PSYCHOMETRIC REVIEW OF NINE APHASIA TESTS

LINDA LILLEY SKENES
VA Medical Center/Medical College of Georgia
Augusta, Georgia

REBECCA JOAN MCCAULEY
John F. Kennedy Institute
Baltimore, Maryland

Nine aphasia tests are reviewed using six psychometric criteria appropriate to standardized tests. The number of criteria met by individual tests ranged from one to five, with a median and mode of three. A seventh criterion was applied to six tests that used normative data. Two of the six reviewed tests met that criterion. The need for test manuals to contain a wide range of psychometric information is discussed and, in particular, the role of subject description in test standardization is emphasized.

INTRODUCTION

Speech-language clinicians administer aphasia tests for many reasons. The reasons may differ from patient to patient and from one stage of a patient’s recovery to the next. For example, Spreen and Risser (1981) list four types of aphasia assessment: screening, diagnostic assessment, assessment for counseling and rehabilitation, and progress evaluation. These types of assessments range from comprehensive descriptions of a patient’s typical communicative behavior to an assessment of optimal performance on a narrowly defined linguistic task.

Despite the different purposes for aphasia assessment, all aphasia tests have as a common goal the valid and reliable measurement of some aspect of the patient’s communication. When the authors of a test describe that test’s psychometric characteristics, such as the procedures followed in writing and standardizing the test, they offer test users an opportunity to evaluate the extent to which the goal of valid and reliable measurement may be achieved. Further, when authors describe the procedures test users should follow in administering and interpreting the test, the authors enable test users to contribute to the attainment of that goal.
The psychometric characteristics of aphasia tests are discussed in numerous texts on aphasia (e.g., Brookshire, 1978; Davis, 1983; Eisenson, 1984; Johns, 1978; Sarno, 1981) and in several collections of test reviews (e.g., Buros, 1972, 1978; Darley, 1979). The discussions contained in these works often provide relatively critical evaluations of the tests reviewed in addition to descriptions of each test’s structure and proposed uses. When a test’s psychometric characteristics are discussed in conjunction with other aspects of the test, the importance of psychometric information can be minimized. Furthermore, aphasia tests are seldom compared in terms of their psychometric characteristics. Therefore, the purposes of this article are to evaluate nine aphasia tests in terms of seven general psychometric characteristics considered relevant to all standardized tests. This review is intended to promote discussion of the psychometric problems facing aphasia tests and to facilitate comparison of such tests in terms of selected psychometric characteristics rather than to serve as a definitive psychometric review.

METHOD

Procedures

Manuals for nine aphasia tests were reviewed using seven psychometric criteria. Six of the tests were among the 10 most frequently used assessment tools for adult language according to a 1980 survey of 76 ASHA-accredited training programs (Muma, Pierce, and Muma, 1983). Muma and coworkers also included the Revised Token Test, Peabody Picture Vocabulary Test, Rosenbek’s apraxia battery, and language sample analysis. These four methods are not traditionally viewed as primary tools for the assessment of aphasia but instead are used as supplementary measures and were not reviewed in this study (Brookshire, 1978). One of the tests examined as part of the present study, the Western Aphasia Battery (Kertesz, 1982), was not published when that survey was conducted. Although not included in the list by Muma and associates, the Functional Communication Profile and the Sklar Aphasia Scale were examined in this review because they are frequently encountered in the literature and because facilities other than training institutions may use them more often. The tests reviewed in this study and their abbreviations are listed in Table 1.

