USE AND MISUSE OF NORM-REFERENCED TESTS IN CLINICAL ASSESSMENT: A HYPOTHETICAL CASE

REBECCA J. McCANLEY    LINDA SWISHER
University of Arizona

The purposes of this paper are to discuss concepts that are fundamental to the proper use of norm-referenced tests in clinical assessment, to consider common errors in the use of such tests, and to suggest alternatives to norm-referenced testing for certain assessment purposes. A hypothetical client is used to illustrate the following errors: the use of age-equivalent scores as the sole summary of test results, the use of individual items to formulate therapy objectives, and the failure to consider the possible effects of measurement error when difference scores are used to assess progress or to examine patterns of impairment.

Assessment, the application of numbers to a performance, is a major component of clinical evaluation, a larger process that requires the integration of information from many sources. The proper use of norm-referenced tests in assessment entails a series of decisions and activities on the part of the speech-language clinician. First, the clinician decides whether or not a norm-referenced instrument is appropriate for a particular assessment purpose, and then which specific test or group of tests of the skill being assessed is psychometrically most acceptable (Buros, 1972, 1978; Darley, 1979; Kilburg, 1982; Launer & Lahey, 1981; McCauley & Swisher, 1984; Weiner & Hoock, 1973), and which reports norms relevant to the client. Next, the clinician administers and scores each test in accordance with the procedures outlined by the test developer. Prior to integrating test results with other sources of information during evaluation, the clinician interprets test results in light of the relevant test norms and the assessment question to arrive at a statement of the test taker’s performance.

Norm-referenced tests provide evidence regarding the existence of a problem. Properly used, they can suggest a need for further assessment or help document a need for the initiation or continuation of therapy. Misused, they can lead to a mistaken understanding of a client’s problem, to inappropriate and fruitless therapy programs, or to inaccurate conclusions regarding the efficacy of therapy. The purposes of this paper are to discuss concepts that are fundamental to the proper use of norm-referenced tests, to consider four common errors in the use of such tests, and to offer alternatives to norm-referenced testing for certain assessment purposes.

CONCEPTS UNDERLYING THE USE OF NORM-REFERENCED TESTS

A brief review of some concepts pertinent to the comparison of a client’s score to the test norms is included to highlight some of the characteristics of norm-referenced tests that make them susceptible to misuse. These concepts are related to psychometric principles that are treated more fully elsewhere (e.g., Anastasi, 1976; Salvia & Ysseldyke, 1981). Such concepts are usually discussed in relation to test construction and the psychometric evaluation of norm-referenced tests; however, an appreciation of them is equally vital to the proper use of such tests.

Although the comparison of a test taker’s score to the relevant norms may at first glance appear to be a straightforward matter, it is complicated by the fact that tests only estimate the underlying skill or behavior they are designed to assess. This means that the comparison of the test taker’s score and the relevant norms involves a comparison of estimated, rather than absolute, or true values. The comparison process is further complicated by the fact that individuals with similar characteristics will vary to some extent in their test performance without such variability being indicative of a problem. These two complications, then, make the comparison of an individual’s score to the relevant test norms a fairly difficult enterprise.

The first of these complications—the fact that test scores are imperfect estimates of whatever they were designed to measure—can be more fully appreciated and responded to following a discussion of the concept of test reliability. Reliability can be defined as the consistency with which a test measures a given attribute or behavior. It can be estimated for a test by examining the consistency of scores obtained by individuals when they are given the same test on two occasions, an equivalent form of the test on the second occasion, or when they are tested under other, variable conditions (Anastasi, 1976, p. 103).

For the clinician who wishes to compare a single test score to the norms, the imperfect reliability of even the most reliable test means that the test taker’s observed score will vary from the score the test taker would receive if the test’s reliability were perfect. That ideal score is referred to as the individual’s true score. When evidence collected by the test designer indicates high reliability, there is probably little difference between the individual’s true and observed scores. When reliability is
low, however, there may be a considerable discrepancy between those scores.

In everyday speech, it is not uncommon to hear a statement such as, "He's probably 50 years old—give or take a few years." With such a statement, the speaker attempts to convey a best estimate of the attribute under consideration, as well as information about the confidence with which the estimate is made. Comparable qualifications are necessary for the more formal estimates produced by standardized tests. In order to qualify the information conveyed by a test score, one calculates an estimated true score (the best estimate) and a confidence interval (the expected reliability of that estimate) based on a statistic known as the standard error of measurement (SEM). The calculation of the test taker's estimated true score and the confidence interval around it requires the following pieces of information: (a) an estimate of the test's reliability, (b) the mean and standard deviation of scores obtained by the normative sample to which the test taker's score is to be compared, and (c) the test taker's observed score (Nunnally, 1978, p. 240; Salvia & Ysseldyke, 1981, pp. 86–94). When all of these pieces of information are available—and often they are not (McCaulley & Swisher, 1984)—they can be used to estimate the standard error of measurement and, then, the confidence interval in which one would expect to find the test taker's true score. As a result of these calculations, one knows considerably more about the precision of the test's measurement than from the observed score alone.

Thus, one can estimate the degree of uncertainty that results from the test's imperfect reliability. In addition, this information allows one to interpret test performance as a range of scores rather than as a specific value (American Psychological Association, 1974, p. 69).

