EXHIBIT TTT-2

Case No. 1:14-cv-00857-TSC-DAR

9. TESTING INDIVIDUALS OF DIVERSE LINGUISTIC BACKGROUNDS

Background

For all test takers, any test that employs language is, in part, a measure of their language skills. This is of particular concern for test takers whose first language is nor the language of the test. Tesr use with individuals who have not sufficiently acquired the language of rhe test may introduce construct-irrelevanr components to rhe testing process. In such instances, test results may not reflect accurately the qualities and competencies intended to be measured. In addition, language differences are almost always associated with concomitant cultural differences that need to be taken into account when tests are used with individuals whose dominant language is different from that of the test. Whether a certain dialect of a language should be considered a differenr language cannot be resolved here, although some aspects of the present discussion are relevant to the debate. In either case, special attention to issues related to language and culture may be needed when developing, administering, scoring, and interpreting test scores and making decisions based on test scores. Language proficiency tests, if appropriately designed and used, are an obvious exception to rhis concern because they are intended to measure familiarity with the language of the test as required in educational and other settings.

Individuals who are bilingual can vary considerably in their ability to speak, write, comprehend aurally, and read in each language. These abilities are affected by the social or functional situations of communication. Some people develop socially and culturally acceptable ways of speaking that combine two or more languages simultaneously. Other individuals familiar with two languages may perform more slowly, less efficiently, and at rimes less accurately on problem-solving tasks that are administered in the less familiar language. Language dominance is not necessarily an indicator of language competence in taking a test, and some accommodation may be necessary even when administering the test in the more familiar language. Therefore ir is important to consider language background in developing, selecting, and administering tests and in interpreting tesr performance. Consequently, for example, test norms based on native speakers of English either should not be used with individuals whose first language is not English or such individuals' test results should be interpreted as reflecting in part current level of English proficiency rather than ability, potential, aptitude or personality characteristics or symptomatology. In cases where a language-oriented test is inappropriate due to the test takers' limited proficiency in that language, a nonverbal test may be a suitable alternative.

Where effective job performance requires the ability to communicate in the language of the test, persons who do not have adequate proficiency in that language may perform poorly on the test, on the job, or both. In that case, the tests used for prediction of future job performance appropriately would be administered in the language of the job, as long as rhe language level needed for the test did not exceed the level needed to meet work requirements. Test users should understand that poor test performance, as well as poor job performance, may result from poor language proficiency rather than other deficiencies.

Many issues addressed in this chapter are also relevant to testing individuals who have unique linguistic characteristics due to disabilities such as deafness and/or blindness. For example, issues regarding test translation and adaptation are applicable to American Sign Language (ASL) versions of traditional tests. It should be noted, however, that ASL is

not only a different language but is also a different mode of communication. Also, individuals with disabilities may require modifications in test administration procedures similar to those required by non-native speakers. A more specific discussion of testing individuals with disabilities is provided in chapter 10.

Issues discussed in earlier chapters, in particular chapters 1-5, including validity of test score inferences, test reliability, and test development and administration are germane to this chapter. The present chapter extends these discussions, emphasizing the importance of recognizing the possible impact of language abilities and skills on test performance. There may be legal requirements relevant to the testing of individuals with different language backgrounds. The standards in this chapter are intended to be applied in a manner consistent with those requirements.

Test Translation, Adaptation, and Modification

Testing test takers in their primary language may be necessary in order to draw valid inferences based on their test scores. Thus, language modifications are often needed. Translating a test to the primary language represents one such modification. However, a number of hazards need to be avoided when doing this sort of translation. One cannot simply assume that such a translation produces a version of the test that is equivalent in content, difficulty level, reliability, and validity to the original untranslated version. Further, one cannot assume that test takers' relevant acculturation experiences are comparable across the two versions. Also, many words have different frequency rates or difficulty levels in various languages. Therefore, words in two languages that appear to be close in meaning may differ significantly in ways that seriously impact the translated test for the intended test use. Additionally, the test content of the translated version may not be equivalent to that of the original version. For example, a test of reading skills in language A that is translated to serve as a test of reading skills in language B may include content not equally meaningful or appropriate for people who read only language B.

For the purposes of test translation and adaptation for use with test takers whose first language is nor the language of the test, back translation is not recommended as a standalone procedure. It may provide an artificial similarity of meaning across languages but not the best version in the new language. In most situations, an iterative process more akin to test development and validation is suggested to ensure that similar constructs are measured across versions. When test forms in two or more languages are developed concurrently, it is generally desirable that some items originate in each of the languages involved. The decision as to whether to use the standard original language test or an adapted version is a complex matter. Issues that may have an impact on this decision are discussed in the next section.