Psychometric Criteria

The seven criteria used in this review were selected from every major division of the standards for test developers proposed by the American Psychological Association (APA) (APA, 1974). These seven criteria are among those listed by the APA as “essential standards” and specify the
Table 1. Reviewed Tests and Their Abbreviations

<table>
<thead>
<tr>
<th>Test Description</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aphasic Language Performance Scale (ALPS)</td>
<td></td>
</tr>
<tr>
<td>(Keenan and Brassel, 1975)</td>
<td></td>
</tr>
<tr>
<td>Boston Diagnostic Aphasia Examination (BDAE)</td>
<td></td>
</tr>
<tr>
<td>(Goodglass and Kaplan, 1972)</td>
<td></td>
</tr>
<tr>
<td>Communicative Abilities in Daily Living (CADL)</td>
<td></td>
</tr>
<tr>
<td>(Holland, 1980)</td>
<td></td>
</tr>
<tr>
<td>Examining for Aphasia (EFA)</td>
<td></td>
</tr>
<tr>
<td>(Eisenson, 1954)</td>
<td></td>
</tr>
<tr>
<td>Functional Communication Profile (FCP)</td>
<td></td>
</tr>
<tr>
<td>(Sarno, 1969)</td>
<td></td>
</tr>
<tr>
<td>Minnesota Test for Differential Diagnosis of Aphasia (MTDDA)</td>
<td></td>
</tr>
<tr>
<td>(Schuell, 1972)</td>
<td></td>
</tr>
<tr>
<td>Porch Index of Communicative Ability (PICA)</td>
<td></td>
</tr>
<tr>
<td>Sklar Aphasia Scale (SAS)</td>
<td></td>
</tr>
<tr>
<td>(Sklar, 1973)</td>
<td></td>
</tr>
<tr>
<td>Western Aphasia Battery (WAB)</td>
<td></td>
</tr>
<tr>
<td>(Kertesz, 1982)</td>
<td></td>
</tr>
</tbody>
</table>

*Tests included among the 10 most frequently used assessment tools for adult language (Muma, Pierce and Muma, 1983).*

types of information to be provided in manuals for educational and psychological tests. The criteria represent every major division of the APA standards. Based on APA specifications, decision rules were constructed and served as guidelines for determining whether a test manual provided sufficient information to meet each criterion. Table 2 lists the criteria and their respective decision rules. These methods are similar to those used by McCauley and Swisher (1984) to examine language and articulation tests for preschool children.

Criterion 1 (test development) requires that a test manual provide information about the development of the test, including descriptions of the guidelines used in developing items and the procedures used in selecting final test items. This criterion allows test users to evaluate the strengths and weaknesses of procedures used during the test’s development. Consequently, the decision rules for this criterion do not address the quality of information provided by the test developer, but merely report the presence or absence of information about the development of the test. Thus, tests may vary in the quality of the information they provide while meeting this criterion. It is the test consumer’s responsibility to evaluate the adequacy of the information presented.

Criterion 2 (tester qualifications) requires that the test manual contain a description of the qualifications a tester should possess in order to ad-
Table 2. Psychometric Criteria and Associated Decision Rules Used in Review

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Information Required for Decision that Criterion Was Met</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. &quot;A test manual should describe fully the development of the test: the rationale, specifications followed in writing items or selecting observations, procedures, results of item analysis or other research (APA, 1974, p. 11).</td>
<td>General description of specifications followed in item-writing process Evidence suggesting the existence of a pool of items Description of item analysis or other procedures used to select final test items from the pool</td>
</tr>
<tr>
<td>2. &quot;The test manual should identify any special qualifications required to administer the test and to interpret it properly&quot; (APA, 1974, p. 15).</td>
<td>Description of tester's required educational background Description of tester's required experience with the test instrument</td>
</tr>
<tr>
<td>3. &quot;The manual should state clearly what interpretations are intended for each subscore as well as for the total test&quot; (APA, 1974, p. 17).</td>
<td>Suggested interpretations of the overall test results Suggested interpretations for all scores or subscores yielded by the test</td>
</tr>
<tr>
<td>4. &quot;The directions should clearly state such critical matters as instructions on guessing, time limits, and procedures for marking answer sheets&quot; (APA, 1974, p. 18).</td>
<td>Directions to test taker that he or she need not be certain of a response to answer Directions to the tester about time limits for each task and an explanation of rationale behind omitting or including time limits on specific tasks Tester's procedures for marking answer sheets</td>
</tr>
<tr>
<td>5. &quot;A manual or research report should present the evidence of validity for each type of inference for which use of the test is recommended&quot; (APA, 1974, p. 31)</td>
<td>Description of test developer's construct of language abilities Explanation of the test's relationship to that construct</td>
</tr>
<tr>
<td>6. &quot;The test manual should indicate to what extent test scores are stable, that is, how nearly constant the scores are likely to be if a parallel form of a test is administered after time has elapsed&quot; (APA, 1974, p. 54).</td>
<td>Evidence of stability for a population described with regard to age and communication status (aphasia or no aphasia)</td>
</tr>
<tr>
<td>7. Norms presented in the test manual should refer to defined and clearly described populations. These populations should be the groups with whom the tester will ordinarily wish to compare the persons tested (APA, 1974, p. 20).</td>
<td>For tests considered norm-referenced, a populations needed to be defined with respect to three of the following four characteristics: age, socioeconomic status, education, gender</td>
</tr>
</tbody>
</table>

