DOCUMENT RESUME

ED 480 038 CG 032 611

AUTHOR	McDivitt, Patricia Jo; Gibson, Donna
TITLE	Guidelines for Selecting Appropriate Tests.
PUB DATE	2003-08-00
NOTE	22p.; In: Measuring Up: Assessment Issues for Teachers, Counselors, and Administrators; see CG 032 608.
PUB TYPE	Information Analyses (070)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Educational Assessment; *Evaluation Methods; *Guidelines; *Student Evaluation; Teacher Competencies; Teacher Education; *Test Selection

ABSTRACT

In 1990 the American Federation of Teachers (AFT), the National Council on Measurement in Education (NCME), and the National Education Association (NEA) published "Standards for Teacher Competence in Educational Assessment of Students." Standard 1 of this document states, "Teachers should be skilled in choosing assessment methods appropriate for instructional decisions." Teachers and all educators involved in the selection and use of tests follow several guidelines when seeking to gain this competence. These guidelines include understanding the purpose of the assessment and determining the quality of the assessment. This chapter reviews these guidelines and provides educators with important information to help them select appropriate tests. (Contains 24 references.) (Author)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 480 038

Guidelines for Selecting Appropriate Tests

By
Patricia Jo McDivitt
Donna Gibson

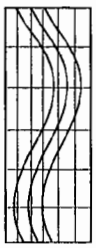
U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- This document has been reproduced as received from the person or organization originating it.
- Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

CG032611



Chapter 3
Guidelines for Selecting Appropriate Tests

Patricia Jo McDivitt & Donna Gibson

In 1990 the American Federation of Teachers (AFT), the National Council on Measurement in Education (NCME), and the National Education Association (NEA) published *Standards for Teacher Competence in Educational Assessment of Students*. Standard 1 of this document states, "Teachers should be skilled in choosing assessment methods appropriate for instructional decisions." (p. 3) Teachers and all educators involved in the selection and use of tests follow several guidelines when seeking to gain this competence. These guidelines include understanding the purpose of the assessment and determining the quality of the assessment. This chapter reviews these guidelines and provides educators with important information to help them select appropriate tests.

Understanding the Purpose of a Test

The first step in attaining competency in selecting appropriate tests involves understanding the purpose or purposes for which an assessment is given. According to Mehrens (2001), in its broadest sense, the purpose of any assessment is to gather data to facilitate decision making. However, many kinds of decisions and many different types of information may be gained from the use of tests and may serve to facilitate decision making. For example, the decision made may involve helping an individual select courses for high school or make wise, realistic career decisions; other decisions might be made to help an individual improve upon his or her strengths and weaknesses in a given subject area; and still others might be made to help an individual build toward mastery of a particular set of content curriculum standards or learning targets. In today's high-stakes arena, still other tests may be used to make important decisions such as whether a particular student should be promoted to the next grade in school or should receive a high

school diploma.

Most tests used in modern educational settings can be categorized into two major types: norm-referenced tests and criterion-referenced tests. These two types of tests differ in purpose, content, and the information gained from their use. The main purpose of a *norm-referenced test* is to compare students' performance and to determine relative strengths and weaknesses of students based upon the generalized skills being measured by the test.

In contrast, *criterion-referenced tests* determine "what test takers can do and what they know, not how they compare to others" (Anastasi, 1988, p. 102). Criterion-referenced tests report how well students are doing relative to a predetermined performance level on a specified set of educational goals or outcomes included in the school, district, or state curriculum. Educators may choose to use a criterion-referenced test when they want to determine how well students have learned the knowledge and skills they are expected to have mastered (Bond, 1996).

When deciding whether to use a norm-referenced or a criterion-referenced test, it is important to know about the content differences between the two. The content of a norm-referenced test is selected according to how well it ranks students from high achievers to low. The content of a criterion-referenced test is determined by how well it matches the learning outcomes deemed most important. Although no test can measure everything of importance, the content of a criterion-referenced test is selected based on its significance in the curriculum, whereas that of a norm-referenced test is chosen by how well it discriminates among students (Bond, 1996). Because the purpose of many norm-referenced tests currently used in the classroom is to measure the academic foundation skills that students need, the test questions are usually designed to measure a generalized set of objectives that are common across the country for a given content area.

