The Composition of Normative Groups and Diagnostic Decision Making: Shooting Ourselves in the Foot

Elizabeth D. Peña
The University of Texas at Austin

Tammie J. Spaulding
Elena Plante
University of Arizona, Tucson

Purpose: The normative group of a norm-referenced test is intended to provide a basis for interpreting test scores. However, the composition of the normative group may facilitate or impede different types of diagnostic interpretations. This article considers who should be included in a normative sample and how this decision must be made relative to the purpose for which a test is intended.
Method: The way in which the composition of the normative sample affects classification accuracy is demonstrated through a test review followed by a simulation study. The test review examined the descriptions of the normative group in a sample of 32 child language tests. The mean performance reported in the test manual for the sample of language impaired children was compared with the sample's norms, which either included or excluded children with language impairment. For the simulation, 2 contrasting normative procedures were modeled. The first procedure included a mixed group of representative cases (language impaired and normal cases). The second procedure excluded the language impaired cases from the norm.
Results: Both the data obtained from test manuals and the data simulation based on population characteristics supported our claim that use of mixed normative groups decreases the ability to accurately identify language impairment. Tests that used mixed norms had smaller differences between the normative and language impaired groups in comparison with tests that excluded children with impairment within the normative sample. The simulation demonstrated mixed norms that lowered the group mean and increased the standard deviation, resulting in decreased classification accuracy.
Conclusions: When the purpose of testing is to identify children with impaired language skills, including children with language impairment in the normative sample can reduce identification accuracy.

Key Words: evidence-based practice, assessment, classification, language impairment

Normative tests are a cornerstone of diagnosis in the field of speech-language pathology. A normative sample provides a standard against which the performance of a particular individual can be compared. However, the nature of the comparison and the inferences to be drawn from that comparison are influenced by the composition of the normative group. In this article, we consider how the composition of normative samples found in tests intended for use with clinical populations can affect the diagnostic process.

Normative samples typically include individuals who represent the age and demographic characteristics of those for whom the test is intended. Most often this sample is drawn from the general population. However, a different reference population may be appropriate depending on the purpose of the test. For example, for tests intended for use with neurogenic disorders, the primary comparative group might be a sample of individuals with diagnoses similar to the test's target population (e.g., traumatic head injury, stroke). In these cases, the purpose of diagnostic testing is not to determine whether a condition is present (that diagnosis is made from medical information) but to describe the severity of the behavioral sequelae. A sample of neurologically normal individuals will not provide this information because test

1058-0360/06/1503-0247

items typically tap skills that neurologically normal individuals can do well, and their performance is typically at ceiling levels. Therefore, this type of normative group would not permit an accurate estimation of severity because their range of scores does not extend low enough to represent the range of severity seen in the target population. In addition, sole reliance on a normal reference during test development may result in the selection of inadequate numbers of items that reflect the lower skill levels necessary to finely differentiate severity in a neurogenic clinical population. The score distribution of a clinical sample, in contrast, can provide insight into severity of deficit and possibly assist in prognostic estimates.

This example highlights the centrality of purpose of a particular test and the clinical decisions it is intended to help the clinician make concerning the individual he or she is testing. Severity estimation is only one possible purpose for a norm-referenced test. In the case of testing children, very often our purpose for testing is not to determine the severity of an impairment but to determine whether an impairment exists. The question then arises of who this child's performance should be compared with in order to make this decision. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) states that "Norms, if used, should refer to clearly described populations. These populations should include individuals or groups to whom test users will ordinarily wish to compare their own examinees" (p. 55). Although not actually stated in the standard, in the field of communication disorders, it is often interpreted to mean that tests should include children with documented language impairment (LI) in the normative group.

The exclusion of children with disabilities from normative groups has been criticized as a limitation of some tests in the area of language assessment (McFadden, 1996). But we suggest that the opposite may be true—that the inclusion of children with disabilities may be at odds with the goal of classification, typically the primary function of the speech pathologist's assessment. In fact, by including such children in the normative sample, we may be "shooting ourselves in the foot" in terms of testing for the purpose of identifying disorders.

