. TNL
9. UTLD–4	9. TOLD–P3	9. TWT–A
	10. TOPS–R	10. TWT–R
	11. TOSS–P	
	12. TOWK	
	13. TOWL–3	
	14. TLT–R	

DISCUSSION

In order for the field of speech-language pathology to move toward evidence-based diagnostic practices, data in support of specific diagnostic practices must be evaluated. Here we evaluated one common clinical practice, that of identifying language impairment by application of an arbitrary low cutoff score to any of the available norm-referenced language tests. We reviewed 43 child language tests, reporting the magnitude of the group differences between the SLI and typically developing (or normative) samples within the test manuals, to determine if the results lend support to the assumption that an arbitrary low cutoff score can diagnose language impairments. Under an evidenced-based practice framework, however, diagnostic accuracy is measured not by mean differences, but by sensitivity and specificity data. Therefore, we also report the number of manuals that include sensitivity and specificity data, the sensitivity and specificity rates, and the test-specific cutoff scores the authors recommend for diagnosing language impairments.

Our review suggests that the practice of applying an arbitrary low cutoff score for diagnosing language impairments is frequently unsupported by the evidence that is available to clinicians in test manuals. The average mean group difference for this sample of tests was –1.34 *SD*. At this level, 43% of all children who have been described as language impaired by the test manuals received scores of ≤1 *SD* from the mean of the normal distribution. Fifty-six percent received scores above –1.5 *SD*. In addition, there were nine tests (Analysis of the Language of Learning [ALL; Blodgett & Cooper, 1987], Comprehensive Assessment of Spoken Language [CASL; Carrow-Woolfolk, 1999a], Comprehensive Receptive and Expressive Vocabulary

Figure 3. Mean group differences (in units of standard deviation) for tests of expressive and receptive language skills.



Test—Second Edition [CREVT–2; Wallace & Hammill, 2002], DELV [Seymour et al., 2003], Expressive Vocabulary Test [EVT; Williams, 1997], The Fullerton Language Test for Adolescents—Second Edition [FLT–A2; Thorum, 1986], Peabody Picture Vocabulary Test—Third Edition [PPVT–III; Dunn & Dunn, 1997], Test of Language Development—Intermediate—Third Edition [TOLD–I3; Hammill & Newcomer, 1997], Utah Test of Language Development—Fourth Edition [UTLD–4; Mecham, 2003]) for which the group mean differences suggest that *most* children with impaired language scored within 1 *SD* of the mean. The mean group differences from data reported in test manuals suggest that scores of children with language impairments on many tests are frequently closer to the normative sample's mean than the commonly applied cutoff scores. Conversely, the overlap for the score distributions for language-impaired and normal children suggests that arbitrary cutoff scores are also likely to identify normal children as impaired in many cases (see also Plante & Vance, 1994).

Figure 4. Mean group differences (in units of standard deviation) for expressive single-word vocabulary test and subtest scores (SW Vocab-E), receptive single-word vocabulary test and subtest scores (SW-Vocab-R), semantic test and subtest scores (Semantic), and morphosyntax test and subtest scores (Morphosyntax).