Other strategies of test modification may be appropriare when the test taker's primary language is not the language of the test. These include modifying aspects of the test or the test administration procedure such as the presentation format, the response format, the time allowed to complete the test, the test setting (individual administration instead of group testing), and the use of only those portions of the test that are appropriate for the level of language proficiency of the test taker. If modifications are made to the presentation or response format of the test, it may sometimes be appropriate for the modified test to be field tested with an adequate population sample prior to use with its intended population.

Issues of Equivalence

The term *equivalence*, as used here, refers to the degree to which test scores can be used to make comparable inferences for different

examinees. When tests are designed for and used with linguistically homogeneous populations, issues of equivalence are relatively straightforward (for example, see chapters 1 and 4). If an individual examinee can be demonstrated to belong to the population for which the test was designed, then adhering to standard procedures of test administration and interpretation is expected to lead to reliable and valid inferences based on the examinee's test score. When a test is intended for use with test takers who differ linguistically from those for whom the test was designed, establishing equivalence poses a greater challenge. In general, the linguistic and cultural characteristics of the intended examinee population should be reflected in examinee samples used throughout the processes of test design, validation, and norming. At each of these stages of test development and standardization, distinct linguistic groups should receive the same level of specific attention. The inclusion of proportional representation of linguistic subgroups in aggregate standardization and validation samples may be insufficient to assure equivalence across linguistic groups.

Issues associated with construct equivalence are perhaps most fundamental. One may question whether the test score for a particular individual represents that individual's standing with respect to the same construct as is measured in the target population. For example, among non-native speakers of the language of the test, one may not know whether a test designed to measure primarily academic achievement becomes in whole or in part a measure of proficiency in the language of the test. There are several psychometric techniques that can be used to determine the equivalence of constructs across groups, including confirmatory factor analysis, analysis of data contained in multimethod-multitrait matrices and the equivalence of responsiveness of the groups to experimental manipulations. These techniques may be supplemented with logical analyses of the results based on knowledge of the linguistic characteristics of the test taker's population of origin.

Other types of equivalence also need to be considered when testing individuals from different linguistic backgrounds. Functional equivalence addresses the question of whether similar activities or behaviors measured by a test have the same meaning in different cultural or linguistic groups. Translation equivalence requires that the translated or adapted test be comparable in content to the original test; it was addressed above in the discussion of test translation and adaptation. Metric equivalence concerns the issue of whether scores from the same test administered in different languages have comparable psychometric properties. For example, with metric equivalence, a score of 50 on test X in language A is interpretable in the same way as a score of 50 on test X in language B. In general, metric equivalence will be limited to particular contexts, examinee groups, and types of interpretations.

Language Proficiency Testing

Consideration of relevant within-linguistic group differences is crucial in determining appropriate test interpretation and decision making in educational programs and in some professional applications of individualized tests. For example, individuals whose first language is not the language of the test may vary considerably in their proficiency along a continuum from those who have no knowledge of the language of the test to those who are fluent in it and knowledgeable of the corresponding culture. Further, a demographic proxy such as Mexican or German is likely to prove insufficient in determining the language of test administration because members of the same cultural group may vary widely in their degree of acculturation, proficiency in the language of the test, familiarity with words and syntax in their native languages,

educational background, familiarity with tests and test-taking skills, and other factors that may significantly affect the reliability and validity of inferences drawn from test scores. Thus, it is essential that individual differences that may affect test performance be taken into account when testing individuals of differing linguistic backgrounds.

The need exists to consider both language dominance and language proficiency. Standardized tests that assess multiple domains in a given language can be helpful in determining language dominance and proficiency. The person conducting the testing first should obtain information about the language in which the examinee is dominant (i.e., the preferred or salient language). Following this determination of dominance, the examinee's level of proficiency in the dominant language should be established. If the languages are similarly dominant, then proficiency should be established for both (or all) languages. Then the test should be administered in the most proficient language if available (unless the purpose of the testing is to determine proficiency in the language of the test). However, testing individuals in their dominant language alone is no panacea because, as suggested above, a bilingual individual's two languages are likely to be specialized by domain (e.g., the first language is used in the context of home, religious practices, and native culture, whereas the second language is used in the context of school, work, television, and mainstream culture). Thus, a test in either language by itself will likely measure some domains and miss out on others. In such situations, testing in both languages (i.e., the dominant language and the language in which the test taker is most proficient) may be necessary, provided appropriate tests are available. If assessment in both languages is carried out, careful consideration should be given to the possibility of order effects.