The practice of developing broad norms that are inclusive of children with a range of abilities reflects two major issues in assessment. The first is the notion of "fairness." This fairness argument takes both social and mathematical forms. First, if those with disabilities are routinely excluded from norms, then the normative group's demographic characteristics will not include the children for whom the test is intended. This would parallel the idea of desiring minority representation for tests intended for use with minority children. It has also been argued that excluding children with LI from tests of child language will truncate the normative distribution, because these children make up the low end of the normal (i.e., bell-shaped) distribution (McFadden, 1996). This argument asserts that the removal of language impaired cases, which represent the low scores in a normal distribution, will artificially skew the normative distribution. This purportedly results in a number of problems, among which is the labeling of children with low but normal scores as impaired because they now fall at the low range of the truncated distribution. In addition, it asserts that children with disabilities will appear more impaired than they are because they are compared with higher scoring children with typical language development.

There are two major problems with this argument. The first is that children with LI do not necessarily score at the low end of the normal distributions of currently available language tests (Spaulding, Plante, & Farinella, 2006). Therefore, eliminating them from the normative sample does not truncate its distribution. The average difference between these children and the normative samples, as reported in the test manuals, is –1.34 *SD,* with score differences normally distributed above and below this mean. Therefore, rather than representing the lower tail of the normal distribution, the distribution of children with LI is somewhat lower overall on average than their typically developing peers but shows substantial overlap with the normal distribution. The second and perhaps more problematic issue with the argument is that it ignores the primary purpose for which a language test might be administered in the field. If the purpose is to identify the presence of LI, then the comparison of interest is whether the performance of the child being tested is consistent with the performance of children who are developing language typically or children known to have LI.

The primary evidence needed to support this diagnostic decision is test specificity and sensitivity. Specificity is the correct classification of typically developing children as having normal language (NL). Sensitivity refers to the correct classification of children with LI as having LI (Dollaghan, 2004; Plante & Vance, 1994). Here we ask an independent question: whether sensitivity and specificity might also be affected by the inclusion of children with LI in norms. We address this issue from two perspectives. First, we turn to data available in test manuals to determine the relative magnitude of this effect on currently available norm-referenced tests. Next, we conduct a simulation to illustrate the impact of including or excluding children with LI from the normative group on diagnostic accuracy.

Test Data: Excluding Versus Including Children With LI in Normative Group

We have made a logical argument that the inclusion of impaired children in normative samples may be ill-advised. We will now present empirical evidence from commercially available tests to support the validity of this argument.

We performed a reanalysis of data originally assembled as part of a review of published tests of language skills (Spaulding et al., 2006). This review addressed the utility of 43 language tests relative to the purpose of identifying LI in children. For the purposes of the current article, 32 tests examined by Spaulding et al. reported data concerning performance of children with LI and information on the

composition of the normative sample. We reanalyzed these data to determine the effect of including children with impairments in the normative sample on the ability to differentiate normal and impaired children. Of the tests examined by Spaulding et al., 32 reported data concerning performance of children with LI and information on the composition of the normative sample relevant to our purpose here. These data form the basis of our empirical example.

We reviewed the descriptions of the normative groups in the sample of the 32 language tests. Of these, 13 test manuals made an unambiguous statement that children with disabilities were excluded from the normative samples. For another 19 tests, the manuals indicated that impaired children were not excluded or were purposefully included in the normative sample. These two groups of tests included in our analyses are listed in Table 1.

For the tests listed in Table 1, we recorded the mean performance reported in the test manual for children with LI. We then determined how discrepant this mean was from that of the normative group using the following simple calculation: group difference = (sample mean – normative mean) / standard deviation.

When a standard deviation was reported for both a language impaired and a normal (or normative) sample, the larger of the two was used in this calculation because it provided the more conservative estimate of group difference.

Figure 1 compares the mean group differences for tests that either include impaired children in the normative sample or exclude them. As this figure shows, there is less separation between the normative sample and impaired samples when children with LI are included (Mixed) in the norms than when they are excluded (Normal Only). Thus, comparing a sample of children with LI to a Mixed normative sample is associated with a reduced difference between the groups. In addition, the standard deviation (indicated by the error bar) is larger for the Mixed than the Normal Only normative sample. Both the smaller group differences and the larger standard deviation should function to reduce the ability to distinguish children with LI from their typically developing peers. As a result, we can expect that sensitivity and specificity would likewise be reduced. Thus, data reported in test manuals support the principle that the inclusion of children with impairments in the normative group has a negative impact on the ability to identify individuals with childhood LI.