An example of the calculations described above is included in the appendix. [Calculations that differ from those described by Salvia and Ysseldyke can be found in Guilford (1954, p. 389) and Anastasi (1976, p. 128).] In the example, a client receives a raw score of 75 on a test for which a relevant normative sample received a mean score of 100 with a standard deviation of 10. A reliability coefficient reported for the same normative sample was .70. This less than ideal, but not uncommon, level of reliability is used to illustrate the loss of precision resulting when tests of even moderate reliability are used. Given these pieces of information, the test taker's estimated true score is 82.5 and the 95% confidence interval around the true score extends from 71.76 to 93.24. Thus, 95 times out of 100, the test taker's true score can be expected to fall between the values of 71.76 and 93.24 and the single best guess one can make is 82.5.

The second complication that can muddle the comparison of a test score to the test norms is that individuals will vary in their abilities, and hence in their test scores, without that difference signaling the existence of a problem. At what point is a score different enough from the range of scores received by the normative sample to suggest a possible problem? Is a 10-point difference from the sample mean large enough? Is a 20-point difference large enough? The concepts of standard deviation and z scores can be used to help answer these questions.

The standard deviation is a measure of the variability of the scores obtained by individuals in the normative sample. It is used in conjunction with the mean and serves to summarize the degree to which test scores cluster around the mean or are scattered over a wide range of values. The adequacy of the mean score as a description of a set of scores is affected by the variability of those scores. If most test scores in the group lie close to the mean, the mean is a more representative value than if most scores lie a considerable distance from the mean.

Consider the following lines which represent two pairs of scores and their mean values:

<table>
<thead>
<tr>
<th>Pair 1</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>35—41—47</td>
<td>41</td>
</tr>
</tbody>
</table>

The mean for each pair of scores is 41. However, despite the fact that both pairs of numbers can be summarized by that same value, the scores 35 and 47 are much more similar to the mean of 41 than are the scores 12 and 70. The standard deviations for these two sets of numbers reflect this difference: The first pair has a standard deviation of 6, whereas the second has the much larger standard deviation of 29. Thus, the larger the standard deviation, the greater the variability of scores around the mean.

Z scores are considered useful because they incorporate information about the standard deviation, as well as about the mean. For example, a score equivalent to the mean would have a z score of 0, and scores 1 standard deviation above or below the mean would have z scores of +1.00 or −1.00, respectively. Z scores can be used to estimate the probability of a given score's occurrence because sets of test scores usually approximate a normal distribution around the mean. One property of normal distributions is that a predictable percentage of values (test scores) will fall within a given distance from the mean value, when the distance is expressed as a multiple of the standard deviation. By converting an individual's raw test score into a z score, one can estimate the percentage of the normative sample that received a test score at least that different from the mean. This permits the test user to estimate how many scores of a particular magnitude should be found and, thus, how unexpected a given score is. For example, scores associated with z scores of less than −2 (i.e., those lying more than two standard deviations below the mean) can be expected to be received by less than 3% of the population and are, therefore, considered by some clinicians to be indicative of nonnormal performance (Ludlow, 1983).

A particularly informative way in which this might be done is shown in Figure 1. In that figure, the scores used in the discussion of confidence intervals have been transformed to z scores, and the confidence interval has been calculated. These values have been plotted along a normal probability curve. The range of values included within the confidence interval illustrates the dangers of taking a raw score at face value. In comparisons of a test taker's score to a cutoff value (such as two standard deviations below the mean or 1.65 standard deviations...
below the mean), the estimated true score and the width of the confidence interval around it should affect the test user's interpretation. Because the estimated true score takes the reliability of the test into account, it is the single best value to compare to the arbitrary cutoff value. The width of the confidence interval around it indicates how much above or below the estimated true value, and by inference how much above or below the cutoff value, the actual true score might be.

In summary, the proper interpretation of a test taker's score on a norm-referenced test requires use of information on the reliability of the test taker's score and that score's probability for the normative sample. Several of the errors in test use that will be discussed in the course of this paper arise when the importance of these pieces of information is not appreciated.

**SPECIFIC PROBLEMS IN THE USE OF NORM-REFERENCED TESTS**

Examples of the four errors in the use of norm-referenced tests that are considered in the course of this paper are illustrated in the hypothetical case that appears in the Exhibit. The case involves a preschool child referred because of a possible language impairment. Although the report refers primarily to the use of language tests, it illustrates errors to be avoided whenever a norm-referenced test of any behavior is used. In the hypothetical case, the errors are presented in the order in which they will be discussed in the following sections. This exhibit contains only the background and assessment sections of a case report. The evaluation section, in which information from different sources would be combined, is excluded from the exhibit.

**AGE-EQUIVALENT SCORES AND THE SUMMARIZATION OF TEST RESULTS**

Age-equivalent scores, such as language ages, are often reported in test manuals when the skills to be assessed change considerably with age during childhood. They are frequently used by clinicians to explain test results to nonprofessionals, such as the parents of a tested child, because it is assumed that they summarize the test results in a form that is easily understood.

In the hypothetical case report (see the Exhibit, Error #1), an age-equivalent score was used to summarize Paul's performance on test B and to support the assertion that Paul's receptive vocabulary was impaired. In spite of the frequency with which age-equivalent scores are used in this manner, there are problems associated both with the use of age-equivalent scores as a summary of test results and with the interpretation of that summary. Several of these problems will be discussed in the section that follows.