Because students are expected to acquire proficiency in the language used in schools that is appropriate to their ages and educational levels, tests suitable for assessing their progress in that language are needed. For example, some tests, especially paper-and-pencil measures, that are prepared for students of English as a foreign language may not be particularly useful if they place insufficient emphasis on the assessment of important listening and speaking skills. Measures of competency in all relevant English language skills (e.g., communicative competence, literacy, grammar, pronunciation, and comprehension) are likely to be most valuable in the school context.

Observing students' speech in naturalistic situations can provide additional information about their proficiency in a language. However, findings from naturalistic observations may not be sufficient to judge students' ability to function in that language in formal, academically oriented situations (e.g., classrooms). For example, it is not appropriate to base judgments of a child's ability to benefit from instruction in one language solely on language fluency observed in speech use on the playground. Nor is it appropriate to base judgments of a person's ability to perform a job on assessments of formal language usage, if formal language usage is not linked to job performance.

In general, there are special difficulties attendant upon the use of a test with individuals who have not had an adequate opportunity to learn the language used by the test. When a test is used to inform a decision process that has a broad impact, it may be important for the test user to review the test itself and to consider the possible use of alternative information-gathering tools (e.g., additional tests, sources of observational information, modified forms of the chosen test) to ensure that the information obtained is adequate to the intended purpose. Reviews of this kind may sometimes reveal the need

to create a formal adaptation of a test or to develop a new test that is suitable for the specific linguistic characteristics of the individuals being tested.

Testing Bilingual Individuals

Test use with examinees who are bilingual also poses special challenges. An individual who knows two languages may not test well in either language. As an example, children from homes where parents speak Spanish may be able to understand Spanish but express themselves best in English. In addition, some persons who are bilingual use their native language in most social situations and use English primarily for academic and work-related activities; the use of one or both languages depends on the nature of the situation. As another example, proficiencies in conversational English and written English can often differ. Non-native English speakers who may give the impression of being fluent in conversational English may not be competent in taking tests that require English literacy skills. Thus, an understanding of an individual's type and degree of bilingualism is important to proper test use.

Administration and Examiner Variables

When an examinee cannot be assumed to belong to the cultural or linguistic population upon which the test was standardized, then use of standardized administration procedures may not provide a comparable administration of the test for that examinee. In this situation, the fundamental principle of sound practice is that examinees, regardless of background, should be provided with an adequate opportunity to complete the test and demonstrate their level of competence on the attributes the test is intended to measure. There may be, however, complex interactions among examiner, examinee, and situational variables that require careful attention on the part of the practitioner administering the test. Factors that may affect the performance of the examinee include the cultural and linguistic background of both the examiner and examinee; the gender and testing style of the examiner; the level of acculturation of the examinee and examiner; whether the test is administered in the original language of the test, the examinee's primary language, or whether both languages are used (and if so in what order); the time limits of the testing; and whether a bilingual interpreter is used.

Use of Interpreters in Testing

Ideally, when an adequately translated version of the test or a suitable nonverbal test is unavailable, assessment of individuals with limited proficiency in the language of the test should be conducted by a professionally trained bilingual examiner. The bilingual examiner should be proficient in the language of the examinee at the level of a professional trained in that language. When a bilingual examiner is not available, an alternative is to use an interpreter in the testing process and administer the test in the examinee's native language. Although a commonly used procedure, this practice has some inherent difficulties. For example, there may be a lack of linguistic and cultural equivalence between the translation and the original test, the translator or the interpreter may not be adequately trained to work in the testing situation, and representative norms may not be available to score and interpret the test results appropriately. These difficulties may pose significant threats to the validity of inferences based on test results.

When the need for an interpreter arises for a particular testing situation, it is important to obtain a fully qualified interpreter to assist the examiner in administering the test. The most important consideration in testing with the services of an interpreter is the inter-

preter's ability and preparedness in carrying out the required duties during testing. The interpreter obviously needs to be fluent in both the language of the test and the examinee's native language and have general familiarity with the process of translating. To be effective, the interpreter also needs to have a basic understanding of the process of psychological and educational assessment, including the importance of following standardized procedures, the importance of accurately conveying to the examiner an examinee's actual responses, and the role and responsibilities of the interpreter in testing. Additionally, it is inappropriate for the interpreter to have any prior personal relationship with the test taker that is likely to jeopardize the objectivity of the test administration. However, in small linguistic or cultural communities, speakers of the alternate languages are often known to each other. Therefore, in such cases, it is the responsibility of the test user or examiner to ensure that the interpreter has received adequate instruction in the principles of objective test administration and to assess preexisting biases so that test interpretations can take such factors into account. If clear biases are evident and cannot be ameliorated, then the examiner should make arrangements to obtain another interpreter.