TABLE 1. Tests of child language categorized according to whether children with impairments were excluded (Normal Only) or included (Mixed) in the normative group.

Test	Author
Normal Only normative group	
Analysis of the Language of Learning	Blodgett & Cooper (1987)
Clinical Evaluation of Language Fundamentals—Preschool	Wiig, Secord, & Semel (1992)
The Expressive Language Test	Huisingh, Bowers, LoGuidice, & Orman (1998)
The Help Test	Lazzari (1996)
The Language Processing Test—Revised	Richard & Hanner (1995)
The Listening Test	Barrett, Huisingh, Zachman, Blagden, & Orman (1992)
Oral and Written Language Scales: Listening Comprehension and Oral Expression	Carrow-Woolfolk (1995)
Oral and Written Language Scales: Written Expression	Carrow-Woolfolk (1996)
Rice/Wexler Test of Early Grammatical Impairment	Rice & Wexler (2001)
Test of Pragmatic Skills—Revised	Shulman (1986)
Test of Semantic Skills—Primary	Bowers, Huisingh, LoGuidice, & Orman (2002)
The Word Test—Adolescent	Zachman, Huisingh, Barrett, Orman, & Blagden (1989)
The Word Test—Elementary—Revised	Huisingh, Barrett, Zachman, Blagden, & Orman (1990)
Mixed normative group	
Boehm Test of Basic Concepts—Preschool, Third Edition	Boehm (2001)
Clinical Evaluation of Language Fundamentals, Fourth Edition	Semel, Wiig, & Secord (2003)
Comprehensive Assessment of Spoken Language	Carrow-Woolfolk (1999a)
Comprehensive Receptive and Expressive Vocabulary Test—Second Edition	Wallace & Hammill (2002)
Diagnostic Evaluation of Language Variation	Seymour, Roeper, & deVilliers (2003)
Expressive Vocabulary Test	Williams (1997)
The Fullerton Language Test for Adolescents, Second Edition	Thorum (1986)
Peabody Picture Vocabulary Test—III	Dunn & Dunn (1997)
Preschool Language Assessment Instrument—Second Edition	Blank, Rose, & Berlin (2003)
Preschool Language Scales, Fourth Edition	Zimmerman, Steiner, & Pond (2002)
Test of Auditory Comprehension of Language—Third Edition	Carrow-Woolfolk (1999b)
Test of Early Language Development, Third Edition	Hresko, Reid, & Hammill (1999)
Test of Language Comprehension—Expanded Edition	Wiig & Secord (1989)
Test of Language Development—Intermediate	Hammill & Newcomer (1997)
Test of Language Development—Primary, Third Edition	Newcomer & Hammill (1997)
Test of Narrative Language	Gillam & Pearson (2004)
Test of Word Knowledge	Wiig & Secord (1992)
Test of Written Language—Third Edition	Hammill & Larsen (1996)
Utah Test of Language Development, Fourth Edition	Mecham (2003)

FIGURE 1. Mean group differences for tests that include or exclude impaired children from the normative sample. Error bars indicate the standard deviation.

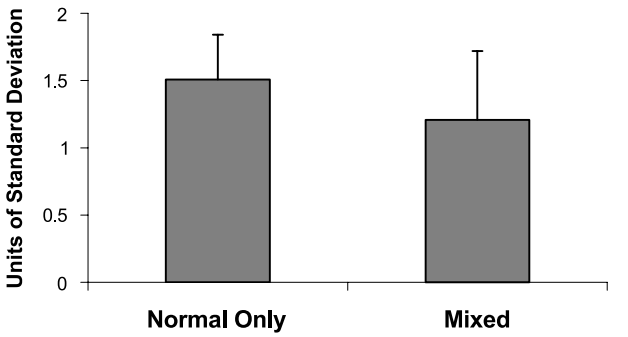


TABLE 2. Mean standard scores, standard deviations, and cut-points for simulated cases.