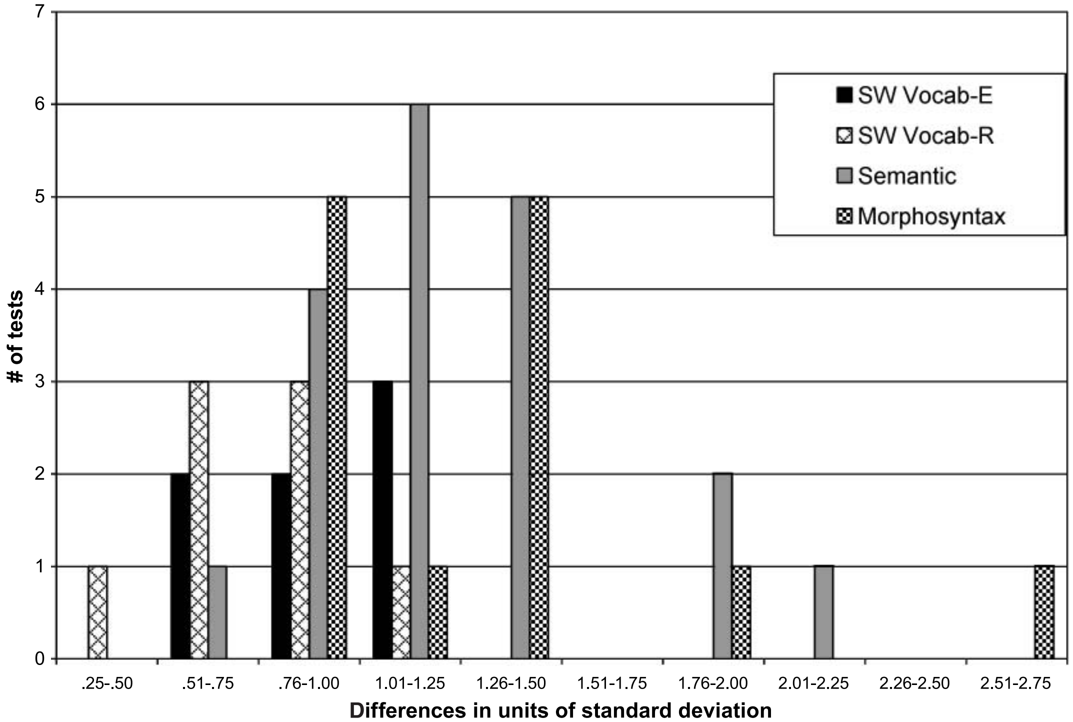
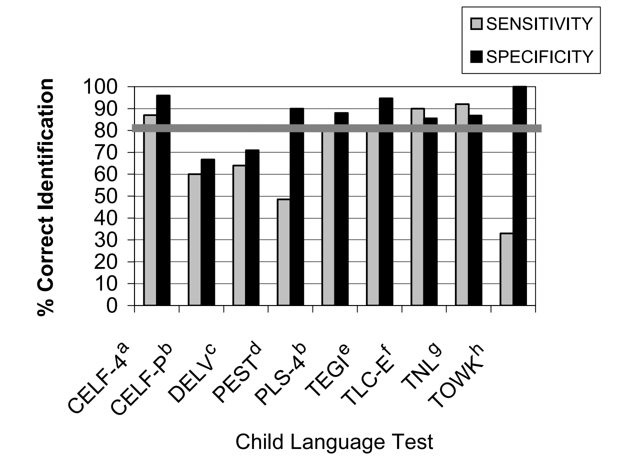


Table 3. Mean difference total test scores between typically developing children and language-impaired children by age group categories in years;months.

	Mean differences for total test scores by age								
Test	*3–3;11*	*4–4;11*	*5–5;11*	*6–6;11*	*7–7;11*	*8–8;11*	*9–9;11*	*10–10;11*	*11–11;11*
BOEHM–P3	1.228	1.143	1.143						
DELV		1.017	.876	1.007	.750	1.005	.874		
ELT			1.455	1.372	1.629	1.916	1.288	2.093	2.398
LPT–R			1.316	1.892	1.185	1.036	1.137	.928	1.490
PLAI–2		1.267	1.267						
PLS–4	1.819	1.810	1.710						
TEGI	2.216	2.023	1.880	1.634					
THT				1.309	2.256	1.644	2.222	1.734	1.574
TLT–R				1.250	1.207	1.649	1.261	1.593	1.608
TOPS–R				1.167	1.003	1.643	1.042	1.309	1.455
TOWK						1.012		1.235	
TWT–R				1.930	1.853	1.557	1.923	1.557	1.585

The probability of obtaining a low score fluctuates across language modalities (receptive vs. expressive) and language domains (e.g., morphosyntax, vocabulary). For example, no test of single-word vocabulary reported a mean group difference of more than 1.5 *SD,* and the majority were ≤1 *SD*. Tests that assessed morphosyntactic skills were more robust than those that assessed other areas of language. These data are consistent with the idea that measuring deficit areas commonly found in children with impaired language will produce robust effects (Rice & Wexler, 2001).

