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
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                                      |
| Analysis of the Language of Learning         | Blodgett & Cooper (1987)                    |
| Clinical Evaluation of Language Fundamentals—Preschool | Wiig, Secord, & Semel (1992)               |
| The Expressive Language Test                 | Huisningh, Bowers, LoGuidice, & Orman (1998)|
| The Help Test                               | Lazzari (1996)                              |
| The Language Processing Test—Revised        | Richard & Hannen (1995)                     |
| The Listening Test                          | Barrett, Huisningh, Zachman, Blagden, & Orman (1992)|
| Rice/Wexler Test of Early Grammatical Impairment | Shulman (1986)                            |
| Test of Semantic Skills—Primary             | Zachman, Huisningh, Barrett, Orman, & Blagden (1989)|
| The Word Test—Adolescent                    | Huisningh, Barrett, Zachman, Blagden, & Orman (1990)|
| The Word Test—Elementary—Revised            |                                            |
| Mixed normative group                        |                                            |
| Boehm Test of Basic Concepts—Preschool, Third Edition | Boehm (2001)                                |
| Comprehensive Assessment of Spoken Language                                         | Carrow-Woolfolk (1999a)                   |
| Comprehensive Receptive and Expressive      | Wallace & Hammill (2002)                    |
| Vocabulary Test—Second Edition              |                                            |
| Diagnostic Evaluation of Language Variation | Seymour, Roepner, & deVilliers (2003)       |
| Expressive Vocabulary Test                  | Williams (1997)                            |
| Peabody Picture Vocabulary Test—III         | Dunn & Dunn (1997)                          |
| Test of Auditory Comprehension of Language—Third Edition | Carrow-Woolfolk (1999b)                   |
| Test of Language Development—Intermediate | Hamill & Newcomer (1997)                   |
| Test of Language Development—Primary, Third Edition | Newcomer & Hammill (1997)               |
| Test of Word Knowledge                      | Wiig & Secord (1992)                       |
Normal cases had a mean of 247 with a standard deviation (SD) of 250. This cutoff was selected because it appeared in the Literature Impairment (LI) = language impairment. Using 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). 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 SDs 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 SDs 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...
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
TABLE 3. Classification accuracy of pretest, modifiability, and posttest measures.

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

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 under-

²Note 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.
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, 1998) 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


Peña et al.: Composition of Normative Groups

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.


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: lzp@mail.utexas.edu

254 American Journal of Speech-Language Pathology • Vol. 15 • 247–254 • August 2006

Downloaded From: http://ajslp.pubs.asha.org/ by a University of Texas, Austin User on 05/20/2015
Terms of Use: http://pubs.asha.org/ss/Rights_and_Permissions.aspx