Group	*N*	*M*	*SD*	–1.14 *SD*[a]	–1.34 *SD*[b]	–1.5 *SD*[c]
Mixed	2,000	101.10	14.52	84.55	81.65	79.33
Normal Only	1,852	102.94	13.40	87.66	84.98	82.84
LI	148	78.11	5.39			

Note. LI = language impairment.

[a]Cut-point based on Tomblin et al.'s (1996) Epi-SLI method.
[b]Cut-point based on Spaulding et al.'s (2006) report of mean score differences for norm-referenced language tests.
[c]Cut-point selected to reflect common clinical practice for identification of children with LI (Spaulding et al., 2006).

Simulation: Excluding Versus Including Children With LI in Normative Group

To better understand the implications of how Mixed versus Normal Only normative samples affect sensitivity and specificity data, we generated a simulated data set that could be used to evaluate classification accuracy. Data for the simulation were created using the random number generator in SPSS Version 13 to generate two normally distributed groups (LI and normal). A total of 2,000 cases were generated using the compute function RV.NORMAL(mean, stddev). Based on prevalence data of 7.4% (Tomblin et al., 1997), 148 of these cases were identified as LI with a mean of 77.0 and standard deviation of 5.2 (Fey, Catts, Proctor-Williams, Tomblin, & Zhang, 2004).[1] Normal cases had a mean of 102.7 and a standard deviation of 13.3 (Fey et al., 2004).

The means and standard deviations for the total group and for the subgroups (LI and normal) were calculated and are displayed in Table 2. Two occurrences can be observed in the conditions of including impaired cases (Mixed group) and excluding them from the norm (Normal Only group). First, the Mixed group mean (101.10) is slightly lower than that of the Normal Only group mean (102.94). Second, the standard deviation for the Mixed group (14.52) is larger than for the Normal Only group (13.40). Diagnostic cut-points were calculated based on the mean and standard deviation of the comparison population (see Table 2).

In our example, we have selected three cut-points to reflect two data-based sources and one that is arbitrary but common in clinical practice. We have adopted the –1.14-*SD* cutoff of the Epi-SLI method developed by Tomblin, Records, and Zhang (1996), which resulted in 86% sensitivity and 99% specificity in their sample. For comparison purposes, we provide data on two other cut-points. The second cut-point reflects the mean score difference for language impaired and normal children on a large sample of child language tests (Spaulding et al., 2006). As such, it estimates the population mean for children broadly selected as language impaired, because the mean of sample means approximates the population mean (the Theorem of Central Limits). In this case, the sample means are obtained from the individual test manuals. Thus, the average sensitivity at this cut-point would approximate 50% across the entire collection of 32 or 43 tests examined by Spaulding et al. that included information on performance of children with and without LI. The third cut-point of –1.5 *SD* was selected because it appeared in the placement criteria for multiple school districts in the United States (Spaulding et al., 2006), although the resulting sensitivity and specificity levels that should result from the use of this criterion with any given test is often unknown. Note that in all three cases, the cut-points change depending on the comparison group. The cut-point is lower when the cut-point is relative to the Mixed group and higher when the cut-point is relative to the Normal Only group.

The comparison between the population curves for the Normal Only (solid line) and impaired cases (dotted line) are compared in Figure 2. The vertical line indicates the conventional cut-points used for clinical identification based on the Epi-SLI cutoff of 1.14 *SD*s below the population mean.