**Error 1: Using Age-Equivalent Scores as Test Summaries**

A sizable number of problems with the use of grade- or age-equivalent scores have been identified by various authors (Anastasi, 1976, pp. 73–76; Mehrens & Lehmann, 1980, pp. 97–99; Salvia & Ysseldyke, 1981, pp. 66–69). Generally, these problems fall into two categories: those that cause such scores to be considered "psychometrically crude" (Anastasi, 1976, p. 74) and those that lead to mistaken inferences.

Two psychometric problems limit the reliability and, therefore, the validity of age-equivalent scores and, thus, cause such scores to be considered "extremely vague indications of developmental level" (Salvia & Ysseldyke, 1981, p. 73). The first psychometric problem concerns the relation of age-equivalent scores and the raw scores on which they are based (Salvia & Ysseldyke, 1981, p. 67). For most tests, as age increases, similar differences in age-equivalent scores are the result of smaller and smaller differences in raw scores. Therefore, an age-equivalent score that is 6 months behind an individual's chronological age may indicate a larger difference in actual test performance for younger test takers than for older ones. This happens because individuals at higher developmental levels, who are quickly approaching adult competence, are making fewer errors. Because of this, the 1-year delay on test B reported for Paul, who is 4 years old, may have resulted from 12 missed items; whereas a 1-year delay for a 10-year-old child taking the test might be due to only one or two missed items. This causes the reliability of age-equivalent scores to be poorer for developmentally more advanced test takers.

A second psychometric problem with age-equivalent scores is that they are not necessarily based directly on evidence collected for children of that chronological age (Mehrens & Lehmann, 1980, pp. 97–98; Salvia & Ys-
EXHIBIT: Sections of a hypothetical case report. The numerals correspond to test use errors and to the numbered sections in the text.

Client: Paul B.  
D.O.B.: 5/27/78  
Parents: John & Mary Ann B.  
Date: 1/20/83  
Address: 1121 Buena Vista  
Clinic: R. F.  

Background

Initial interview, Paul B., age 4:2, was first seen at this office on 7/23/82 for evaluation. His parents had been referred by a teacher at Paul’s preschool. Both parents reported that Paul did not talk as much as his older brother and sister did at his age, and that when he talked, he used mostly two- and three-word utterances. His parents volunteered that they considered Paul a “bright” child and that they were satisfied with his development in other areas. Information from other sources . . . . In summary, then, the nonverbal intelligence and audiometric assessments conducted prior to our initial evaluation indicated normal function.

Initial Speech and Language Assessment

Norm-referenced assessment. Three norm-referenced tests were administered: a test of articulation (test A), a test of single-word receptive vocabulary (test B), and a test of receptive and expressive syntax (test C).

Test A . . . . Therefore, Paul’s performance on test A indicated normal test performance.

Test B. Paul received a score of 45 on test B. This corresponded to a language age of 3:2, indicating a delay in receptive vocabulary of 1 year when compared with his chronological age of 4:2 . . . .

Test C . . . . On the 36-item subtests of test C, Paul received a score of 30 on receptive subtest #1 compared with his score of only 25 on expressive subtest #2. These scores corresponded to z scores of –2.00 and –2.40, respectively. Thus, it appeared that Paul exhibited a greater deficit in expressing than in understanding the syntax of language . . . . Because of Paul’s difficulty with the plural –s on test C, these structures were included in the initial therapy objectives . . . .

Current Assessment Information

Results obtained when tests B and C were readministered indicated little progress. Thus, therapy does not appear to have been effective for those areas.

McCauley & Swisher: Norm-Referenced Tests and Assessment 341

seldyke, 1981, p. 68). Instead, a given age-equivalent score is often calculated indirectly either by interpolating between two ages for which data are available or by extrapolating from ages for which data are available to older or younger ages, for which data were not gathered. For example, Paul’s raw score on test B probably fell between the average scores of the 3-year-olds and 4-year-olds in the normative sample, and he was assigned the age-equivalent of 3:2 by interpolation although no children of age 3:2 were sampled in the normative studies. Dependence on indirectly estimated (interpolated or extrapolated) scores, therefore, may involve assumptions about the continuity of language development across chronological age that cannot be justified. Thus, extrapolated and interpolated age-equivalents are likely to be less stable estimates of the test taker’s behavior than other kinds of derived scores, such as z scores (Mehrens & Lehmann, 1980, p. 98).

Because of their psychometric problems, age-equivalent scores often serve as the basis for misinterpretations. Two inferences which foster such misinterpretations occur with sufficient frequency to warrant attention here. One inference is that current functioning is accurately described by an age-equivalent description (Mehrens & Lehmann, 1980, p. 98 Salvia & Ysseldyke, 1981, pp. 67-68). Someone faced with the hypothetical case of Paul and his age-equivalent score of 3:2 on a vocabulary test might infer that he functions like a 3-year-old in his understanding of single words. This inference is not well-founded, however, because there is as yet no evidence that a 4-year-old—who has a year more experience with the world and with language than the younger children on whom the normative data were obtained—approaches a language test in the same way or makes the same kind of errors as a younger child (Mehrens & Lehmann, 1980, p. 98; Salvia & Ysseldyke, 1981, p. 68; Thorndike, 1982, p. 108). Similarly, a 60-year-old suffering from aphasia might receive an age-equivalent score of 10 years on the vocabulary test. It is unlikely, however, that such a client would make the same kind of errors as the 10-year-old or that he would exhibit similar communication skills.