Whenever proficiency in the language of the test is essential to job performance, use of a translator to assist a candidate with licensure, certification, or civil service examinations should be permitted only when it will not compromise standards designed to protect public health, safety, and welfare. When a translator is permitted, it also is essential that the candidate not receive help interpreting the content of the test or any other assistance that would compromise the integrity of the licensure or certification decision. Creation of audio tapes that enable a candidate to listen to each question being read in the language of the test may be more appropriate when such an accommodation is justified.

In educational and psychological testing, it may be appropriate for an interpreter to become familiar with all details of test content and administration prior to the testing. Also, time needs to be provided for the interpreter to translate test instructions and items, if necessary. In psychological testing, it is often desirable for the examiner to demonstrate for the interpreter how certain test items are administered and explain what to expect during testing. In addition, it is important that, prior to the testing, the examiner and the interpreter become familiar with each other's style of speaking and the speed at which they work. Immediately prior to the assessment, the role of the interpreter needs to be explained clearly to the examinee. It is essential that the interpreter make all efforts to provide accurate information in translation. The interpreter must reflect a professional attitude and maintain objectivity throughout the testing process (e.g., not interject subjective opinions, not give cues to the examinee). Once the testing is completed, the examiner is responsible for reviewing the test responses with the assistance of the interpreter. Responses that are difficult to interpret (e.g., vocabulary words), nontest behaviors that might have special meanings (e.g., body language), as well as language factors (e.g., mixed use of two languages) and cultural factors that might have an effect on testing results need to be discussed fully. This information is to be used then by the examiner in carefully evaluating the test results and drawing inferences from the results.

Cultural Differences and Individual Testing

Linguistic behavior that may appear eccentric or be judged to be less appropriate in one culture may be seen as more appropriate in another culture and may need to be taken into account during the testing process. For example, children or adults from some cul-

tures may be reluctant to speak in elaborate language to adults or people in higher status roles and instead may be encouraged to speak to such persons only in response to specific questions or with formulaic utterances. Thus, when tested, such test takers may respond to an examiner probing for elaborate speech with only short phrases or by shrugging their shoulders. Interpretations of scores resulting from such testing may prove to be inaccurate if this tendency is not properly taken into consideration. At the same time, the examinet should not presume that their reticence is necessarily a cultural characteristic. Additional information (e.g., prior observations or a family member's consultation) may be needed to discuss the extent of culture's possible influence on linguistic performance.

The values associated with the nature and degree of verbal output also may differ across cultures. One cultural group may judge verbosity or rapid speech as rude, whereas another may regard those speech patterns as indications of high mental ability or friendliness. An individual from one culture who is evaluated with values appropriate to another culture may be considered taciturn, withdrawn, or of low mental ability. Resulting interpretations and prescriptions of treatment may be invalid and potentially harmful to the individual being tested.

Standard 9.1

Testing practice should be designed to reduce threats to the reliability and validity of test score inferences that may arise from language differences.

Comment: Some tests are inappropriate for use with individuals whose knowledge of the language of the test is questionable. Assessment methods together with careful professional judgment are required to determine when language differences are relevant. Test users can judge how best to address this standard in a particular testing situation.

Standard 9.2

When credible research evidence reports that test scores differ in meaning across subgroups of linguistically diverse test takers, then to the extent feasible, test developers should collect for each linguistic subgroup studied the same form of validity evidence collected for the examinee population as a whole.

Comment: Linguistic subgroups may be found to differ with respect to appropriateness of test content, the internal structure of their test responses, the relation of their test scores to other variables, or the response processes employed by individual examinees. Any such findings need to receive due consideration in the interpretation and use of scores as well as in test revisions. There may also be legal or regulatory requirements to collect subgroup validity evidence. Not all forms of evidence can be examined separately for members of all linguistic groups. The validity argument may rely on existing research literature, for example, and such literature may not be available for some populations. For some kinds of evidence, separate linguistic subgroup analyses may not be feasible due to the limited number of cases available. Data may sometimes be accumulated so that these

analyses can be performed after the test has been in use for a period of time. It is important to note that this standard calls for more than representativeness in the selection of samples used for validation or norming studies. Rather, it calls for separate, parallel analyses of data for members of different linguistic groups, sample sizes permitting. If a test is being used while such data are being collected, then cautionary statements are in order regarding the limitations of interpretations based on test scores.

Standard 9.3

When testing an examinee proficient in two or more languages for which the test is available, the examinee's relative language proficiencies should be determined. The test generally should be administered in the test taker's most proficient language, unless proficiency in the less proficient language is part of the assessment.