> When standardized tests are norm-referenced, it means that national samples of students have been used as the norming group for interpreting relative standing. Because these tests are designed to be used in different schools throughout the country, they tend to provide broad coverage of each content area to maximize potential usefulness in as many schools as possible. Thus, close inspection of the objectives and types of test questions is needed to determine how well the test matches the emphasis in the local curriculum. (McMillan, 1997, pp. 79–80)

Evaluating Test Quality

The second step in selecting an appropriate test is to evaluate its quality. Evaluating the quality of a test involves a careful analysis of the characteristics of the population to be tested; the knowledge, skills, abilities, or attitudes to be assessed; and the eventual use and interpretation of the test scores (ACA & AAC, 1987). The following list outlines major quality criteria that teachers, counselors, and other test users should consider when selecting a test. These criteria are relevant for many kinds of tests not strictly those used in educational settings or classrooms. This information is based upon Klein and Hamilton (1999, Table 1), the *Code of Fair Testing Practices in Education* (JCTP, 2002), and *Responsibilities of Users of Standardized Tests* (ACA & AAC, 1987).

Purpose. Compare the purpose and recommended use of the assessment against your assessment goals.

Validity. Check for evidence of validity, that is, the degree to which an assessment measures what it is intended to measure.

Reliability. Check the consistency and dependability of the assessment results. Select only tests that have documented evidence of reliability, that is, consistency.

Alignment with curriculum. For tests intended to measure students' mastery of learning targets, check for instructional validity, or the degree to which the test questions measure what is actually taught in the classroom.

Equity and fairness. Check to be sure that the test meets appropriate standards for bias, fairness, and cultural sensitivity, and is fair and equitable for all test takers in your setting.

Technical standards. If the assessment is norm-referenced, check for norming procedures that are relevant to the local population and intended use of the data; also check for the types and quality of norms.

Costs and feasibility. Check for practical constraints due to cost, conditions, and time required for administration.

Consequences. Check what inferences and actions might result from the use of the test scores.

Timeliness of score reports. Check on the length of time between the test administration and the receipt of score reports.

Motivation. Check for the degree to which examinees will be motivated to do their best.

Quality of the administrative, interpretative, and technical manuals. Check to see that supportive materials are high in quality, user friendly, and readily available.

Each of these issues will be described in more detail in the remainder of this chapter. The selection of a test should be guided by established criteria for technical quality recommended by measurement professionals, including validity and reliability. Therefore, we begin with a discussion of technical qualities, including validity and reliability.

Validity

Assessments need to be fair, reliable, defensible, and free of bias. They also need to be valid. In fact, validity is at the core of the test development process for any assessment. One common definition of validity is contained in Cronbach (1971): Test validation is a process in which evidence is collected by the developer of a test to support the types of inferences that may appropriately be drawn from test scores. A more recent definition of validity is cited in the 1999 version of the *Standards for Educational and Psychological Testing:*

> Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself. When test scores are used or interpreted in more than one way, each intended interpretation must be validated. (AERA, APA, & NCME, 1999, p. 9)

When gathering and examining evidence of validity, the first question to ask is, Validity for what purpose? For example, career interest inventories have been in use for a number of years, and many of these instruments have well-documented validity. The validity of such interest inventories has commonly been determined by comparing individuals' interests with their occupational choices and then determining the rate of correct predictions over a specified period of time (Seligman, 1980). When predicting the occupation that a person is likely to enter in the future, an interest inventory may be valid because the person's answers to the questions will probably relate to career interests. When it comes to predicting whether this person will be successful in the given occupation, however, a career interest inventory may lack validity. Persons who enter an occupation for which they receive a low score on a given career interest inventory may well not stay in the occupation, whereas people who score high are much more likely to stay in the occupation. The low scorers who stay in that field are just as likely to be successful as the high scorers, however. Therefore, a score on a given interest inventory may have some validity for predicting whether people will enter an occupation, and how long they will stay in it, but may have little validity when predicting success in the occupation (Hood & Johnson, 1991).