Perhaps the most perilous inference in the interpretation of age-equivalent scores involves their use to indicate impairment (Anastasi, 1976, p. 76; Bloom & Lahey, 1978, p. 352). Here the inference is made that an age-equivalent score that is a certain number of months lower than chronological age necessarily indicates a delay or even an impairment for the tasks addressed by the test. Although there are minor differences in the way in which test developers calculate age-equivalent scores, such scores do not take into account the normal individual differences to be expected within any of the age groups. Thus, they do not provide evidence as to whether a child who scores below the average obtained by his age peers did so by an amount that represents normal variation or by an amount that is sufficiently unexpected to suggest impairment. If considerable variation is common for normal children within an age group, even a large age-equivalent delay may not imply impairment.

A graphic illustration may help make this point in relation to the age-equivalent score used to summarize Paul’s performance on test B. Figure 2 shows two idealized curves that represent test scores received by a large number of 3- and 4-year-olds on the test. The range of scores shown by children in each group is indicated by the width of the bell-shaped curve. Normal variation resulted in some overlap in the scores received by the two groups. Paul—and all of those 4-year-olds whose scores fall in the shaded area of the curve—would be said to have received an age-equivalent score of 3 years or less. However, when the variability shown by the 4-year-olds as a group is taken into account using z scores, the same score may not be found to be sufficiently different.
from the average obtained by this group to suggest impairment. Certainly, if a $z$ score better than $-2.0$ were considered within the normal range, almost all of the scores falling in the shaded area on the figure would be considered normal, despite their correspondence to age-equivalent delays of up to a year or more. Thus, age-equivalent delays do not necessarily indicate delay or impairment.

In summary, age-equivalent scores are psychometrically imprecise. In addition, they may encourage misinterpretation by any but the most cautious consumer, and they may be particularly misleading to the nonprofessionals to whom they are frequently presented.

**Suggestions Regarding the Proper Use of Age-Equivalent Scores**

Summarizing test results with standard scores (e.g., $z$ scores) or with percentile ranks is recommended by the American Psychological Association (1974, p. 23) as an alternative to the use of developmental scores such as age-equivalent scores.

Unfortunately, the calculation of these standard scores is not always possible because it requires access to the mean and standard deviation of the scores obtained by a relevant normative sample. Despite its importance, this information is often missing from speech and language test manuals (McCaulley & Swisher, 1984). Clinicians may wish to seek out tests that include this information and to avoid tests for which only age-equivalent scores are available.

If age-equivalent scores are used, awareness of their limitations is an important step toward minimizing errors in interpretation. It has been suggested that age-equivalent scores be used in conjunction with other, better descriptions of test results, such as $z$ scores or percentile ranks (Mehrens & Lehmann, 1980, p. 99). Their misinterpretation can further be avoided if age-equivalent scores are accompanied by a thorough description of the reasonable inferences they support.

**Profiles and the Description of Pattern of Impairment**

Profiles are typically constructed by graphing two or more different test scores for the same individual using the same unit of measurement [e.g., $z$ scores (Salvia & Ysseldyke, 1981, p. 479)]. Because such profiles consolidate test information, they may be used by clinicians to facilitate visual comparisons of test scores and to extract information about an individual’s pattern of impairment. Their use sometimes involves the mistaken assumption that a seemingly significant difference in test scores indicates a corresponding difference in associated behaviors.

The discussion of Paul’s performance on the two subtests of test C (see the Exhibit, Error #2) is an example of the way in which test profiles are often interpreted. Figure 3 is a hypothetical profile that could be constructed from Paul’s performance on test C. In the example, Paul’s lower score on subtest 2 than on subtest 1 was considered to be an indication that he was more impaired on behaviors assessed by subtest 2. However, unless steps are taken to demonstrate that the difference between Paul’s two scores is probably not due to chance, this interpretation cannot be justified. Because such steps are relatively time consuming, are not commonly appreciated, and are not always possible, they are often not taken.

**Error 2: Profiles as Descriptions of Pattern of Impairment**

The complications that make profile analysis time consuming are closely related to the complications accompanying the comparison of an individual’s test score to the norms. The scores to be compared in a profile are only estimates of the ideal or true scores one would obtain if the scores were free from error, just as a single score and the test norms are only estimates of the values one would
obtain if the scores were free from error. Because error contributes to the magnitude of each score, it can affect the size of the difference between pairs of scores. Therefore, two scores within a test profile can seem quite different from one another due to measurement error rather than to real differences in the behaviors being measured.

To emphasize the possible effect of misinterpretation in profile analysis, Salvia and Ysseldyke (1981) wrote that “simplistic and unsophisticated profile analyses . . . are likely to be interpretations more of error than of real differences” (p. 499). In other words, one can jump to the wrong conclusion about an individual’s relative strengths and weaknesses by assuming that all apparent differences in test scores represent real differences in behavior. Other writers (Anastasi, 1976, pp. 129–131; Mehrens & Lehmann, 1980, pp. 56–57, 101–105) make somewhat different suggestions regarding the procedures to be used in profile analysis. Their suggestions, however, share the common goal of avoiding the mistaken interpretation of error as indicative of real differences in behavior.

Suggestions Regarding the Use of Profiles

Salvia and Ysseldyke (1981, pp. 483–487) discuss at length three steps to avoid simplistic profile analysis. These steps will be described as they relate to scores obtained on different tests, but similar procedures could be used to compare subtests of the same test instrument.