Figure 5. Sensitivity (percentage correctly identified as language impaired) and specificity (percentage correctly identified as typically developing) information from test manuals. [a]Cutoff reported for 1.0, 1.5, and 2.0 *SD* below the mean. Represents 2.0 *SD* below the mean cutoff because this represents the best balance between sensitivity and specificity; [b]Cutoff of less than 1.0 *SD* below the mean; [c]Cutoff is based on the raw scores determined for each age across each domain assessed; [d]Cutoff is 1.28 *SD* below the mean; [e]Cutoff varies by age and provides this information for the range of possible scores for each subtest and composite; [f]Provided information from a regression analysis that weighs the subtest scores to maximize identification accuracy, but does not report the regression formula to determine cutoff score; [g]Cutoff is based on 1.0 *SD* below the mean; [h]Cutoff is based on 1.5 *SD* below the mean.



However, there are pronounced differences in performance across tests that reflect similar constructs (e.g., receptive language, morphosyntactic skills, semantics) (see Table 1). For example, 11 out of 13 tests of morphosyntax (Bankson Language Test—Second Edition [BLT–2; Bankson, 1990], CASL [Carrow-Woolfolk, 1999a], Clinical Evaluation of Language Fundamentals—Preschool [CELF–P; Wiig, Secord, & Semel, 1992], DELV [Seymour et al., 2003], The Expressive Language Test [ELT; Huisingh, Bowers, LoGuidice, & Orman, 1998), FLT–A2 [Thorum, 1986], Test of Auditory Comprehension of Language—Third Edition [TACL–3; Carrow-Woolfolk, 1999b], The Help Test [THT; Lazzari, 1996], TOLD–I3 [Hammill & Newcomer, 1997], Test of Language Development—Primary [TOLD–P3; Newcomer & Hammill, 1997], UTLD–4 [Mecham, 2003]) reported mean score differences of less than –1.5 *SD,* and just over a third of these tests (DELV, FLT–A2, TACL–3, TOLD–I3, UTLD–4) reported mean score differences of ≤–1.0 *SD*. This indicates that some sets of test items are more effective than others for differentiating performance, even within the same language domain (Merrell & Plante, 1997).

The existence of such discrepancies undermines a clinician's ability to compare scores across tests both across and within domains. The discrepancies could be the result of normative samples that are not actually equivalent across tests, which leads to very different standard scores for the same language skill (Merrell & Plante, 1997; Plante & Vance, 1994). Likewise, score variation can result from sometimes subtle differences in how the skill domain is sampled across tests. A child may pass a language target on one test and fail it on another by virtue of how the targets are represented or how the responses are elicited across language measures (Merrill & Plante, 1997).

Although group mean differences for tests did change with age, this change was not predictable enough to support an age-adjusted cutoff score that could be applied across tests. Indeed, test score differences generally increased with age for some tests and decreased or fluctuated with age for others. Because children with language impairments score differently at different ages, applying a single cutoff score for diagnosing language impairment will vary in its accuracy rate, even within a single test, depending on the age of the child. Likewise, it is the case that different tests may be more

Table 4. Identification accuracy for currently available tests.

	Identification accuracy		
Test	*Sensitivity*[a]	*Specificity*[b]	*Cutoff score*[c] *(standard score)*
CELF-4	87%	96%	70
CELF-P	60%	67%	85
CELF-P[d]	80%	89%	96
DELV[e]	64%	71%	
EVT[f]	71%	68%	97
EOWPVT[f]	71%	71%	96
PEST	49%	90%	
PEST[g]	90%	95%	59.95
PLS-4	80%	88%	85
PPVT-3[f]	74%	71%	104
TEGI[h]	81%	95%	
ROWPVT[f]	77%	77%	97
SPELT-3[i]	90%	100%	95
SPELT-P[d]	83%	95%	79.15
TEEM[g]	90%	95%	75
TLC-E	90%	86%	n/a
TNL	92%	87%	85
T0WK	33%	100%	85