As a result, determining whether or not a test is valid involves a process of gathering evidence to support a specific interpretation of the test scores. Many different methods for gathering evidence exist, and the evidence gathered establishes what kinds of inferences are appropriate to make (Osterlind, 1989). In looking at validity, educators must keep in mind what specific inferences will be drawn from the scores, then look for and gather evidence to support such inferences. Mehrens (2001) identifies two general types of inferences: (1) inferences about performance other than that measured, and (2) inferences about a characteristic (construct) of the person measured.

When gathering evidence, it is important to note that there are several types of validity. These are discussed in the sections that follow.

Face Validity

Face validity asks the question, Based upon a surface examination, does the test look like it measures what it is intended to measure, with test questions that appear to provide an adequate measure of what the test as a whole is intended to measure? Face validity is simply a matter of whether or not the test questions on the surface seem to be relevant to the person taking the test (Hood & Johnson, 1991). Some would

argue that face validity is really not valid at all, especially if the process of examining an assessment is haphazard or not very systematic. For example, when examining a mathematics test consisting of word problems, teachers might ask themselves whether the test items do in fact appear to measure the defined mathematics objectives, or instead measure reading comprehension ability. A quick look at the test may lead them to conclude that the test does not have face validity because it appears to measure reading comprehension more than the mathematics skills or objectives it purports to measure.

Content Validity

Although it is important that an assessment does have some face validity, it is more important that the evidence of validity be documented, or have content validity.

Content validity indicates whether the material in the test is related to what is being measured and reflects the level of learning or development of that skill (Seligman, 1980). Content validity asks the fundamental question, How well does the assessment measure what it is intended to measure? For example, if a high school end-of-course biology test purports to measure the curriculum standards and core skills outlined for the course, then each test item or question must show a close correspondence to those curriculum standards and core skills. This close correspondence must be documented through a content validation study, which seeks to establish a consensus of informed opinions about the degree of congruence between particular test items and specific descriptions of the content domain to be assessed by those items. A content validation study requires convening a panel of expert judges who rate the item-to-content congruence according to established criteria (Osterlind, 1989).

In the development of current criterion-referenced statewide assessment programs, the content validation study often involves educators, including curriculum experts, subject-area teachers, and others. These educators, who are experts in the subject area, are asked to use their professional judgment to determine whether or not the test questions on a given criterion-referenced test do in fact measure the designated curriculum content standards or learning targets. This process depends on the development of clear learning targets. Based upon the learning targets for a given program or subject area, a test blueprint for the assessment is developed. The blueprint outlines the number of items a given test will include, mapped directly to the learning targets. The blueprint also provides information concerning the relative emphasis

assigned to particular learning targets.

Instructional Validity

For many criterion-referenced tests used in the schools today, one aspect of content validity is the extent to which the test has instructional validity.

Instructional validity relates to the match between what is taught in the classroom and what is actually assessed. When examining instructional validity, the major questions to ask are, How closely do the test questions correspond to what has actually been taught in the classroom? Have students had the opportunity to learn what is being assessed? Instructional validity is also determined by teachers' professional judgments (McMillan, 1997).

Criterion-Related Validity

Validity also refers to the extent to which the test is related to defined criterion measures. Establishing criterion-related validity involves accumulating various types of evidence: "Evidence of the relation of test scores to a relevant criterion may be expressed in various ways, but the fundamental question is always: How accurately do test scores predict criterion performance? The degree of accuracy deemed necessary depends on the purpose for which the test is used" (AERA et al., 1999, p. 14).

Test developers and researchers seek to establish criterion-related evidence that a test is measuring the same trait, knowledge, or attitude by calculating a correlation coefficient, which measures the relationship between the test and the criterion. Unlike in content-validation studies, teachers and subject-area experts typically do not conduct formal studies to obtain correlation coefficients that will provide evidence of criterion-related validity. However, understanding the principles of establishing criterion-related validity is important. Where there are two or more measures of the same thing, and these measures provide similar results, criterion-related evidence can be established informally (McMillan, 1997). For example, consider the development of a test of computer word processing skills that measures speed and accuracy of key entry. The test might be given to a student who is taking a word processing course. The classroom teacher might then be asked to observe the student's word processing skills and rate the student using a rating sheet. The teacher's rating sheet would be compared with the student's test results, to determine how closely related the two are. If the teacher's observational ratings coincide with the student's score on the test, then

criterion-related validity has been established. This type of validity is also called *concurrent validity.* Measures of concurrent validity are usually obtained when the test is going to be used in the future to estimate some type of behavior—such as the ability to do the work of a key-entry word processor.