As the first step in a complete profile analysis, the test user determines whether the observed differences between two scores, called a difference score, is likely to be reliable. That is, if the tests (or subtests) were readministered, would the difference between test scores tend to remain the same or would it change? Answering this question requires the calculation of the standard error of measurement of the difference (SEMdiff) between the two scores to be compared. That value is comparable to the standard error of measurement for a single score and can be used to determine whether the observed difference was likely to have occurred by chance. The required calculations make use of information about the reliability of each test (subtest) and about the intercorrelation of each pair of tests (subtests) for which scores will be compared. Because the calculations are too lengthy to summarize here, the reader is encouraged to consult the detailed examples provided by Salvia and Ysseldyke (1981, pp. 487–505).

If results of the first step indicate the existence of a reliable difference, the test user then undertakes the second step in the analysis. In this step, one examines the characteristics of the normative sample used for each test (subtest) to see if differences in the characteristics of the two samples can explain that reliable difference. The following is an extreme example in which differences between the normative samples are largely responsible for a reliable difference in test scores. Suppose that Paul had taken two receptive vocabulary tests which were identical except that one had been standardized on children who had participated in an academic enrichment program for gifted children and the other had been standardized on children who were acquiring English as a second language. When Paul is compared to these two groups, we might expect him to perform poorly when compared to the first group of children but well when compared to the second group. Hence, a reliable, and possibly large, difference between the z scores obtained by Paul on the two tests might be found, yet the difference would be due to the normative samples, not to real differences in Paul’s vocabulary skills from one testing to the next.

The second step described above is not needed for profiles of subtest scores which make use of a single normative sample (Salvia & Ysseldyke, 1981, p. 504) unless there is some evidence that different children were used in the standardization of different subtests. In actual practice, one may wish to examine differences in normative samples as the first step in profile analysis in order to avoid determining the reliability of an observed difference only to find it is probably due to differences between the normative samples.

The third step in profile analysis—concluding that a difference between two scores suggests a real difference in underlying skills—is taken only if the reliable difference between two test scores does not seem to be explained by differences in normative samples. Many so-called profile analyses include only this last step. However, this step is better regarded as a possible outcome of the procedures described above than as an automatic conclusion that can be reached through visual inspection of the profile alone.

Despite its usefulness, the test user will frequently be unable to perform a profile analysis because of the absence of information on reliability or the intercorrelation of the test instruments. As a conservative, but reasonable and relatively easy alternative to the complete profile analysis discussed by Salvia and Ysseldyke, one might confine observations about patterns of impairment to observations regarding the presence or absence of impairment in different areas.

Test Items and the Planning of Therapy Objectives

Clinicians sometimes base therapy objectives on an analysis of errors a client has made on a norm-referenced test. This practice is probably motivated in part by a desire to make the most of already collected information. It is almost certainly based on the premise that missed items indicate significant and real areas of weakness.

In the Exhibit (Error #3), Paul’s performance on the plural -s on test C provided the rationale for including those structures in initial therapy objectives. Unfortunately, the use of this kind of analysis may lead the clinician to miss important deficits while focusing therapy objectives on less important or even falsely identified deficits.
Error 3: Performance on Individual Test Items as Indications of Deficit

To begin with, norm-referenced speech and language tests do not describe all skills that might need to be addressed in therapy (Leonard, Prutting, Perozzi, & Berkley, 1978; Miller, 1983; Schery, 1981; Snyder, 1983). Because of gaps in the skills assessed by available tests, the use of norm-referenced tests to plan therapy precludes the identification of problems in many areas.

Using the results of test items to plan therapy objectives also presents grave difficulties for skills that are covered by existing tests. First, the relatively small number of items included in a norm-referenced test cannot adequately sample all of the specific forms and developmental levels that might be appropriate (McLean & Snyder-McLean, 1978, p. 129). For example, language tests may include no items or only a small number of items for a given form (Owens, Haney, Giesow, Dooley, & Kelly, 1983) or a given developmental level (Leonard et al., 1978). Norm-referenced tests, after all, are designed to function in somewhat the way that headlines do in a newspaper—they give the user the quickest possible summary of a story (e.g., Paul B. May Have Language Problem), but they fail to tell the story itself.

An example may emphasize the potential impact of this lack of detail on therapy planning. Suppose Paul had acquired the plural -s morpheme in some developmentally early contexts, but not in later ones. This pattern might not be evident from his performance on test C which would only include a few such items, almost all of which might be developmentally similar. As this example illustrates, important information about the scope of impairment and the nature of existing skills will usually be missing from the results of a norm-referenced test.

A second, related problem with planning therapy objectives from test items is that this approach can result in an inaccurate conceptualization of the individual’s impairment. A single error on a standardized test may have any of several causes ranging from the test taker’s momentary lapse in attention to lack of mastery of the skill tapped by that item. Similarly, correct answers on individual items may result from an age-appropriate mastery of the linguistic structure being probed or from a potentially unproductive familiarity with the structure in an isolated context. A correct answer may even result from a lucky guess for certain kinds of tests [e.g., receptive tasks with a limited number of alternative responses (Burns, 1978)]. In short, individual errors represent only possible deficits and correctly answered items only possible competencies. Only multiple observations of similar phenomena cause an individual’s performance on the test as a whole to be less subject to alternative explanations.

A third problem that limits the utility of norm-referenced tests in planning therapy is that the tasks included in such tests assess behaviors only within a very restricted range of communicative contexts (Bloom & Lahey, 1978, pp. 313–314; Danwitz, 1981; Muma, Lubinski, & Pierce, 1982). Restrictions on nonlinguistic contexts and the purpose of communication can result in an incorrect characterization of a child’s functional communication skills (Bloom & Lahey, 1978, p. 492; Launer & Lahey, 1981; Prutting, Gallagher, & Mulac, 1975).