*Tests were considered norm-referenced if norms were used in instructions regarding the interpretation of the test or if, according to a statement of purpose, the test is to be used to identify aphasia.
minister the test. Both the tester's educational background and experience with the test are incorporated into decision rules for this criterion. When tester qualifications are reported, a prospective test user is informed of the test author's expectations about tester qualifications and can act to assure that those qualifications are met. Conversely, when tester qualifications are not reported, a test user may assume that anyone can read the manual, review the administration procedures, and then reliably and validly administer and interpret the test.

Criterion 3 (test interpretation) specifies that test manuals should suggest appropriate interpretations for test scores and subscores. The inclusion of such information would be expected to reduce the likelihood of misinterpretations and to increase the probability of consistent interpretation across test users. Comments on the interpretation of subscores, e.g., whether subtests can be administered and/or interpreted independently, may reduce the misuse involved when subtests are used outside of the context in which they were developed. Without explicit and detailed information about appropriate interpretations for all scores associated with a test, test users may be encouraged to interpret measures for which there is no standard interpretation.

Criterion 4 (directions) emphasizes the inclusion in the test manual of information on test instructions and scoring because such information can be useful to test users in the administration and appraisal of a test. During test administration, instructions on scoring and time limits clarify the tasks required of the tester and the patient. Additionally, this information may help test users evaluate the test by revealing the author's attempts to identify and control nonlanguage variables (e.g., dysarthria, apraxia, dementia, intelligence) that may affect performance of brain-injured patients (Wertz, 1978). Thus, when test authors state time limits—or explain why time limits are not imposed—and when they provide response instructions for the patient, they are providing indirect evidence of construct validity.

The decision rules for Criterion 5 (validity) focus on one type of validity, construct validity, because evidence of other kinds of validity are dependent upon the purposes of a test and may vary appropriately among tests. Although evidence suggesting construct validity is not equivalent to "evidence of validity for each type of inference for which use of the test is recommended," which the APA suggests is necessary (APA, 1974, p. 31), construct validity is essential for every test (Anastasi, 1976; APA, 1974, pp. 31, 46). In order to be consistent among tests, measures of construct validity serve as decision guidelines for this review. Evidence of construct validity provides the consumer with a measure of a test developer's reasoning about a theoretical trait (Anastasi, 1976). Further, Messick (1980) has proposed that evidence of construct validity is provided by evidence of other kinds of validity, e.g., criterion-related validity.
and content validity. Because aphasia is a language disorder, language was the construct selected for examination in this study.

Evidence of test–retest reliability, such as that required in Criterion 6 (score stability), provides information relative to a very important psychometric characteristic (APA, 1974, p. 54). Measures of score stability suggest the degree to which measurement error may intrude upon the measurement for which the test is used. The decision rules for this criterion require evidence of reliability for a population defined with respect to chronological age and communication status. These two characteristics of the test population are specified because they may affect score stability. Factors such as the time elapsed since the onset of aphasia and medical stability are also desirable information about an aphasic normative sample because of the rapid changes in language abilities observed during the period of spontaneous recovery (Culton, 1969; LaPointe, 1978). However, information about such factors was not required by this criterion.