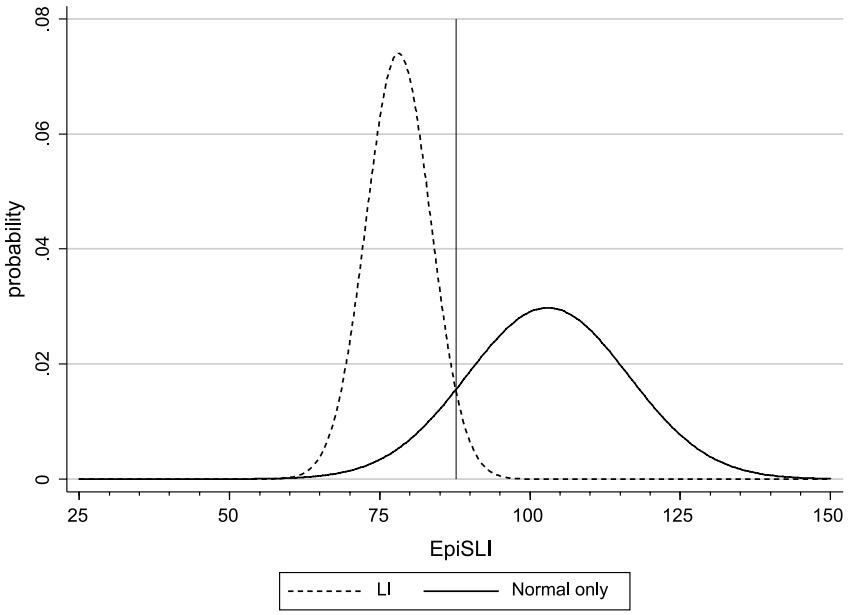
Figure 3 additionally shows the population curve for the Mixed group (dashed line). Although there is a great deal of overlap with the Normal Only group, the mean shifts to the left, and there is more overlap with the LI group. Here, a vertical line is drawn 1.14 *SD*s below the Mixed group mean. Note that because the standard deviation is larger, the cut-point is farther to the left, potentially resulting in an increase of false negatives.

Classification Analysis

It is evident that including cases of LI in the comparison group shifts the norm slightly downward and results in a lower cut-point. What does this mean in practical terms? Using the simulated data set described, we ran classification analyses to project sensitivity and specificity using the two types of norms. As before, the Mixed group includes the language impaired cases whereas the Normal Only group excludes those cases. The cut scores for each set were

[1]Fey et al. (2004) was selected because identification of LI was based on the Epi-SLI composite cutoff of 1.14.

FIGURE 2. Comparison of Normal Only and language impaired (LI) groups.



those provided in Table 2, and the estimated sensitivity and specificity were derived from the mean and standard deviation of the Mixed and Normal Only normative groups.

Table 3 displays the classification analyses based on these different criteria for setting the cut score. As seen here, characteristics of the normative sample directly affect diagnostic accuracy. Plante and Vance (1994) suggest that tests with 80% specificity and sensitivity have "fair" classification accuracy and those with 90% or more have "good" classification accuracy. In this simulation, sensitivity changes from good to fair at a cut score of –1.14 by including individuals with LI in the normative sample. At other cut scores, sensitivity drops below

FIGURE 3. Comparison of Normal Only, Mixed, and LI groups.

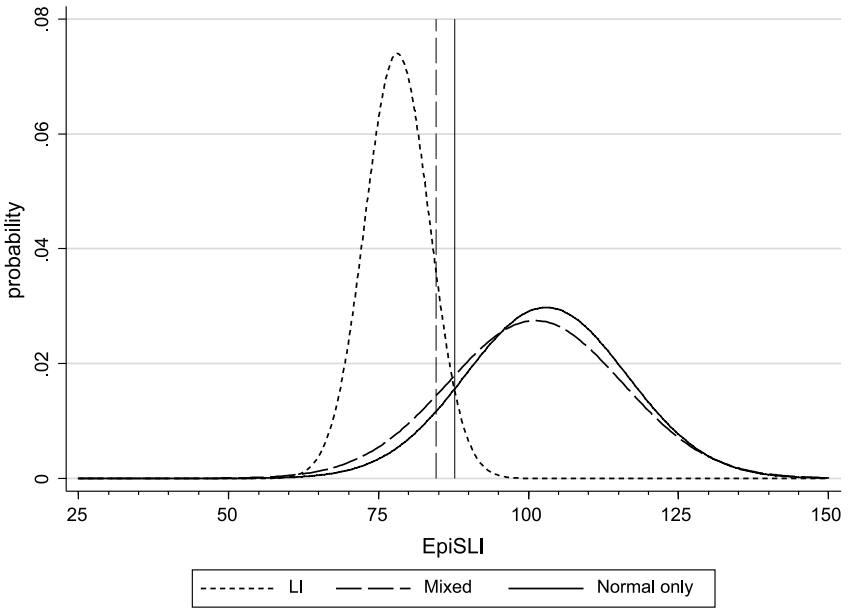


TABLE 3. Classification accuracy of pretest, modifiability, and posttest measures.