Note. Data reported comes from the test manual unless otherwise specified. [a]Rate at which children with language impairment were correctly identified; [b]Rate at which normal children were correctly identified; [c]Score that maximally differentiates between language-impaired and normal distributions; [d]From Plante & Vance, 1995; study included only preschool children; [e]Sensitivity and specificity data are provided by age group for each language domain; See text for how overall sensitivity and specificity was derived; [f]From Gray et al, 1999; study included only preschool children; [g]From Merrell & Plante, 1997; study included only preschool children; [h]Sensitivity and specificity data are available for individual scores; See text for a description of how overall sensitivity and specifity was derived; [i]From Perona et al., in press; study included only preschool children.

effective for identification at different ages. Therefore, it is not just necessary to know the overall score differences for a test, but also how these play out across the range of ages covered within the test.

Although we wanted to evaluate group score differences with respect to demographic variables including gender, minority representation, and socioeconomic status, insufficient information was available in the vast majority of tests to permit such comparisons.[4] However, this information is integral to diagnostic practices because clinicians need to be able to judge how representative the test groups are to their local population of children.

The results of this review indicate that the overall distribution of performance across language task, as seen in Figure 2, belies the notion that these children represent the low end of the normal distribution (McFadden, 1996). Instead, they seem to represent a distribution that is shifted downward and certainly extends to the low end in some cases but also includes language scores that would be considered within normal limits relative to a normal distribution. Because the mean of this shifted distribution falls just above where a "low score cutoff" would be applied (e.g., –1.5 *SD*), the clinical consequence is that a child who truly has a language impairment has a roughly equal chance of being correctly or incorrectly identified, depending on the test that he or she is given. In this review, such differences were apparent when independent samples from each test were compared. However, similar results can also be found within a single sample of children who receive multiple tests (Plante & Vance, 1994).

These data certainly suggest that the wholesale application of a low cutoff score for diagnosing impairment will result in inconsistent identification when it is applied across the currently available language tests. Therefore, the ability to identify language impairment accurately using an arbitrary low cutoff score varies appreciably, depending on the test employed. Even if a child is diagnosed accurately as language impaired at one point in time, future diagnoses may lead to the false perception that the child has recovered, depending on the test(s) that he or she has been given. Although even the tests that result in the most accurate identification still produce some errors, it is clear that an injudicious selection of tests can significantly increase the rate of misidentification. If the cutoff criteria do not match the test selected, typically developing children may be misdiagnosed as language impaired, and children with language impairments may go undiagnosed. The consequences are biased sample composition for researchers and denial of services for children with language impairments.

It is possible that more specifically defined samples of children might have produced a better outcome for individual tests. For example, one might argue that matching the type of test to the type of deficit shown by a child would increase the discrepancy from the typically developing average score. However, this supposition has some noteworthy limitations. First, within a framework of evidence-based practice, the responsible clinician cannot disregard data that are available in favor of potentially stronger outcomes that exist in theory only. To perpetuate a professional practice in the face of *counterevidence* because one can imagine circumstances for which a better outcome might occur (but is unproven to occur) is an unsuitable solution. Second, clinicians often do not know where a child's deficits will lie before testing in order to match the type of test to the child's deficits (and if they did, it would negate the need to test). Therefore, clinicians are left with the prospect of selecting one or more tests at random, or based in part on personal preferences or recommendations by others (Wilson et al., 1991), with the hope that a low score will be obtained on at least one of the tests selected. The de facto interpretation of high and low scores across tests may misrepresent a child's true status if the child's scores are not considered in light of the test's own data on how children with impairment typically score on each test. Failure to consider the available data also raises ethical concerns (see the American Speech-Language-Hearing Association Code of Ethics 2003, Principle 1G concerning evaluation of the effectiveness of services and Principle IID concerning misrepresentation of diagnostic information).