Criterion 7 (normative sample) requires that the manual contain a clear description of any sample for which norms were reported. A detailed description of a normative sample is important for test users who want to make a norm-referenced interpretation and who, therefore, must decide whether the norms are appropriate for the patient and for the testing purpose (APA, 1974, pp. 20–21; Brookshire, 1983; Weiner and Hoock, 1973; McCauley and Swisher, 1984). To aid test users in these decisions, test authors should identify pertinent subject variables and, subsequently, measure those variables—a potentially difficult task.

If information pertinent to a criterion was not available in the manual or if it differed from the information specified in the associated decision rules, the test being reviewed was judged not to meet the criterion. Information available in other sources may not be immediately accessible to potential test users; therefore, such information was not included in the initial review but was noted for discussion.

RESULTS

Reliability of Judgments

Each author independently examined the tests in a random order and then rescored the first two tests. The results of this procedure were used to obtain an estimate of intraexaminer reliability. One examiner obtained perfect agreement between the two sets of judgments. The other examiner differed on 1 of 14 decisions from one scoring to the next, resulting in an intraexaminer agreement of 93%. Before interexaminer comparisons were made, the second author corrected the single discrepant judgment.

Comparisons of the authors' judgments for all nine tests revealed 82.5%
interexaminer agreement on the 63 ratings (9 tests × 7 criteria). Agreement was poorest for criterion 3 (55%) and best for criteria 1, 4, 5, 6, 7 (89%). For each judgment on which the authors disagreed, the test manual was consulted, and the disagreement was reconciled prior to further analysis.

Overall Results

Table 3 presents the criteria met by each of the reviewed test manuals for criteria 1 through 6. Because criterion 7 was relevant only to tests that were considered norm-referenced, results for this criterion are not included in Table 3 but are discussed later in the results section.

As shown in Table 3, test manuals were judged to perform quite differently on the six criteria considered in this review. No test met all of the criteria. The number of criteria met by the test manuals ranged from five for the PICA to one for the EFA and the WAB. The median and modal number of criteria met was three.

The number of tests meeting each criterion also varied greatly. The criterion pertinent to test score interpretation (criterion 3) was met by eight of the nine tests reviewed, whereas the criterion concerned with test validity (criterion 5) was met by seven. In contrast, the criteria related to the description of test development (criterion 1) and to testing directions (criterion 4) were met by only one test each. Another infrequently met criterion required evidence of test score stability (criterion 6); it was met by only two tests.

<table>
<thead>
<tr>
<th>Test</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALPS</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>BDAE</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>CADL</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>EFA</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>FCP</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>MTDDA</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>PICA</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>SAS</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>WAB</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Totals for each criterion</td>
<td>1</td>
<td>5</td>
<td>8</td>
<td>1</td>
<td>7</td>
<td>2</td>
<td>24</td>
</tr>
</tbody>
</table>

x indicates tests passing the criterion.
Results for Each Criterion

In this section, results for each criterion are described in detail. The reasons that each test met or did not meet the criterion are reported, including whether information supplied in the test manual was incomplete or differed from that specified by the decision rules.

Criterion 1 (Test Development). Only the manual for the CADL (pp. 2–3, 12, 16) contained all of the information required by criterion 1 for the description of test development. In contrast, the manuals for the EFA (pp. 29–31), the FCP (pp. 1–2, 4), and the WAB (p. 1) provided little or no information about the following aspects of test development: the existence of a pool of possible items, the procedures used to develop test items, or the methods used to choose final test items from the pool of potential items.