Comparison group	Cut score	LI as LI	Sensitivity (%)	False neg.	Error (%)	NL as NL	Specificity (%)	False pos.	Error (%)
Mixed	–1.14	126	85.1	22	14.9	1705	92.1	147	7.9
Normal Only	–1.14	143	96.6	5	3.4	1628	87.9	224	12.1
Mixed	–1.34	101	68.2	47	31.8	1754	94.7	98	5.3
Normal Only	–1.34	130	87.8	18	12.2	1685	91.0	167	9.0
Mixed	–1.5	88	59.5	60	40.5	1786	96.4	66	3.6
Normal Only	–1.5	109	73.6	39	26.4	1733	93.6	119	6.4

Note. NL = normal language.

acceptable levels altogether. On the other hand, specificity increases, because the cutoff for impairment has been shifted downward.[2]

Implications for Test Interpretation

Test scores are most typically interpreted in one of two ways: relative or absolute. Relative interpretations provide information about the test taker or group of test takers relative to a population (American Educational Research Association et al., 1999). This type of score interpretation is made when clinicians wish to gauge severity of impairment or to develop a profile of relative skill levels across domains. Relative score interpretations are of primary interest when tests are norm-referenced. On the other hand, an absolute interpretation is used to interpret an individual test score (or group of test scores) in comparison with defined standards. This type of score interpretation is common when using test scores to indicate the presence or absence of LI or to determine whether performance matches a criterion for age- or grade-appropriate expectations. This type of score is typically of primary importance for criterion-referenced testing. It is up to the test developer to determine the primary interpretation of the test, and this interpretation will determine the appropriate comparison group. Therefore, the purpose of testing and the types of interpretations that a clinician intends to make are of central importance in the diagnostic process.

When the purpose is identification, clinicians make an absolute interpretation of test scores. Subsequently, decisions concerning the psychometric characteristics of the test a clinician selects must be made relative to whether it supports or undermines the nature of the intended interpretation. In other words, a test's psychometric characteristics, including the composition of its normative group, are not always inherently good or bad. Instead, they may be better or worse relative to whether they support the clinician's ability to make specific types of diagnostic decisions. The inclusion of impaired children in normative samples shifts that distribution in ways that undermine a clinician's ability to identify LI. Tests that include mixed norms are more likely to underidentify children with LI and should not be the primary choice for such a purpose. Conversely, tests that exclude children with LI from their sample are more likely to accurately identify children with LI. Note, however, that in this simulation an increase in accurate identification of language impaired cases also increased false positives. But the increase in false positives is relatively small (between 2.8% and 4.2%) compared with the decrease in false negatives (between 11.5% and 14.1%). This small increase in false positives in the simulation is likely due to our assumption of a normal distribution for both LI and NL data sets. This shift highlights the idea that test selection for diagnostic purposes most appropriately considers a balance of specificity and sensitivity (discussed below).

When might it be appropriate to include a broad representative sample in a norm? A broad representative sample of children of all capabilities in a norm is likely to be appropriate when the intended interpretation of the test scores is relative rather than absolute. Documentation about severity of the LI is one example. Here, the purpose is to determine the degree of difference from the mean of the general (rather than the normal) population. Intelligence tests are also used in a way that is relative rather than absolute in that IQ scores place an individual's cognitive skills in a context that is relative to the general population. A mixed group norm is advantageous specifically because the addition of impaired cases to the distribution broadens the variability within the normative group (see Figure 3). Therefore, a wider range of scores is represented, and the gradations in score level are increased relative to the Normal Only distribution.

We have demonstrated here that test diagnostic accuracy can be affected by the composition of the normative group. If the normative group includes a mixed sample, measurement precision decreases for absolute score interpretation but may increase for relative interpretations. Given that clinicians may intend to make either type of interpretation based on a norm-referenced test, they should be mindful of two principles: First, clinicians and researchers must consider how they wish to interpret test results in order to determine whether test characteristics are appropriate for that purpose. Second, a single test may not be able to support multiple diagnostic purposes. Here we demonstrate that the composition of the normative group can either support the use of a test for absolute score interpretations or undermine it for relative interpretations. Elsewhere, we (Merrell & Plante, 1997; Plante & Vance, 1994) and others (McCauley &

[2]Note that there are trade-offs between specificity and sensitivity; as one increases, the other decreases. These trade-offs depend in part on the base rate for the impairment in the population (in this case 7.4% for LI). A higher base rate, for example, would have a greater effect on specificity because in mixed norms it would further lower both the mean and cut-point.