Another type of criterion-related validity is called *predictive validity.* For example, if a classroom teacher is interested in the extent to which students' test preparation, as indicated by scores on a final examination in mathematics, predicts how well those students will do next year, he or she might examine the grades of students who took the class previously, then determine informally if students who scored high on the final examination are getting high grades, and students who scored low on the final examination are getting low grades, in the current year's math class. If a correlation is found, then an inference predicting how the students in the class will perform, based on the final exam, might be valid (McMillan, 1997).

Construct Validity

Construct validity is determined by gathering evidence that there is a relationship between the content of a test and the construct it is intended to measure. Construct validity demonstrates two points: (1) that the construct measured by the test is required for success on the criterion of interest, and (2) that the specific test under consideration is a good measure of the theoretical construct or trait (Bennett, Seashore, & Wesman, 1991).

> Test content refers to the themes, wording, and format of the items, tasks, or questions on a test, as well as the guidelines for procedures regarding administration and scoring. Test developers often work from a specification of the content domain. The content specification carefully describes the content in detail, often with a classification of areas of content and types of items. Evidence based on test content can include logical or empirical analyses of the adequacy with which the test content represents the content domain and of the relevance of the content domain to the proposed interpretation of test scores. (AERA et al., 1999, p. 13)

Construct validity evidence relies on both logical and statistical means to justify the use of a test. Evidence of construct validity is usually gathered by collecting criterion-related validity evidence, content validity evidence, and information about the test development process.

Construct validity also involves gathering evidence or information about the test's overlap with other tests. Convergent validity and divergent validity also provide evidence of construct validity. *Convergent validity* means that an assessment shows a substantial correlation with other tests and assessments that measure similar characteristics. For example, students ought to score similarly on measures of mathematical aptitude and on the mathematics section of an achievement test. *Divergent validity,* on the other hand, is shown when an assessment does not correlate highly with a test or a variable that measures different constructs. For example, a student's score on a test of perceptual speed and accuracy in all likelihood would not show a strong correlation with a test of academic achievement.

According to Hood and Johnson (1991, p. 37), construct validity is a complex concept that encompasses several questions:

- Do the test results make psychological sense?
- Are the test results related to things that they ought to be related to and unrelated to things that they ought not to be related to?
- Do the results on the test change according to what we know about developmental changes?
- Do older students do better on the test than younger students; for example, on an arithmetic test, do sixth graders score higher than third graders do?
- Does the test pick up the kinds of changes known to occur as people develop?

Validity Checklist

The validity checklist in Figure 1 is designed to help test users determine whether or not a given test is valid. Because test selection should be guided by established criteria for technical quality recommended by measurement professionals, including validity, the items on the checklist address what types of validity information are available and whether or not the validity information provided is relevant to the purposes of the test. For example, content validity and instructional validity are important if you are using a criterion-referenced test to determine whether students have mastered specific learning targets. On the other hand, criterion-related validity is important if you are using the test for employee selection purposes.

Figure 1. Validity checklist

Type of Validity	Ask Yourself	Yes/No
Face validity	Does the assessment appear to measure what it is intended to measure?	Y N
Content validity	Is there documentation that the assessment measures what it is intended to measure?	Y N
Instructional validity	Do the assessment questions correspond to what has actually been taught in the classroom?	Y N
Criterion-related validity	Do the test scores predict future performance on a specific criterion?	Y N
Predictive validity	Is there evidence showing that the test accurately predicts future performance?	Y N
Concurrent validity	Is there evidence showing that the test measures performance on the relevant behaviors?	Y N
Construct validity	Does the assessment represent the theoretical entity it is intended to represent?	Y N
Convergent validity	Is there evidence showing that the test results are similar to results on other measures that should be related?	Y N
Divergent validity	Is there evidence showing that the test results are unlike those obtained on other, unrelated measures?	Y N

Reliability

In order to have good validity, a test must be reliable (Lyman, 1998). In general terms *reliability* refers to how consistently a test measures what it is purported to measure (Hood & Johnson, 1997). In fact, a test can be highly reliable (i.e., give consistent results) and not measure what it is purported to measure (i.e., not be valid). Therefore, a good understanding of reliability is required for appropriate testing.