Comment: Unless the purpose of the testing is to determine proficiency in a particular language or the level of language proficiency required for the test is a work requirement, test users need to take into account the linguistic characteristics of examinees who are bilingual or use multiple languages. This may require the sole use of one language or use of multiple languages in order to minimize the introduction of construct-irrelevant components to the measurement process. For example, in educational settings, testing in both the language used in school and the native language of the examinee may be necessary in order to determine the optimal kind of instruction required by the examinee. Professional judgement needs to be used to determine the most appropriate procedures for establishing relative language proficiencies. Such procedures may range from self-identification by examinees through formal proficiency testing.

Standard 9.4

Linguistic modifications recommended by test publishers, as well as the rationale for the modifications, should be described in detail in the test manual.

Comment: Linguistic modifications may be recommended for the original test in the primary language ot for an adapted version in a secondary language, or both. In any case, the test manual should provide appropriate infotmation regarding the recommended modifications, their rationales, and the appropriate use of scores obtained using these linguistic modifications.

Standard 9.5

When there is credible evidence of score comparability across regular and modified tests or administrations, no flag should be attached to a score. When such evidence is lacking, specific information about the nature of the modification should be provided, if permitted by law, to assist test users properly to interpret and act on test scores.

Comment: The inclusion of a flag on a test score where a linguistic modification was provided may conflict with legal and social policy goals promoting fairness in the treatment of individuals of diverse linguistic backgrounds. If a score from a modified administration is comparable to a score from a nonmodified administration, there is no need for a flag. Similarly, if a modification is provided for which there is no reasonable basis for believing that the modification would affect score comparability, there is no need for a flag. Furthet, reporting ptactices that use asterisks or other non-specific symbols to indicate that a test's administration has been modified provide little useful information to test users.

Standard 9.6

When a test is recommended for use with linguistically diverse test takers, test developers and publishers should provide the information necessary for appropriate test use and interpretation.

Comment: Test developers should include in test manuals and in instructions for score interpretation explicit statements about the applicability of the test with individuals who are not native speakers of the original language of the test. However, it should be recognized that test developers and publishers seldom will find it feasible to conduct studies specific to the large number of linguistic groups found in certain countries.

Standard 9.7

When a test is translated from one language to another, the methods used in establishing the adequacy of the translation should be described, and empirical and logical evidence should be provided for score reliability and the validity of the translated test's score inferences for the uses intended in the linguistic groups to be tested.

Comment: For example, if a test is translated into Spanish for use with Mexican, Puerto Rican, Cuban, Central American, and Spanish populations, score reliability and the validity of test score inferences should be established with members of each of these groups separately where feasible. In addition, the test translation methods used need to be described in detail.

Standard 9.8

In employment and credentialing testing, the proficiency level required in the language of the test should not exceed that appropriate to the relevant occupation or profession.

Comment: Many occupations and professions require a suitable facility in the language of the test. In such cases, a test that is used as a part of selection, advancement, or credentialing may appropriately reflect that aspect of performance. However, the level of language proficiency required on the test should be no greater than the level needed to meet work requirements. Similarly, the modality in which language proficiency is assessed should be comparable to that on the job. For example, if the job requires only that employees understand verbal instructions in the language used on the job, it would be inappropriate for a selection test to require proficiency in reading and writing that particular language.

Standard 9.9

When multiple language versions of a test are intended to be comparable, test developers should report evidence of test comparability.

Comment: Evidence of test comparability may include but is not limited to evidence that the different language versions measure equivalent or similar constructs, and that score reliability and the validity of inferences from scores from the two versions are comparable.

Standard 9.10

Inferences about test takers' general language proficiency should be based on tests that measure a range of language features, and not on a single linguistic skill.

Comment: For example, a multiple-choice, pencil-and-paper test of vocabulary does not indicate how well a person understands the language when spoken nor how well the person speaks the language. However, the test score might be helpful in determining how well a person understands some aspects of the written language. In making educational

placement decisions, a more complete range of communicative abilities (e.g., word knowledge, syntax) will typically need to be assessed.

Standard 9.11

When an interpreter is used in testing, the interpreter should be fluent in both the language of the test and the examinee's native language, should have expertise in translating, and should have a basic understanding of the assessment process.

Comment: Although individuals with limited proficiency in the language of the test should ideally be tested by professionally trained bilingual examiners, the use of an interpreter may be necessary in some situations. If an interpreter is required, the professional examiner is responsible for ensuring that the interpreter has the appropriate qualifications, experience, and preparation to assist appropriately in the administration of the test. It is necessary for the interpreter to understand the importance of following standardized procedures, how testing is conducted typically, the importance of accurately conveying to the examiner an examinee's actual responses, and the role and responsibilities of the interpreter in testing.