The remaining five test manuals provided some but not all of the information required by criterion 1. The manuals for both the ALPS (pp. 9–12) and the MTDDA (pp. 20–24) provided information about item writing and the methods used to choose final test items. For both tests, however, it was unclear whether the test authors began with a pool of items from which final items were chosen or whether items were added during test development that had not undergone the same scrutiny as earlier items. The SAS manual (p. 11) contained a description of the item-writing process and implied the existence of a pool of items at the beginning of the test’s development. However, it did not contain a description of the method by which items were chosen for the final version of the test. The manuals for the BDAE (pp. 2–3) and the PICA (p. 6) provided adequate descriptions of the item-writing process; however, neither a description of the methods used to choose final test items nor information about the existence of an item pool were provided.

Criterion 2 (Tester Qualifications). Five of the nine test manuals (ALPS, CADL, FCP, MTDDA, and PICA) provided adequate descriptions of tester qualifications that addressed both the tester’s educational background and the tester’s familiarity with the test. The BDAE, EFA, SAS and WAB manuals did not meet criterion 2.

The five tests that met this criterion differed in their descriptions of the tester’s expected educational background. Three of these five test manuals (ALPS, p. 32; MTDDA, p.4; PICA, p. 1) referred to the “clinician,” which was interpreted to mean “speech–language clinician.” The manual for the CADL (p. 44) stated that no educational criteria need be met as long as one is trained in the testing procedure. This suggests that there are no educational requirements for testers. The FCP manual was the fifth manual that met this criterion. The author included the statement that the
FCP is not recommended for "the clinically inexperienced working in settings where few aphasics are seen" (pp. 2-3).

For the five test manuals that met criterion 2, information about the tester's familiarity with the test was stated more specifically than the tester's educational background. The types of information supplied by these manuals included the number of times the test should be presented in a practice setting prior to use with a patient (ALPS, p. 32), the necessary degree of familiarity with test materials (FCP, p. 11; ALPS, p. 32), and descriptions of specific training procedures for testers (CADL, p. 49; PICA, p. 1).

The four tests that did not meet criterion 2 were the BDAE, EFA, SAS, and WAB. Manuals for the BDAE, EFA, and SAS provided information about desired educational characteristics of testers, but did not state how familiar the tester need be with the test instrument. The WAB manual did not include information about either type of tester qualification.

Criterion 3 (Test Interpretation). All of the test manuals, except for the EFA manual, provided the information required for this criterion. The manual for the EFA did not meet this criterion because there was no clear statement of how overall test results were to be interpreted. It was stated, without providing additional instructions, that the summary sheet "permits the examiner to obtain an overall and profile view of the subject's abilities and disabilities" (p. 32). Unlike the other tests that met this criterion, the CADL has no subscores; therefore, it met this criterion on the basis of its instructions for interpretation of the overall test score.

Criterion 4 (Directions). The PICA was the only test whose manual provided evidence of information required by all three decision rules regarding instructions on guessing, time limits, and procedures for marking answer sheets. All nine test manuals included directions for scoring their respective response sheets; however, only the PICA (vol. 2, pp. 13-19), CADL (p. 48), and SAS (p. 7) manuals included directions and rationales for time limits applied to specific tasks. Discussion of response latency was included in the FCP (pp. 15-16) and the EFA (p. 32) manuals, but no rationale for its use was provided. The BDAE manual was not specific in terms of response latency, but instructions for scoring some task relative to promptness of response are included in the answer booklet. The PICA was the only aphasia test manual that suggested that the clinician tell the patient that he or she need not be absolutely certain of a response to answer (vol. 2, p. 5).

Criterion 5 (Validity). Seven of the nine test manuals provided evidence of construct validity as described in the decision rules listed in Table 2. The BDAE (pp. 2, 5-11), CADL (pp. 1-2), EFA (pp. 1-2), FCP (pp. 1-
all described the test authors’ constructs of language and stated how their respective tests related to those constructs. Although the authors of the ALPS mentioned five modalities, they did not present further discussion of language (p. 3). The author of the WAB defined language in the test manual (p. 1) but did not explain the WAB’s relationship to that construct. That information can be found in an article by Shewan and Kertesz (1980). Additional information relative to the validity of the WAB is available in the text *Aphasia and Associated Disorders: Taxonomy, Localization, and Recovery* (Kertesz, 1979).