In essence, a reliable test can be depended on to measure the same trait or variable each time it is used.

When a test is reliable, the results can be generalized in several different ways. First, the test user can assume that items that are similar but not identical to those on the original test would produce similar results (called alternate-form reliability). For example, a teacher may test students on their recognition of single-digit numerals by testing recognition of five different single-digit numerals. Because this measure is highly reliable, the teacher may assume that students would receive the same score if they were tested on other single-digit numerals. Hence, the teacher can generalize from one sample of items from the single-digit numeral domain to any other samples from the single-digit numeral domain (Salvia & Ysseldyke, 2001).

Second, results can be generalized from one time to another; that is, the same testing behavior or results will occur again if the students are tested with the same test at a different time (called test-retest reliability). For example, if the teacher gave the single-digit numeral test to students in the morning, he or she should see the same results upon administering the test in the afternoon (provided that no teaching of numerals has occurred in the interim).

Third, there should be consistency in results among testers (called inter-rater or interscorer reliability). If one teacher scores students on their recognition of single-digit numerals, then a second teacher scores the same students on the same measure, the two teachers' results should be similar. If they are, the assumption is made that scorers are consistent and results are reproducible among the scorers. These three types of reliability are discussed in more detail in the following sections.

Alternate-Form Reliability and Internal Consistency

Alternate-form reliability is determined by comparing the consistency of one individual's testing behavior on two equivalent forms of the same test (Hood & Johnson, 1997). Both forms of the test must be constructed to measure the same trait or construct and look similar in terms of format, number of items, and directions (Ponterotto, 1996). If necessary, the individual being assessed can be given both forms of the test without concern that results will reflect being exposed to the same test items. Often school systems will use two forms (e.g., Form A and Form B) of a standardized achievement test to accommodate students being served in special education programs. Alternate-form reliability is particularly important when the test users will need to test individuals or groups several times on the same content or trait, as

might occur in research, in examining the effectiveness of teaching methods, or in examining student achievement.

Internal consistency is calculated on only one form of a test and is used to estimate the generalizability of results to different test items (Salvia & Ysseldyke, 2001). Specifically, the *reliability coefficients* obtained through this process indicate the consistency with which the items sample the trait being measured (Hood & Johnson, 1997). This type of reliability is important for tests that are not timed and are not completed under time pressure (Lyman, 1998).

Stability or Test-Retest Reliability

Stability and *test-retest reliability* are often used synonymously because test-retest reliability is an index of stability (Salvia & Ysseldyke, 2001). This method of evaluating reliability involves administering the same test instrument to one group or sample at two points in time (Ponterotto, 1996). Calculating test-retest reliability allows the user to know if the test produces the same results over time.

There are several considerations when evaluating the test-retest reliability of a particular measure. The first is to determine the interval between the two administrations of the test. Reliability coefficients can be expected to decrease as the length of the interval increases. If the interval is too long, maturation of the test takers and events they have experienced (learning) may influence the results. Conversely, test-retest coefficients can be inflated if the interval is too short. When the interval is brief, memory and practice may influence the test takers' results.

Inter-rater or Interscorer Reliability

When establishing *inter-rater* or *interscorer reliability,* two or more scorers score a set of tests independently and their scores are correlated to establish the reliability coefficient (Salvia & Ysseldyke, 2001), or degree to which the two scorers agree. This type of reliability is important when there is an element of subjectivity in scoring tests or rating behaviors (Lyman, 1998). A test that can be scored objectively has perfect interscorer reliability; however, many tests include items that are scored by subjective criteria. For example, individually administered achievement and aptitude tests require the test administrator to evaluate responses for score assignment. Additionally, behavior often must be rated on a subjective basis. With subjective evaluations, there is more variation in how items are rated among the raters (e.g., a student self-report of specific behaviors versus ratings of those behaviors by a parent, a teacher, and an administrator). This is

the source of error or error variance in reliability coefficient calculations. Steps should be taken to minimize the error variance for these tests to increase reliability.