10. TESTING INDIVIDUALS WITH DISABILITIES

Background

With the advancement of scientific knowledge, medical practices, and social policies, increasing numbers of individuals with disabilities are participating more fully in educational, employment, and social activities. This inceased participation has resulted in a gteater need for the testing and assessment of individuals with disabilities for a variety of purposes. Individuals with disabilities are defined as petsons possessing a physical, mental, or developmental impairment that substantially limits one or more of their major life activities. Although the *Standards* focus on technical and professional issues regarding the testing of individuals with disabilities, test developers and users are encouraged to become familiar with federal, state, and local laws, and court and administrative rulings that regulate the testing and assessment of individuals with disabilities.

Tests are administeted to individuals with disabilities in various settings and for diverse purposes. For example, tests are used for diagnostic purposes to determine the existence and nature of a test taker's disabilities. Testing is also conducted for prescriptive purposes to determine intervention plans. In addition, tests are administered to persons who have been diagnosed with identified disabilities for educational and employment purposes to make placement, selection, or other similar decisions, or for monitoring performance as a tool for educational accountability. These uses of tests for persons with disabilities occur in a variety of contexts including school, clinical, counseling, forensic, employment, and credentialing.

Issues Regarding Accommodation When Testing Individuals With Disabilities

A major issue when testing individuals with disabilities concerns the use of accommodations, modifications, or adaptations. The purpose of these accommodations or modifications is to minimize the impact of test-taker attributes that are not relevant to the construct that is the primary focus of the assessment. The terms *accommodation* and *modification* have varying connotations in different subfields. Here accommodation is used as the general term for any action taken in response to a determination that an individual's disability requires a departure from established testing protocol. Depending on circumstances, such accommodation may include modification of test administration processes or modification of test content. No connotation that modification implies a change in the construct(s) being measured is intended.

A standardized test that has been designed fot use with the general population may be inappropriate fot use for individuals with specific disabilities if the test requires the use of sensory, motor, language, or psychological skills that are affected by the disability and that are not relevant to the focal construct. For example, a person who is blind may read only in Braille format, and an individual with hemiplegia may be unable to hold a pencil and thus would have difficulty completing a standard written exam. In addition, some individuals with disabilities may possess other attendant characteristics (e.g., a person with a physical disability may fatigue easily), causing them to be further challenged by some standardized testing situations. In these examples, if reading, use of a pencil, and fatigue are incidental to the construct intended to be measured by the test, modifications of tests and test administration procedures may be necessary for an accurate assessment.

Note also that accommodations are not needed or appropriate under a variety of circumstances. First, the disability may, in fact, be directly relevant to the focal construct. For example, no accommodation is appropriate for a person who is completely blind if the

test is designed to measure visual spatial ability. Similarly, in employment testing it would be inappropriate to make test modifications if the test is designed to assess essential skills required for the job and the modifications would fundamentally alter the constructs being measured. Second, an accommodation for a particular disability is inappropriate when the purpose of a test is to diagnose the presence and degree of that disability. For example, allowing extra time on a timed test to assess the existence of a specific learning disability would make it very difficult to determine if a processing difficulty actually exists. Third, it is important to note that not all individuals with disabilities require special provisions when taking all tests. Many individuals have disabilities that would not influence their performance on a particular test, and hence no modification is needed.

Professional judgment necessarily plays a substantial role in decisions about test accommodations. Judgment comes into play in determining whether a particular individual needs accommodation and the nature and extent of such accommodation. In some circumstances, individuals with disabilities request testing accommodations and provide appropriate documentation in support of the request. Generally the request is reviewed by the agency sponsoring the assessment or an outside source knowledgeable about the assessment process and the type of disability. In either case, a conclusion is drawn as to what constitutes reasonable accommodation. Disagreement may arise between the accommodation requested by an individual with a disability and the granted accommodation. In these situations, and to the extent permitted by law, the overarching concern is the validity of the inference made from the score on the modified test: fairness to all parties is best served by a decision about test modification that results in the most accurate measure possible of the construct of interest. The role of professional judgment is further complicated by the fact that empirical research on test accommodations is often lacking.

When modifying tests it is also important to recognize that individuals with the same type of disability may differ considerably in their need for accommodation. A central consideration in determining a test modification for a disability is to recognize that the modifications should be tailored directly to the specific needs of individual test takers. As an example, it would be incorrect to make the assumption that all individuals with visual impairments would be successfully accommodated by providing testing materials in Braille format. Depending on the extent of the disability, it may be more appropriate for some individuals to receive testing materials written in large print, while others might need a tape cassette or reader.