**Criterion 6 (Score Stability).** Only the test manual for the ALPS (pp. 15, 17) and the PICA (pp. 44–45) provided evidence of test score stability required by this criterion. Two studies of test–retest reliability were reported by the authors of the ALPS, one for 22 aphasia patients and one for 61 normal subjects. The reliability coefficient for the four subtests of the ALPS ranged from 0.83 to 0.94 for the aphasic patients who were retested from 3 to 5 wk after an initial testing. The PICA manual reported a reliability correlation coefficient of 0.99 for 40 “patients.”

Score stability information was available for the WAB in two sources (Kertesz, 1979; Shewan and Kertesz, 1980), but it was not reported in the manual. Some test–retest reliability data were included in the FCP manual (p. 24), but the ages of subjects were not reported. Therefore, the FCP manual did not meet this criterion. Furthermore, the data that were reported were obtained from acute aphasic patients whose performances could be expected to change during the 2-wk interval between administrations as a consequence of spontaneous recovery alone (Culton, 1969; LaPointe, 1978). The MTDDA (pp. 22, 101) manual suggested that stability data had been obtained, but the evidence presented was vague and seemed to apply to an earlier version of the test. The authors of the BDAE acknowledged that stability data had not been obtained for their test (p. 12). Reliability data were also not reported for current versions of the CADL, EFA, or SAS.

**Criterion 7 (Normative Sample).** The ALPS (p. 107), EFA (p. 31), and SAS (p. 1) were not considered norm-referenced according to the guidelines stated in Table 2. That is, they neither indicated that the identification of aphasia was an intended purpose of the test, nor did they make use of norms in the interpretation of scores. Consequently, the ALPS, EFA, and SAS tests were not scored with respect to this criterion.

Manuals for the CADL (pp. 25, 26, 35) and PICA (vol. 2, pp. 70–116) provided information about the age, education, gender, and socioeconomic status (occupation) for the normative sample and, therefore, met this criterion. Criterion 7 was considered pertinent to, but was not met by, the manuals of the BDAE, FCP, MTDDA, and WAB. According to
the statement of purpose, the BDAE is to be used to diagnose "the presence and type of aphasic syndrome" (p. 1). Similarly, test scores were used during test interpretation for the FCP (pp. 18, 22). The MTDDA also used norms and provided minimal evidence about the normative sample (p. 21). No information regarding the normative sample associated with the WAB was reported in the test manual.

DISCUSSION

Manuals for nine aphasia tests were examined using seven psychometric criteria that were selected from APA standards for psychological tests (1974). When criteria 1 through 6 were applied to all tests, the number of criteria met by individual tests ranged from one to five, with a median and mode of three criteria. The PICA manual met the greatest number of criteria, five; whereas the manuals for the EFA and WAB met only one criterion each. Although the PICA met more of the psychometric characteristics deemed "essential" by the APA (1974), we are not suggesting that the PICA is a better assessment tool. Psychometric characteristics are but one group of characteristics to consider in test selection and interpretation. Other practical issues include, but are not limited to, test length, variability and difficulty of test items, linguistic sophistication, and measurement of functional communication.

Of the characteristics reviewed in this study, the criteria on test interpretation (criterion 3) and on construct validity (criterion 5) were met most frequently. These criteria were met by seven and eight tests, respectively. The least frequently met criteria pertained to the description of test development (criterion 1) and the directions for the test's administration (criterion 4). These criteria were met once each. Criterion 7 (normative sample) was only applied to six of the nine tests, and the CADL and the PICA were the only tests of those six that met this criterion.

Foremost among the characteristics investigated in this review were validity and reliability. When empirical evidence of validity is not reported in a test manual, the usefulness of the test in assessing the performance of a given patient cannot be estimated. Therefore, although construct validity alone was examined in this review, evidence of criterion-related and content-related validity are also crucial to test assessment. Only such evidence in its entirety supports specific uses of a test.