A fourth problem with planning objectives from errors on specific test items concerns the typical scoring systems used for speech and language tests. With a few exceptions (e.g., Porch, 1967, 1974), most scoring systems fail to provide subtle descriptive clues which, in sufficient numbers, would prove useful in therapy planning (McLean & Snyder-McLean, 1978, p. 131). Description of an individual’s responses—both correct and incorrect—may help differentiate between different kinds or degrees of impairment. For instance, it might be useful to distinguish errors indicating response bias from those indicating the client’s mastery of some aspect of the tested skill or a complete absence of the tested skill. Similarly, it might be desirable to separate correct responses that are immediate from those that are delayed. This information would prove valuable to the clinician, who could incorporate in early therapy objectives a behavior that had been associated with self-corrected responses rather than one that had evoked no observed response.

When test items are translated too literally into therapy objectives, their use may also result in a fifth problem—the invalidation of the norm-referenced test as a test of ability. For example, subsequent administrations of a vocabulary test to a child who has been drilled on items from that test will be almost useless in indicating the child’s general receptive vocabulary skills.

Because of the problems noted above, the use of test items in the planning of therapy may often result in an impoverished and potentially inappropriate set of objectives. There is probably no circumstance in which norm-referenced test items can profitably be used for this purpose.

Suggestions Regarding Tests and Therapy Objectives

As argued above, the determination of therapy objectives requires a more comprehensive description of a client’s current level of skills than is typically obtained from a norm-referenced test. In the case of a developing child, this detailed description needs to be made with reference to stages of normal development in order to determine which skills are emerging and, thus, which are appropriate therapy objectives (McLean & Snyder-McLean, 1978, p. 253). For language deficits, a detailed description would address both the expression and comprehension of language in both spoken and written form. At present, however, available data permit a detailed description of only some of the important components of expressive, spoken language. These data have been used to develop procedures to analyze spontaneous language samples (Bloom & Lahey, 1978, pp. 445–488) and to choose the content of criterion-referenced tests (e.g., McCarthy, 1974).
Systematic language sampling and analysis. Although still in need of further study, language analysis (e.g., Bloom & Lahey, 1978, pp. 445–488; Lee, 1974; Lee & Canter, 1971; Tyack & Gottlebeben, 1974) is an important clinical tool for the description of children's expressive language. The purpose of such procedures is to provide a detailed description of the language used by the client in relation to the normal stages of language acquisition. This detailed description offers a fertile source of suitable therapy objectives.

Although the virtues of expressive language sampling are everywhere praised, the procedures themselves may be infrequently used in some settings. Two recent surveys (Kilburg, cited in Kilburg, 1982; Muma, Pierce, & Muma, 1983) suggest that language sampling is currently more popular among ASHA-accredited educational institutions than it is among ASHA-accredited clinics. Kilburg described the results of a national survey in which she asked speech-language pathologists in ASHA-accredited speech-language clinics what kind of language measures they preferred and which methods they actually used in the assessment of children up to 3 years of age. She noted that although most respondents preferred the analysis of spontaneous behavior, less than one third of them reported that they actually made use of that kind of analysis.

Muma, Pierce, and Muma (1983) conducted a survey of ASHA-accredited educational institutions and found that language sampling and analysis procedures were the most frequently used tools in the assessment of child language among 76 programs. Future surveys will be needed to determine what accounts for the differences between these two surveys. For example, are apparent differences in the frequency of language sampling in these two settings due to differences in survey methods, in the education of the two groups of respondents, or in time constraints that might prove more compelling in the clinical setting observed by Kilburg than in the setting observed by Muma and his coauthors?

The reportedly limited use of language analysis methods in clinical settings may stem from the relatively lengthy and complex procedures they require. Nonetheless, in light of the meager information provided by norm-referenced tests for the planning of therapy, the value of language analysis outweighs the difficulties it presents. Furthermore, recent advances in computer-assisted language analysis (e.g., Hixson, 1983; Mordeciai & Palin, 1983) may dramatically reduce these practical difficulties.

Although the concepts of reliability and validity may be more easily overlooked for language analysis procedures, these concepts nonetheless apply to any behavioral measure. For example, language analysis has been shown to be sensitive to the child’s activity while the sample is collected (Longhurst & Grubb, 1974), as well as to the identity of the speaker’s interlocuter (Olswang & Carpenter, 1978). Therefore, this assessment method, as others, requires additional study.

Criterion-referenced tests. Whereas norm-referenced tests are designed to compare the test taker’s performance to that of the normative sample, criterion-referenced tests are designed to determine which skills a test taker has attained along a continuum of skills judged by the test developer to be important. Instead of being used to answer questions such as “how does this child perform compared to his age peers?,” criterion-referenced tests are used to answer questions such as “does this child have these skills?”

Although criterion-referenced tests represent a theoretical alternative for planning therapy, they do not yet present a practical alternative (Schery, 1981). To date, few criterion-referenced tests have been designed to address content areas of interest to speech-language clinicians (e.g., McCarthy, 1974). Such criterion-referenced tests have usually been developed in association with therapy programs and are limited in scope to the therapy objectives of a specific program.