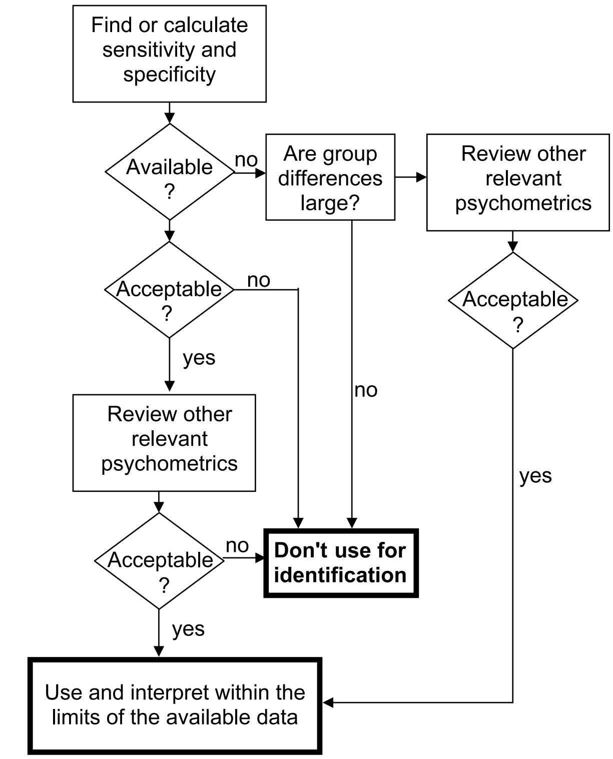
Fortunately, clinicians have data-based alternatives to relying on the "low end of normal" assumption when testing. Current best practices in diagnostics require clinicians to select procedures that are supported by data validating the intended use. When the purpose for testing is to identify impairment, the primary evidence

[4] Tests reporting gender, minority representation, and socioeconomic data for the clinical and typically developing samples include the Comprehensive Assessment of Spoken Language (Carrow-Woolfolk, 1999a), Clinical Evaluation of Language Fundamentals—Fourth Edition (Semel, Wiig, & Secord, 2003), OWLS Listening Comprehension and Oral Expression Scale (Carrow-Woolfolk, 1995), OWLS Written Expression Scale (Carrow-Woolfolk, 1996), and Test of Early Grammatical Impairment (Rice & Wexler, 2001).

required is good sensitivity and specificity. Although these data were virtually absent in manuals only a decade ago, the review of the current tests revealed a core group of tests for which the sensitivity and specificity information is reported (see Table 4). For 5 tests (CELF-4, PLS-4, TEGI, TLC-E, TNL), the accuracy levels reflected by the sensitivity and specificity data support their use for identification of language impairments. Four of these tests (CELF-4, PLS-4, TEGI, TNL) provide sufficient information for clinicians to apply a data-driven method of identification using these tests. In addition, it is sometimes the case that sensitivity and specificity data for a particular test are available in the research literature. To our knowledge, this includes an additional 5 norm-referenced tests (CELF-P, PEST, SPELT-3, SPELT-P, Test of Examining Expressive Morphology [TEEM; Shipley, Stone, & Sue, 1983]) having acceptable sensitivity and specificity (Gray et al., 1999; Merrell & Plante, 1997; Perona et al., in press; Plante & Vance, 1994, 1995). Clinicians may still need to evaluate other aspects of these tests' properties to determine whether additional factors (e.g. population representation, dialect or language representation) or psychometric properties (e.g., reliability) would undermine each of these tests for use with their particular clients. However, a more detailed review that includes these additional characteristics is a poor use of time if the primary evidence of sensitivity and specificity is lacking.