Swisher, 1984b) have criticized the practice of drawing therapy targets from norm-referenced test items from both psychometric and evidence-based perspectives. Therefore, it is typically not the case that the psychometric characteristics of a test will be optimal for multiple diagnostic purposes.

Implications for Test Development

The standards for test development within speech-language pathology have evolved with the field. This evolution is driven, in part, by clinical training that emphasizes psychometric review of test properties (e.g., McCauley & Swisher, 1984a, 1984b) and the importance of the diagnostic purpose and the related clinical decisions in test validity (Merrell & Plante, 1997; Plante & Vance, 1994, 1995) and evidence-based practice (Dollaghan, 2004). These trends have decreased the emphasis on general psychometric checklists in favor of evidence of strong sensitivity and specificity for tests intended for use in identifying cases of child LI (Plante & Vance, 1994). More recently, researchers have placed an emphasis on the content representation of tests intended for the diagnosis of particular forms of LI. For example, researchers are identifying specific types of deficits that serve as "clinical markers" of specific language impairment (SLI), including deficits involving grammatical morphology (Bedore & Leonard, 1998; Rice & Wexler, 1996), nonword repetition (Dollaghan & Campbell, 1998; Simkin & Conti-Ramsden, 2001), or narratives (Botting, 2002; Liles, Duffy, Merritt, & Purcell, 1995; Reilly, Bates, & Marchman, 1998). As such, tests built around areas of known deficit have a higher probability of resulting in good sensitivity to the targeted impairment (in this example, SLI) than tests built to reflect broader language skills. For example, the Rice/Wexler Test of Early Grammatical Impairment (Rice & Wexler, 2001) and the Test of Narrative Language (Gillam & Pearson, 2004) are two recent tests whose content is derived directly from a research base on SLI. In both cases, the test sensitivity and specificity support use of these tests to identify LI.

We would add to these trends a need to carefully consider the composition of the normative group relative to the test's intended purpose. When combined with careful selection of test content and the development of empirically derived cutoff scores for group differentiation, the exclusion of language impaired children from test norms used for identification of LI will improve a test's ability to accurately classify children. Improved tools for identification will help diagnose children with LI who often go undetected, especially in the early elementary years (cf. Tomblin et al., 1997). Identification of children with LI at an earlier—rather than a later—age is preferable because of the association of LI and later difficulties such as reading (Catts, Fey, Zhang, & Tomblin, 2001; Law & Durkin, 2000). Intervention potentially improves educational outcomes and teacher perceptions in later grades (Urwin, Cook, & Kelly, 1988) and improves literacy skills (Bernhardt & Major, 2005) as well as social skills and self-esteem (Law & Sivyer, 2003). These outcomes are predicated on early and accurate identification. Therefore, it is to the field's advantage that the characteristics that maximize correct identification be built into norm-referenced tests.

Acknowledgments

Funding was provided to the second author by the Bamford-Lahey Children's Foundation. The authors would like to thank Lynn Gale, PhD, for the production of Figures 2 and 3. This work was completed while the first author was a Fellow at the Center for Advanced Study in the Behavioral Sciences.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

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

Bedore, L. M., & Leonard, L. B. (1998). Specific language impairment and grammatical morphology: A discriminant function analysis. *Journal of Speech, Language, and Hearing Research, 41,* 1185–1192.

Bernhardt, B., & Major, E. (2005). Speech, language and literacy skills 3 years later: A follow-up study of early phonological and metaphonological intervention. *International Journal of Language & Communication Disorders, 40*(1), 1–27.

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

Botting, N. (2002). Narrative as a tool for the assessment of linguistic and pragmatic impairments. *Child Language Teaching & Therapy, 18*(1), 1–22.