When evaluating the reliability of tests, it is important to understand the meaning of the reliability coefficients that are reported. Both validity and reliability coefficients are reported as a correlation coefficient with a range from 0.00 to ±1.00. Reliability coefficients of +1.00 or 1.00 indicate a perfect relationship. A reliability coefficient of 0.00 indicates no relationship or no reliability. Additionally, reliability coefficients provide the cap for validity coefficients, meaning that validity coefficients for a particular test cannot be higher than the reliability coefficients for that test.

What is an acceptable level of reliability? Ponterotto (1996, p. 80) states that there "is no absolute answer to this question." When selecting a test, you need to determine the purpose of the test and implications of the test results. If the results will have significant, life-altering consequences (e.g., decisions about educational placement, admissions, or medical interventions), then high levels of reliability are necessary (Walsh & Betz, 1995). On the other hand, midlevel coefficients may be appropriate for research purposes with large samples. Lyman (1998) has proposed that the following factors affect reliability:

How long is the test? A test with many items that assess a construct or trait is more reliable than one with only a few items, unless the test is so long as to induce fatigue in the test taker.

Who made up the group of people studied in the test construction process? Review the test publisher's description of the groups that were tested and for whom reliability coefficients were calculated. In general, the more group members vary in ability or behavior, the higher the reliability coefficients are likely to be.

How much time elapsed between test and retest sessions? The more time that elapses between sessions, the more likely reliability coefficients are to be low. A two-week time period is considered preferable (Salvia & Ysseldyke, 2001) because the period is long enough that test takers are unlikely to remember specific items from the previous administration but not long enough for significant maturation to have occurred.

What types of reliability are reported? A test publisher may provide coefficients for all the different types of reliability or only certain ones, and the coefficients for the various types of reliability will differ. Remember, consider the purpose of the testing to evaluate which types of reliability are most essential for your purposes.

The validity and reliability of a test are the essential psychometric properties you should review when selecting the appropriate assessment instrument for your needs. Practical considerations related to the usability of a test instrument also factor in to the decision, however.

Usability of the Test Instrument

What happens when the most reliable and valid test instrument is too expensive for an organization to use? What should the test users do when a valid and reliable test is affordable but the test publisher requires six months to score it? What should a principal do if the school district is using a group-administered test that was developed with Caucasian children only, but his or her school is 80 percent African American?

These are a few of the dilemmas that surface when evaluating the usability of a test. Many test publishers facilitate the process of evaluating test usability by including information about the test construction process in the test manual. For norm-referenced tests, characteristics about the norming sample are usually provided. Here are some questions to consider when evaluating the usability of a test for a specific population of test takers. In general the answers to these questions will be found in the test manual or information provided by the publisher.

What is the age group of the test takers? Test publishers provide information about the age range of the group on whom the test was normed. Look for a match between the age range of the normative sample and of your test takers. If a test taker's age falls outside of the normed age range for the test, then the results will not be reliable or valid for that individual.

Is the test designed for both genders? In general, males and females are represented in the norming group for most tests. Certain tests, however, may be designed for males or females exclusively. If a male is given a test created for and normed on females only, then his results will not be valid or reliable.

Where do the test takers reside? In what part of the United States (or what country outside the United States) do your test takers live? In recruiting norm groups, test developers attempt to include a cross section of individuals from various regions of the country. Before choosing a test, you should ensure that your region is represented in the norming sample.

What racial and ethnic groups are represented in the norming sample of the test? For a variety of reasons, different races and ethnic groups perform differently on tests of achievement and intelligence (Salvia & Ysseldyke, 2001). Under-representation or over-representation of specific groups can reflect bias in the construction of the test. Therefore, you need to determine that the norming sample is representative of your population of test takers, in order for the results to be comparable.