We suggest the following guidelines for clinicians, based on the results of our review. These guidelines are displayed graphically in Figure 6. First and foremost, clinicians should identify their specific purpose in administering a norm-referenced test. If that purpose is to identify the presence (or absence) of language impairment, then the clinician should seek information in the manual that will permit calculation of sensitivity and specificity data. Our experience is that this information is presented in numerous ways within test manuals (under sections dealing with test validity). However, if the available information is placed into the cells of Figure 1, sensitivity and specificity can be readily calculated. If the results support use of the test, then the clinician may wish to review other psychometric properties (e.g., normative information, reliability) that might influence their interpretation or confidence in the child's score. If sensitivity and specificity data are not available, and the clinician is determined to use the test anyway, then data concerning what scores can be expected from children with impaired language should be used as a benchmark for interpreting the score of the child tested. If there is little difference between the language-impaired and normative samples in the manual, then the clinician should have little confidence that a score obtained by the child tested can be interpreted as reflecting either normal or impaired status. However, if the mean differences between groups reported in the test manual are extreme (e.g., >2 *SD* below the cutoff score in use by the school or research criteria), then the probability of correctly identifying children with language impairments is likely to be good. Note that this would require mean score differences of between 3 and 4.5 *SD* below the mean (for a cutoff score criteria of 1 to 1.5 *SD* below the mean), which is a much more robust difference than was reported for any of the tests reviewed for this study. In the end, the weight that a clinician gives to a test score in making his or her final diagnostic decision must be modulated by the strength of the data available to support that decision. For example, if sensitivity and specificity data are strong, and these data were derived from subjects who are comparable to the child tested, then the clinician can be relatively confident in relying on the test score data to aid his or her diagnostic decision. However, if the data are weak, then more caution is warranted and other sources of information on the child's status might have primacy in making a diagnosis.

Figure 6. A decision tree for the review of tests for use in identifying language impairment.


Find or calculate sensitivity and specificity
Available ?
no
Are group differences large?
Review other relevant psychometrics
Acceptable ?
no
Acceptable ?
yes
no
Review other relevant psychometrics
yes
Acceptable ?
no
Don't use for identification
yes
Use and interpret within the limits of the available data

In summary, a simplified review of critical information in test manuals (e.g., sensitivity and specificity data, mean group differences) can serve to determine whether the interpretations that a clinician intends to make are empirically justified. Furthermore, these data can also assist the clinician in determining the degree of confidence that a clinician should place in the interpretation that he or she makes. For example, if the sensitivity and specificity of a test are high, confidence in the classification of typically developing versus language disordered should increase. However, to the extent that the child in question is similar to or different from the sample from which these data were derived, a clinician may need to adjust his or her confidence level appropriately. This consideration of both the interpretation of test data and the confidence in that interpretation reflects the probabilistic nature of diagnostics. Test results can only indicate the likelihood, rather than the certainty, that an impairment is present. A simple review of the currently available evidence can greatly improve the clinician's certainty in this clinical determination.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Deafness and Other Communication Disorders Grant R01 DC04726. Additional support was provided through a fellowship to the first author from the Bamford-Lahey Children's Foundation.

REFERENCES

American Speech-Language-Hearing Association. (2003). *Code of ethics*. Rockville, MD: Author.

Bankson, N. W. (1990). *Bankson Language Test—Second Edition*. Austin, TX: Pro-Ed.

Barrett, M., Huisingh, R., Zachman, L., Blagden, C., & Orman, J. (1992). *The Listening Test—Revised*. East Moline, IL: LinguiSystems.

Blank, M., Rose, S. A., & Berlin, L. J. (2003). *Preschool Language Assessment Instrument—Second Edition*. Austin, TX: Pro-Ed.

Blodgett, E. G., & Cooper, E. B. (1987). *Analysis of the Language of Learning*. East Moline, IL: LinguiSystems.

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.

Bowers, L., Huisingh, R., LoGiudice, C., & Orman, J. (2002). *Test of Semantic Skills—Primary*. East Moline, IL: LinguiSystems.

Bracken, B. A. (1998). *Bracken Basic Concept Scale—Revised*. San Antonio, TX: The Psychological Corporation.

Brackenbury, T., & Pye, C. (2005). Semantic deficits in children with language impairments: Issues for clinical assessment. *Language, Speech, and Hearing Services in Schools, 36,* 5–16.