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

Carrow-Woolfolk, E. (1995). *Oral and Written Language Scales: Listening Comprehension and Oral Expression*. Circle Pines, MN: AGS.

Carrow-Woolfolk, E. (1996). *Oral and Written Language Scales: Written Expression*. Circle Pines, MN: AGS.

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

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

Catts, H. W., Fey, M. E., Zhang, X., & Tomblin, J. B. (2001). Estimating the risk of future reading difficulties in kindergarten children: A research-based model and its clinical implementation. *Language, Speech, and Hearing Services in Schools, 32,* 38–50.

Dollaghan, C. A. (2004). Evidence-based practice in communication disorders: What do we know, and when do we know it? *Journal of Communication Disorders, 37,* 391–400.

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

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test—III*. Circle Pines, MN: AGS.
Fey, M. E., Catts, H. W., Proctor-Williams, K., Tomblin, J. B., & Zhang, X. (2004). Oral and written story composition skills of children with language impairment. *Journal of Speech, Language, and Hearing Research, 47,* 1301–1318.
Gillam, R., & Pearson, N. (2004). *Test of Narrative Language*. Austin, TX: Pro-Ed.
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*. Austin, TX: Pro-Ed.
Hresko, W. P., Reid, D. K., & Hammill, D. D. (1999). *Test of Early Language Development, 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.
Law, J., & Durkin, C. (2000). The literacy skills of language-impaired children: Time for 'joined up' thinking? *Educational Psychology in Practice, 16*(1), 75–87.
Law, J., & Sivyer, S. (2003). Promoting the communication skills of primary school children excluded from school or at risk of exclusion: An intervention study. *Child Language Teaching & Therapy, 19*(1), 1–25.
Lazzari, A. M. (1996). *The Help Test*. East Moline, IL: LinguiSystems.
Liles, B. Z., Duffy, R. J., Merritt, D. D., & Purcell, S. L. (1995). The measurement of narrative discourse ability in children with language disorders. *Journal of Speech and Hearing Research, 38,* 415–425.
McCauley, R. J., & Swisher, L. (1984a). Psychometric review of language and articulation tests for preschool children. *Journal of Speech and Hearing Disorders, 49,* 34–42.
McCauley, R. J., & Swisher, L. (1984b). Use and misuse of norm-referenced tests in clinical assessment: A hypothetical case. *Journal of Speech and Hearing Disorders, 49,* 338–348.
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.
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.
Reilly, J. S., Bates, E. A., & Marchman, V. A. (1998). Narrative discourse in children with early focal brain injury. *Brain & Language, 61,* 335–375.
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.
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.
Shulman, B. B. (1986). *Test of Pragmatic Skills—Revised*. Tucson, AZ: Communication Skill Builders.
Simkin, Z., & Conti-Ramsden, G. (2001). Non-word repetition and grammatical morphology: Normative data for children in their final year of primary school. *International Journal of Language & Communication Disorders, 36,* 395–404.
Spaulding, T. J., Plante, E., & Farinella, K. A. (2006). Eligibility criteria for language impairment: Is the low end of normal always appropriate? *LanguageSpeech, and Hearing Services in Schools, 37,* 61–72.
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.
Tomblin, J. B., Records, N. L., & Zhang, X. (1996). A system for the diagnosis of specific language impairment in kindergarten children. *Journal of Speech and Hearing Research, 39,* 1284–1294.
Urwin, S., Cook, J., & Kelly, K. (1988). Preschool language intervention: A follow-up study. *Child: Care, Health & Development, 14*(2), 127–146.
Wallace, G., & Hammill, D. D. (2002). *Comprehensive Receptive and Expressive Vocabulary Test—Second Edition*. San Antonio, TX: The Psychological Corporation.
Wiig, E. H., & Secord, W. (1989). *Test of Language Comprehension—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: AGS.
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 Scale, Fourth Edition*. San Antonio, TX: The Psychological Corporation.

Received April 11, 2005
Revision received July 20, 2005
Accepted February 20, 2006
DOI: 10.1044/1058-0360(2006/023)

Contact author: Elizabeth Peña, 2504-A Whitis, Room 7.214, University of Texas at Austin, Austin, TX 78712.
E-mail: lizp@mail.utexas.edu