There are several additional practical criteria to consider. First, expense is a concern for many test users. If the most valid and reliable test is desired but is too expensive to be practical, then compromise may be the answer. The test user may have to compromise on the standards for choosing the test and seek a more cost-effective one with acceptable levels of reliability and validity.

Second, ease of use is an important criterion to consider when many different people will be administering or scoring the test. Particularly if the test will be administered to large groups, another consideration is the clarity of the administration instructions and directions for the test taker. Finally, scoring procedures need to be clear, and you will need to determine whether the test can be scored on-site or requires off-site scoring, and how much time is required for the scoring process. If you need immediate results, a test that does not require a lengthy off-site scoring process is the best choice.

Third, the amount of time allotted for administration and completion of the test is an important factor to consider, especially for large groups of test takers. For example, most school districts schedule a set number of days for group test administrations. In addition, counselors and psychologists may need to consider when and where students can complete individually administered tests or behavior checklists in order to achieve the maximum level of effort and performance.

Overall, choosing an appropriate assessment instrument can be a complex process. Validity and reliability criteria are essential in

determining that a test has been constructed properly. In addition, there are many practical criteria to consider, such as the norming group and logistical issues. Test publishers often provide this information, but other references are available that compare various tests on key parameters. In the next section, we provide several resources to help prospective test users choose appropriate assessment instruments.

Resources for Test Information

The following resources are but a few of the tools available for selecting and evaluating tests. This list is not inclusive and we encourage you also to review test publishers' brochures and Internet resources.

Nonevaluative Descriptive Resources

Several resources assist test users in finding assessment instruments that measure specific traits. These resources provide information only about the test instrument itself, without any reviews or critiques. Hence, these resources are often used in conjunction with evaluative descriptive resources.

The newest edition of *Tests: A Comprehensive Reference for Assessments in Psychology, Education, and Business* (Maddox, 1996) is available from Pro-Ed, Inc. (website: www.proedinc.com). Currently in its fourth edition, this reference provides updated information on approximately 2,000 assessment instruments in the fields of psychology, education, and business. The following information is provided for each test: purpose, a concise description, scoring procedures, cost, and publisher contact information. A second nonevaluative resource is *Tests in Print,* a bibliography of all commercially available tests currently in print and available to users. The current edition, *Tests in Print VI* (Murphy, Plake, Impara, & Spies, 2002), is available through the Buros Institute of Mental Measurements at the University of Nebraska in Lincoln (website: www.unl.edu/buros/).

Evaluative Descriptive Resources

Once you have located specific tests that may fit your needs, we recommend you locate critiques of the tests. The nonevaluative descriptive resources provide information about psychometric properties (i.e., reliability and validity) of the test, but reviews and critiques provide information about the pros and cons of the use of the test. There are several convenient test-review resources available, two of which were mentioned previously.

As a joint project, the ERIC Clearinghouse on Assessment and Evaluation, the Library and Reference Services Division of the Educational Testing Service, the Buros Institute, the Region III Comprehensive Center at George Washington University, and Pro-Ed test publishers have created the Test Locator (available from www.ericae.net/testcol.htm). It contains descriptions of more than 10,000 tests and research instruments that are available through test publishers, and in journal articles or book chapters, as well as reviews and critiques. (A test review search is also offered at the Buros Institute website.)

Another resource available from the Buros Institute (www.unl.edu/buros/) is the *Mental Measurements Yearbook* (Plake, Impara, & Spies, 2003), which is available in hardback, on CD-ROM, and as Silver Platter services for libraries. This resource is a compilation of reviews and critiques for current assessment instruments.

Additionally, several publishing companies publish reviews of test instruments. For example, Pro-Ed, Inc. (www.proedinc.com) publishes *A User's Guide to Tests in Print,* currently in its second edition (Hammill, Brown, and Bryant, 1992). This book includes objective test evaluations with recommendation ratings based on accepted psychometric principles. It lists more than 250 tests, with more than 2,000 test scores reviewed. Another resource from the same publisher is *Test Critiques* (Keyser and Sweetland, 1994). This compilation contains reviews and in-depth studies of more than 800 of the most widely used assessment instruments. Each entry provides the reader with information on the practical applications and uses of the test; settings in which the test is used; appropriate and inappropriate subjects for the test; and guidelines for administration, scoring, and interpretation. Additional resources and references for information about test and testing issues can be found in chapter 53.