As test modifications involve altering some aspect of a test originally developed for use with a target population, it is important to recognize that making these alterations has the potential to affect the psychometric qualities of the test. There have been few empirical investigations into the effects of various accommodations on the reliability of test scores or the validity of inferences drawn from modified tests. Due to a number of practical limitations (e.g., small sample size, nonrandom selection of test takers with disabilities), there is no precise, technical solution available for equating modified tests to the original form of these tests. Thus it is difficult to compare scores from a test modified for persons with disabilities with scores from the original test.

Modifications designed to accommodate persons with disabilities also may change the construct measured by the test, or the extent to which it is fully measured. For example, a test of oral comprehension may become a test of reading comprehension when administered in written format to a person who is deaf or hard of hearing. Such a change in test administration may alter the construct being measured by the original test. When this occurs, the scores on the standard and modified versions of the test will not have the same meaning. Similarly, modification of test administration may also

alter the predictive value of test scores. For example, when a speed test is administered with relaxed time requirements to a person with a disability, the relationship of test scores to criteria such as job performance may be affected. Appropriate professional judgment should be exercised in interpreting and using scores on modified tests.

Some modified tests, with accompanying research to support the appropriate modifications, have been available for a number of years. Although the development of tests and testing procedures for individuals with disabilities is encouraged by the *Standards*, it should be noted that all relevant individual standards given elsewhere in this document are fully applicable to the testing applications and modifications or accommodations considered in this chapter. Issues of validity and reliability are critical whenever modifications or accommodations occur.

Strategies of Test Modification

A variety of test modification strategies have been implemented in various settings to accommodate the needs of test takers with disabilities. Some require modifying test administration procedures (e.g., instructions, response format) while others alter test medium, timing, settings, or content. Depending on the nature and extent of the disability, one or more test modification procedures may be appropriate for a particular individual. The listing here of a variety of modification strategies should not suggest that the full array of strategies is routinely available or appropriate; the decision to modify tests on a determination that modification is needed to make valid inferences about the individual's standing on the construct in question.

Modifying Presentation Format

One modification option is to alter the medium used to present the test instructions and items to the test takers. For example, a test booklet may be produced in Braille or large print for individuals with visual impairments. When tests are computer-administered, larger fonts or oversized computer screens may be used. Individuals with a hearing disability may receive test instructions through the use of sign communication or writing.

Modifying Response Format

Modifications also can be made to allow individuals with disabilities to respond to test items using their preferred communication modality. For example, an individual with severe language deficits might be allowed to point to the preferred response. A test taker who cannot manually record answers to test items or questions may be assisted by an aide who would mark the answer. Other ways of obtaining a response include having the respondent use a tape recorder, a computer keyboard, or a Braillewriter.

Modifying Timing

Another modification available is to alter the timing of tests. This may include extended time to complete the test, more breaks during testing, or extended testing sessions over several days. Many national testing programs (e.g., achievement, certification) allow persons with disabilities additional time to take the test. Reading Braille, using a cassette recorder, or having a reader may take longer than reading regular print. Reading large type may or may not be more time-consuming, depending on the layout of the material and on the nature and severity of the impairment.

Modifying Test Setting

Tests normally administered in group settings may be administered individually for a variety of purposes. Individual administration may avoid interference with others taking a test in a group. Some disabilities (e.g., attention deficit disorder) make it impractical to test in a group setting. Other alterations may include changing the testing location if it is not wheelchair accessible, providing tables or chairs that provide greater physical support, or altering the lighting conditions for individuals who are visually impaired.

Using Only Portions of a Test

Another strategy of test accommodation involves the use of portions of a test in assessing persons with disabilities. These procedures are sometimes used in clinical testing when certain subparts of a test require physical, sensory, language, or other capabilities that a test taker with disabilities does not have. This approach is commonly used in cognitive and achievement testing when the physical or sensory limitations of an individual interfere with the ability to perform on a test. For example, if a cognitive ability test includes items presented orally combined with items presented in a written fashion, the orally-presented items might be omitted when the test is given to an individual with a hearing disability as they will not provide an adequate assessment of that individual's cognitive ability. Results on such items are more likely to reflect the individual's hearing difficulty rather than his or her true cognitive ability. Although omitting test items may represent an effective accommodation technique, it may also prevent the test from adequately measuring the intended skills or abilities, especially if those skills or abilities are of central interest. For example, it should be noted that eliminating a portion of the test may not be appropriate in situations such as certification testing and employment testing where the construct measured by the each portion may represent a separate and necessary job or occupational requirement.