At least in part, the limited number of available criterion-referenced tests is due to the fact that such tests are difficult to develop. They require a detailed description of the content area they are meant to assess—a requirement that can lead to a cumbersome instrument that might not gain wide acceptance. From a psychometric point of view, they present difficulties to the test designer because of ongoing controversies about appropriate standardization methods [e.g., the appropriate way in which to determine reliability for such instruments (Carver, 1974; Millman, 1979; Proger & Mann, 1973)]. Nonetheless, because norm-referenced tests provide incomplete and possibly misleading information for the formulation of language objectives, language analysis and the development of criterion-referenced tests require increased clinical and research attention.

Multiple Test Scores and The Assessment of Progress

Spurred on by demands for increased accountability (e.g., Caccamo, 1973; Mowrer, 1972), some clinicians have begun using periodic administrations of norm-referenced tests to document the effects of therapy. The rationale behind this practice is presumably that test instruments that are appropriate for the identification of impairment must be equally appropriate for the measurement of progress. Their use seems to allow clinicians to adapt measures that they consider objective and manageable to the troublesome task of validly assessing progress.

In the hypothetical case (see the Exhibit, Error #4), tests B and C were readministered after a period of therapy for the purpose of measuring Paul’s progress in therapy. In spite of clinical observations to the contrary, no substantial change in test scores was obtained, and the clinician concluded that Paul was not making much progress. Despite its apparent efficiency, however, this practice can result in spurious conclusions regarding changes in behavior during therapy.

Downloaded From: http://jshd.pubs.asha.org/ by a Ohio State University - Library User on 12/06/2015
Error 4: Repeated Testing as a Means of Assessing Progress

The use of norm-referenced tests to assess progress can result in the underestimation or overestimation of change. It can result in the underestimation of change because the purpose for which they are designed is largely incompatible with the assessment of progress. To allow the comparison of a child to his peers, a test needs to examine gross, relatively stable behavior patterns (e.g., single-word vs. more complex stages of production). Norm-referenced tests are typically designed with this purpose in mind. On the other hand, to examine progress in therapy, a test needs to examine the specific behaviors that will be affected by learning within as well as between major developmental stages (Salvia & Ysseldyke, 1982, p. 544; Schery, 1981). Criterion-referenced tests are more often designed for this second purpose. Because norm-referenced tests are designed to compare individuals, they are likely to be less sensitive to changes in behavior over time than criterion-referenced tests.

Guidelines used in choosing items during the standardization process differ depending upon which of these two purposes the test is intended to serve (Carver, 1974). Items that result in a broad range of performances are more appropriate for a test designed to look at differences between individuals. On the other hand, items that result in a narrow range of performances (e.g., those passed by most individuals who have had a particular course of intervention) are more appropriate for tests designed to look at progress over time. Thus, norm-referenced tests, which by definition are designed to look at differences between individuals, will almost always be composed of items that can be expected to result in a broad range of performances and, therefore, will rarely be sensitive enough to behavioral change to document progress in therapy.

Besides its potential for the underestimation of change, the use of norm-referenced test scores to assess change can result in the overestimation of progress in therapy. This occurs when positive changes in test scores, called gain scores, are automatically interpreted as progress. When any positive change, regardless of magnitude, is considered significant, an individual may seem to improve on the test without showing improved language skills in therapy sessions or everyday talking situations.

Determining the amount of change that is noteworthy requires a detailed analysis similar to that used in profile analysis. One must remember that some amount of change in test scores can occur as a result of the test’s imperfect reliability. Because gain scores can be quite unreliable (Mehrens & Lehmann, 1980, p. 57; Salvia & Ysseldyke, 1981, p. 544), even a very large difference in test scores might occur by chance, making improvement difficult to document with norm-referenced tests.

Still another reason that the use of norm-referenced tests can result in the overestimation of progress is that the client may simply “learn” the test because of repeated administrations. Also, the clinician may intentionally or unintentionally teach correct responses to some items during intervention. Using a norm-referenced test no more frequently than suggested by the test developers reduces the likelihood of invalidating later test results in this way.

Thus, because norm-referenced tests were constructed to serve a different purpose and because gain scores can be particularly unreliable, norm-referenced tests do not lend themselves to use in monitoring an individual’s performance over time. Their use can engender inflated illusions of success or unwarranted delusions of failure and can invalidate their future use as tests of skill.

Suggestions Regarding Multiple Test Scores and Assessing Progress

Repeated, relatively infrequent administrations of a norm-referenced test may be a valid component of an ongoing assessment program if the intent of readministering the test is to obtain evidence about the continued existence of an impairment rather than to obtain evidence about the amount of change observed in the test taker’s behavior. When a norm-referenced test is used in this way, it can yield indirect information about an individual’s progress. That is, if the test taker’s test performance no longer differs significantly from that of his peers, then one can infer that therapy is no longer needed.

Because the use of norm-referenced tests offers, at best, a rough and potentially inaccurate measure of behavioral change, a finer description of the effects of therapy must be found elsewhere. The assessment of behavioral change is more appropriately made with measures such as criterion-referenced tests and within-subject experimental designs.

Criterion-referenced tests. Criterion-referenced tests were discussed previously in regard to their use in the description of impaired functioning. They are also useful in the assessment of progress (Carver, 1974) because they can be designed to be especially sensitive to changes in behaviors for which therapy has been undertaken. As mentioned earlier, however, criterion-referenced tests are in relatively early stages of development and so present a theoretical rather than practical alternative.