Carrow-Woolfolk, E. (1995). *OWLS Listening Comprehension and Oral Expression Scale*. Circle Pines, MN: American Guidance Service.

Carrow-Woolfolk, E. (1996). *OWLS Written Expression Scale*. Circle Pines, MN: American Guidance Service.

Carrow-Woolfolk, E. (1999a). *Comprehensive Assessment of Spoken Language*. Circle Pines, MN: American Guidance Service.

Carrow-Woolfolk, E. (1999b). *Test of Auditory Comprehension of Language—Third Edition*. Austin, TX: Pro-Ed.

Cowley, J., & Glasgow, C. (1994). *The Renfrew bus story*. Centreville, DE: The Centreville School.

Dawson, J. I., Stout, C. E., & Eyer, J. A. (2003). *The Structured Photographic Expressive Language Test—Third Edition*. Dekalb, IL: Janelle Publications.

DiSimoni, F. (1978). *Token Test for Children*. Hingham, MA: Teaching Resources.

Dollaghan, C. A. (2004). Taxometric analyses of specific language impairment in 3- and 4-year-old children. *Journal of Speech, Language, and Hearing Research, 47,* 464–475.

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

Flax, J. F., Realpe-Bonilla, T., Hirsch, L. S., Brzustowitz, L. M., Bartlett, C. W., & Tallal, P. (2003). Specific language impairment in families: Evidence for co-occurrence with reading impairments. *Journal of Speech, Language, and Hearing Research, 46,* 530–543.

Ford, J. A., & Milosky, L. M. (2003). Inferring emotional reactions in social situations: Differences in children with language impairment. *Journal of Speech, Language, and Hearing Research, 46,* 21–30.

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

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

Gillam, R., & Pearson, N. (2004). *Test of Narrative Language*. Austin, TX: Pro-Ed.

Gray, S. (2003). Word learning by children with specific language impairment: What predicts success? *Journal of Speech, Language, and Hearing Research, 46,* 56–67.

Gray, S., Plante, E., Vance, R., & Henrichsen, M. (1999). Performance of SLI and NL children on four tests of single-word vocabulary. *Language, Speech, and Hearing Services in Schools, 30,* 196–206.

Hammill, D., & Larsen, S. (1996). *Test of Written Language—Third Edition*. Austin, TX: Pro-Ed.

Hammill, D. D., & Newcomer, P. L. (1997). *Test of Language Development—Intermediate—Third Edition*. Austin, TX: Pro-Ed.

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

Huisingh, R., Bowers, L., LoGuidice, C., & Orman, J. (1998). *The Expressive Language Test*. East Moline, IL: LinguiSystems.

Hresko, W. P., Reid, D. K., & Hammill, D. D. (1999). *Test of Early Language Development—Third Edition*. Austin, TX: Pro-Ed.

Lazzari, A. M. (1996). *The Help Test*. East Moline, IL: LinguiSystems.

Leonard, L. B. (1991). Specific language impairment as a clinical category. *Language, Speech, and Hearing Services in Schools, 22,* 66–68.

Leonard, L. B. (1998). *Children with specific language impairment*. Cambridge, MA: MIT Press.

Leonard, L. B., Deevy, P., Miller, C. A., Rauf, L., Charest, M., & Kurtz, R. (2003). Surface forms and grammatical functions: Past tense and passive participle use by children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 46,* 43–55.

Maillart, C., Schelstraete, M., & Hupet, M. (2004). Phonological representations in children with SLI: A study of French. *Journal of Speech, Language, and Hearing Research, 47,* 187–198.

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

Mecham, M. J. (2003). *Utah Test of Language Development—Fourth Edition*. Austin, TX: Pro-Ed.

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

Newcomer, P. L., & Hammill, D. D. (1997). *Test of Language Development—Primary, Third Edition*. Austin, TX: Pro-Ed.

Paradis, J., Crago, M., Genesee, F., & Rice, M. (2003). French–English bilingual children with SLI: How do they compare with their monolingual peers. *Journal of Speech, Language, and Hearing Research, 46,* 113–127.