The importance of reliability stems both from its role as a prerequisite to validity (measurement must be consistent if it is to be valid) and from its assurance that a representative sample has been obtained. If a measure is reliable for individuals who are similar to a given patient, then an examiner can assume that repeated testing of the patient would probably yield similar results each time. Without such evidence, an examiner could not be sure that repeated testings would result in comparable test scores (Anastasi, 1976). Thus, demonstrating test-retest reliability is crucial for
any test author (APA, 1974), and such data should be obtained for individuals who present characteristics similar to those on whom the test is likely to be administered (Anastasi, 1976, p. 127). However, obtaining consistent measures for acute aphasic patients presents problems because of spontaneous recovery (Culton, 1969; LaPointe, 1978) and because of practical variables such as medical stability, medication, and patient fatigue.

Although only criteria on seven psychometric characteristics were addressed in this review, numerous other characteristics are considered important for standardized tests (APA, 1974). All such psychometric characteristics contribute to the test’s validity. Thus, all of a test’s psychometric characteristics affect the degree to which it can be used to measure what it is intended to measure for a particular patient.

When assessing the psychometric adequacy of aphasia tests, test users confront difficult decisions, even when all of the information discussed in this review is provided in the test manual. No standardized test can reasonably include all of the specific psychometric characteristics needed for all patients and for all testing purposes (Messick, 1980). A clinician may wish, for example, to assess the language abilities of a noninstitutionalized male who is 85 yr of age and taking medication for hypertension, yet there exists no aphasia test that provides information on the stability of the test for that type of individual. Therefore, the subjects on whom standardization measures (e.g., norms or evidence of reliability or validity) were obtained affect the interpretation of that test. A consideration of the functions of test norms may help explain why the subjects on whom the test is developed are so important.

Norms generally are used to compare an individual’s performance on a test with the performance of other similar individuals (Anastasi, 1976). Aphasia tests often are administered to compare a given patient with groups of individuals who represent different degrees of aphasic impairment or different types of aphasia. When normative samples are obtained, subject variability should be controlled, or it will not be apparent whether differences in performance may be attributed to aphasia or to other variables.

Brookshire (1983) has expressed related concerns about the description of aphasic subjects in experimental studies. He noted that more complete descriptions of subject characteristics allow readers to judge whether the observed results of a study are more likely to be caused by the manipulated, independent variables or by uncontrolled subject variables. In aphasia testing, uniform reporting of subject characteristics would serve a similar function. The presence of such information would indicate to users the likelihood that characteristics other than aphasia would be responsible for observed differences in test performances between the test taker and the test norms. Therefore, the absence of data about characteristics of the normative sample significantly lessens the utility of a test.
Brookshire (1983) proposed that experimental reports involving aphasic subjects include the following information: age, education, source of subjects, gender, lesion location, handedness, etiology, time postonset, and severity of aphasia. Test authors and users would be wise to consider including information about these eight characteristics for subjects used in test standardization to insure appropriate use of tests. In addition, they should obtain standardization information on homogeneous subgroups categorized according to those characteristics.

The evaluation of the adequacy of a test's validity, reliability, and other psychometric characteristics should be implemented in the specific context presented by the patient and the assessment purpose. Thus, crucial questions to be asked follow: What is the purpose of this assessment? What decisions and interpretations will be made based on the results of the test administration? What are the special characteristics of this patient that may confound assessment efforts? Only when these questions are answered for each patient and each assessment purpose should a clinician undertake assessment.

This research was conducted while R. McCauley was supported by National Institute of Neurological and Communicative Disorders and Stroke, National Research Service Award, 5 F32 NS06390 CMS. The authors would like to thank Daniel Boone, Richard Curlee, Theodore Glattke, and Linda Swisher for their suggestions on an earlier version of this article.

REFERENCES