Summary

In the current educational environment, teachers are not only being challenged to become more knowledgeable about tests and test interpretation, but also being required to gain the knowledge and skills to select tests appropriately. Competency in test selection depends upon understanding the test's purpose, as well as knowing how to evaluate its quality. It is also important to research the usability of the instrument and its applicability in the particular setting where it will be used.

References

ACA & AAC. [American Counseling Association & the Association for Assessment in Counseling]. (1987). *Responsibilities of users of standardized tests: RUST statement revised.* Alexandria, VA: Authors.

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

▼AFT, NCME, & NEA [American Federation of Teachers, National Council on Measurement in Education, & National Education Association]. (1990) *Standards for teacher competence in educational assessment of students.* Washington, DC: Authors.

Anastasi, A. (1988). *Psychological testing.* New York: Macmillan.

Bennett, H. G., Seashore, H. G., & Wesman, A. G. (1991). *Differential Aptitude Tests for Personnel and Career Assessment technical manual.* San Antonio: The Psychological Corporation.

Bond, L. (1996). Norm- and criterion-referenced testing. In *Practical assessment research and evaluation.* College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Hammill, D., Brown, L., & Bryant, B. (1992). *A user's guide to tests in print* (2nd ed.). Austin, TX: Pro-Ed.

Hood, A .B., and Johnson, R. W. (1991). *Assessment in counseling: A guide to the use of psychological assessment procedures.* Alexandria, VA: American Association for Counseling and Development.

Hood, A. B., & Johnson, R. W. (1997). *Assessment in counseling: A guide to the use of psychological assessment procedures* (2nd ed.). Alexandria, VA: American Counseling Association.

✦JCTP [Joint Committee on Testing Practices]. (2002). *Code of fair testing practices in education.* Washington, DC: American Psychological Association.

Keyser, D., & Sweetland, R. (1994). *Test critiques* (Vols. 1–10). Austin, TX: Pro-Ed.

Klein, S. P., & Hamilton, L. (1999). *Large-scale testing: Current practices and new directions.* Santa Monica, CA: RAND.

Lyman, H. B. (1998). *Test scores and what they mean* (6th ed.). Boston: Allyn and Bacon.

Maddox, T. (1996). *Tests: A comprehensive reference for assessments in psychology, education, and business* (4th ed.). Austin, TX: Pro-Ed.

McMillan, J. H. (1997). *Classroom assessment: Principles and practice for effective instruction.* Needham Heights, MA: Allyn & Bacon.

Mehrens, W. A. (2001). Selecting a career assessment instrument. In J. T. Kapes and E. A. Whitfield (Eds.), *A counselor's guide to career assessment instruments* (4th ed.). Alexandria, VA: National Career Development Association.

Murphy, L. L., Plake, B. S., Impara, J. C., & Spies, R. A. (Eds.) (2002). *Tests in print VI.* Lincoln, NE: Buros Institute of Mental Measurements.

Osterlind, S. J. (1989). *Constructing test questions.* Dordrecht, The Netherlands: Kluwer Academic.

Plake, B. S., Impara, J. C., & Spies, R. A. (Eds.). (2003). *The fifteenth mental measurements yearbook.* Lincoln, NE: Buros Institute of Mental Measurements.

Ponterotto, J. G. (1996). Evaluating and selecting research instruments. In F. T. L. Leongs & J. T. Austin (Eds.), *The psychology research handbook: A guide for graduate students and research assistants* (pp. 72–82). Thousand Oaks, CA: Sage Publications.

Salvia, J., & Ysseldyke, J. E. (2001). *Assessment* (8th ed.). New York: Houghton Mifflin.

Seligman, L. (1980). *Assessment in developmental career counseling.* Cranston, RI: The Carroll Press.

Walsh, W. B., & Betz, N. E. (1995). *Tests and assessment* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

✦ Document is included in the Anthology of Assessment Resources CD
▼ Document is available on a website

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

☐ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☒ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").