Within-subject experimental designs. Within-subject or single-subject experimental designs, as they are also called, represent a particularly powerful method of assessing progress in therapy (Costello, 1979; McReynolds, 1983; McReynolds & Kears, 1983). These designs offer several distinct advantages over other methods. First, they can permit one to determine more directly whether or not therapy is responsible for changes in behaviors occurring during therapy. Second, because they involve the continuous collection of data, they facilitate the alteration of therapy techniques in response to client behaviors throughout therapy.

McReynolds (1974, p. 131) discussed four components of systematic procedures for training children’s language that are particularly applicable to within-subject experi-
mental designs: (a) the isolation and description of a behavior that will be the focus of therapy; (b) the specification of stimuli, responses, and procedures used in therapy; (c) the collection of data sensitive to changes in the behavior that is the target of therapy; and (d) the assessment of the effectiveness of therapy through the use of experimental control. Although none of the four components is completely alien to speech-language clinicians, the final component is perhaps the least familiar. The assessment of therapy effectiveness is achieved through the clinician’s demonstration that (a) changes in behavior observed during therapy are reliable and (b) that the therapy procedures, stimuli, situation, and so forth have been held constant.

SUMMARY AND CONCLUSIONS

In the body of this paper, four errors were considered that arise when norm-referenced tests are used in the following ways: the use of age-equivalent scores as the only summary of test results, the use of test profiles to compare an individual’s strengths and weaknesses, the use of test items to formulate specific therapy objectives, and the use of repeated testing to assess changes in client behaviors during therapy. Only one of these practices was considered a proper use of norm-referenced test results—the use of test profiles to describe a client’s strengths and weaknesses. Yet even that use was considered acceptable only when rigorous steps are taken to rule out the interpretation of error as real differences in performance. For each misuse of a norm-referenced test that was considered in this paper, however, alternative methods that entail different assessment tools or a somewhat different application of the norm-referenced test are available.

One factor contributing to the errors outlined in this paper may be an insufficient appreciation of certain essential psychometric concepts and of the proper role of norm-referenced tests in assessment. However, another is certainly the dearth of adequately developed alternatives to norm-referenced testing. In particular, criterion-referenced tests and within-subject design techniques have not received sufficient research attention. To some degree, the correct use of norm-referenced tests will depend upon the accessibility of these and other, appropriate assessment methods.

The proper use of norm-referenced tests as part of a clinical assessment cannot be regarded as a simple process that begins with the administration of the test. When the information being sought concerns the formulation of specific therapy objectives or the assessment of a client’s behavior over time, methods other than norm-referenced testing are more appropriate. When the information being sought concerns the existence of a language impairment or the assessment of an individual’s pattern of impairment, the correct use of norm-referenced tests begins with the choice of specific tests that are accompanied by the data required to obtain this information accurately. In short, the use of norm-referenced tests must be incorporated within the broader context of clinical evaluation. In that broader context, specific assessment questions are matched with appropriate tools, and assessment results are evaluated by the clinician who weighs the merits and limitations of each source of information.

ACKNOWLEDGMENTS

This paper was written while R. McCauley was supported by National Institute of Neurological and Communicative Disorders and Stroke, National Research Service Award 5 F32 NS05390 CMS and L. Swisher was partially supported by Department of Education Grant G008301459. Portions of this paper were presented at the Annual Convention of the American Speech-Language-Hearing Association in Cincinnati, 1983. The authors wish to thank Anthony DeFon, Linda Lilley, James Nation, Joseph Perozzi, and Glorjaean Wallace for their helpful comments on various versions of this manuscript. Special thanks are extended to Richard Curlee for his questions and criticisms.

REFERENCES


Received October 25, 1983
Accepted August 14, 1984

Requests for reprints should be sent to Rebecca J. McCauley, Ph.D., Communication Sciences and Disorders, The John F. Kennedy Institute for Handicapped Children, 707 North Broadway, Baltimore, MD 21205.

APPENDIX

Calculations leading to a test taker’s estimated true score and the 95% confidence interval around the estimated true score. The equations used in this example and a more detailed discussion of this topic can be found in Salvia and Ysseldyke (1981, pp. 86-92).

Example

1. Obtain the following information (if available) from the test manual and the test taker’s corrected test form:

   Test taker’s score: X = 75
   Mean score of relevant normative sample: M = 100
   Standard deviation of scores obtained by sample: SD = 10
   Reliability coefficient reported for similar sample: r = .70

2. Calculate the estimated true score (X’

   Using the following equation:

   \[ X' = M + \left( r \times \frac{S}{M - X} \right) \]

   \[ = 100 + \left( .7 \times \frac{10}{100 - 75} \right) \]

   \[ = 100 + 2.8 \]

   \[ = 102.8 \]

3. Calculate the 95% confidence interval using the following steps:

   a. Estimate the standard error of measurement (SEM) using the following equation:

   \[ SEM = SD \sqrt{1 - r} \]

   \[ = 10 \sqrt{1 - .7} \]

   \[ = 5.48 \]

   b. Calculate the lower and upper limits of the 95% confidence interval using the following equations:

   Lower limit = X’ - [1.96]\times(0.48) \[ = 102.8 - (1.96)(5.48) \]

   \[ = 91.76 \]

   Upper limit = X’ + [1.96]\times(0.48) \[ = 102.8 + (1.96)(5.48) \]

   \[ = 113.84 \]

These calculations indicate the test taker’s true score can be expected to fall between the values of 91.76 and 113.84 95 times out of 100 and that the single best guess one can make as to the test taker’s true score is 102.8.

1 This value is used to produce a 95% confidence interval.