Perona, K., Plante, E., & Vance, R. (in press). Diagnostic accuracy of the Structured Photographic Expressive Language Test: Third Edition. *Language Speech, and Hearing Services in Schools.*

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,* 70–76.

Rescorla, L. (1989). The Language Development Survey: A screening tool for delayed language in toddlers. *Journal of Speech and Hearing Disorders, 54,* 587–599.

Rescorla, L., & Alley, A. (2001). Validation of the Language Development Survey (LDS): A parent report tool for identifying language delay in toddlers. *Journal of Speech, Language, and Hearing Research, 44,* 434–445.

Rice, M. L. (1994). Grammatical categories of children with specific language impairment. In R. V. Watkins & M. L. Rice (Eds.), *Specific language impairments in children* (pp. 69–90). Baltimore: Brookes.

Rice, M. L., & Wexler, K. (1996). Toward tense as a clinical marker of specific language impairment in English-speaking children. *Journal of Speech and Hearing Research, 39,* 1239–1257.

Rice, M. L., & Wexler, K. (2001). *Rice/Wexler Test of Early Grammatical Impairment*. San Antonio, TX: The Psychological Corporation.

Rice, M. L., Wexler, K., Marquis, J., & Hershberger, S. (2000). Acquisition of irregular past tense by children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 43,* 1126–1145.

Richard, G. J., & Hanner, M. (1995). *The Language Processing Test—Revised*. Austin, TX: Pro-Ed.

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

Seymour, H. N., Roeper, T. W., & deVilliers, J. (2003). *Diagnostic Evaluation of Language Variation*. San Antonio, TX: The Psychological Corporation.

Shipley, K. G., Stone, T. A., & Sue, M. B. (1983). *Test for Examining Expressive Morphology*. Tucson, AZ: Communication Skill Builders.

Shulman, B. B. (1986). *Test of Pragmatic Skills—Revised*. Tucson, AZ: Communication Skill Builders.

Thorum, A. R. (1986). *The Fullerton Language Test for Adolescents—Second Edition*. Austin, TX: Pro-Ed.

Tomblin, J. B., Records, N. L., Buckwalter, P., Zhang, X., Smith, E., & O'Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language, and Hearing Research, 40,* 1245–1260.

Wallace, G., & Hammill, D. D. (2002). *Comprehensive Receptive and Expressive Vocabulary Test—Second Edition*. San Antonio, TX: The Psychological Corporation.

Wells, B., & Peppé, S. (2003). Intonation abilities of children with speech and language impairments. *Journal of Speech, Language, and Hearing Research, 46,* 5–20.

Werner, E., & Kresheck, J. D. (1983). *Structured Photographic Language Test—Preschool*. Sandwich, IL: Janelle Publications.

Wiig, E. H., & Secord, W. (1989). *Test of Language Competence—Expanded Edition*. San Antonio, TX: The Psychological Corporation.

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

Wiig, E. H., Secord, W. A., & Semel, E. (1992). *Clinical Evaluation of Language Fundamentals—Preschool*. San Antonio, TX: The Psychological Corporation.

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

Wilson, K. S., Blackmon, R. C., Hall, R. E., & Elcholtz, G. E. (1991). Methods of language assessment: A survey of California public school clinicians. *Language, Speech, and Hearing Services in Schools, 22,* 236–241.

Young, E. C., & Perachio, J. J. (1993). *Patterned Elicitation Syntax Test*. Tucson, AZ: Communication Skill Builders.

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. (2002). *Preschool Language Scales—Fourth Edition*. San Antonio, TX: The Psychological Corporation.

Received October 7, 2004
Accepted April 19, 2005
DOI: 10.1044/0161-1461(2006/007)

Contact author: Tammie J. Spaulding, Department of Speech, Language, and Hearing Sciences, P.O. Box 210071, The University of Arizona, Tucson, AZ 85721-0071. E-mail: spauld20@email.arizona.edu