Using Substitute Tests or Alternate Assessments

One additional modification is to replace a test standardized on the general population with a test or alternate assessment that has been specially designed for individuals with disabilities. More valid results may be obtained through the use of a test specifically designed for use with individuals with disabilities. Although a substitute test may represent a desirable accommodation solution, it may be difficult to find an adequate replacement that measures the same construct with comparable technical quality, and for which scores can be placed on the same scale as the original test.

Using Modifications in Different Testing Contexts

There are important contextual differences between the individualized use of tests, as in the case of clinical diagnosis, and group or large-scale testing, as in the case of testing for academic achievement, employment, credentialing, or admissions.

Individual diagnostic testing is conducted typically for clinical or educational purposes. In these contexts a highly qualified test professional (e.g., a licensed or certified psychologist) is responsible for the entire assessment process of test selection, administration, interpretation, and reporting of results. The test professional seeks to gather appropriate information about the client's specific disability and preferred modality of communication and uses this information to determine the accommodations appropriate for the test taker. During the assessment process, any modified tests are used along with other assessment methods to collect data about the client's functioning in relevant areas. Inferences are then made based on this multitude of information. Test modifications may be used during assessment not only out of necessity but also as a source of clinical insight about the client's functioning. For example, a test taker with obsessive compulsive disorder may be allowed to continue to complete a test item, subtest, or a total test beyond the standardized time limits. Although in such cases the performance of the test taker cannot be judged according to the standardized scoring standards, the fact that the test taker could produce a successful performance with extra time often aids clinical interpretation.

The use of test modifications in large-scale testing is different, however. Large-scale testing is used for purposes such as measurement of academic achievement, program evaluation, credentialing, licensure, and employment. In these contexts, a standardized test usually is

administered to all test participants. Large numbers of test takers are not uncommon, and decisions may in some cases be made solely on the basis of test information, as in the case of a test used as an initial screening device in an employment context. In some cases, decision making requires the comparison of test takers, as in selection or admission contexts where the number of applicants may greatly exceed the number of available openings. This context highlights the need fot concern for fairness to all parties, as comparisons must be made between tesr scores obtained by individuals with disabilities taking modified tests and scores obtained by individuals under regular conditions. While test takers should not be disadvantaged due to a disability not relevant to the construct the test is intended to assess, the resulting accommodation should not put those taking a modified test at an undue advantage over those tested under regular conditions. As research on the comparability of scores under regular and modified conditions is sometimes limited, decisions about appropriate accommodation in these contexts involve important and difficult professional judgments.

Reporting Scores on Modified Tests

The practice of reporting scores on modified tests varies in different contexts. In individual testing, the test professional commonly reports when tests have been administered in a nonstandardized fashion when providing test scores. Typically, the steps used in making test accommodations or modifications are described in the test report, and the validity of the inferences resulting from the modified test scores is discussed. This practice of reporting the nature of modifications is consistent with implied requirements to communicate information as to the nature of the assessment process if the modifications impact the reliability of test scores or the validity of inferences drawn from test scores.

On the other hand, the reporting of test scores from modified tests in large-scale testing has created considerable debate. Often when scores from a nonstandardized version of a test are reported, the score report contains an asterisk next to the score or some other designation, often called a *flag*, to indicate that the test administration was modified. Sometimes recipients of these special designations are informed of the meaning of the designation; many times no information is provided about the nature of the modification made. Some argue that reporting scores from nonstandard test administrations without special identification misleads test users and perhaps even harms test takers with disabilities, whose scores may not accurately reflect their abilities. Others, however, argue that identifying scores of test takers with disabilities as resulting from nonstandard administrations unfairly labels these test takers as persons with disabilities, stigmatizes them, and may deny them the opportunity to compete equally with test takers without disabilities when they might otherwise be able to do so. Federal laws and the laws of most states bar discrimination against persons with disabilities, require individualized reasonable accommodations in testing, and limit practices that could stigmatize persons with disabilities, particularly in educational, admissions, credentialing, and employment testing.

The fundamental principles relevant here are that important information about test score meaning should not be withheld from test users who interpret and act on the test scores, and that irrelevant information should not be provided. When there is sufficient evidence of score comparability across regular and modified administrations, there is no need for any sort of flagging. When such evidence is lacking, an undifferentiated flag provides only very limited information to the test user, and specific information about the nature of the modification is preferable, if permitted by law.

Standard 10.1

In testing individuals with disabilities, test developers, test administrators, and test users should take steps to ensure that the test score inferences accurately reflect the intended construct rather than any disabilities and their associated characteristics extraneous to the intent of the measurement.

Comment: Chapter 1 (Validity) deals more broadly with the critical requirement that a test score reflects the intended construct. The need to attend to the possibility of construct-itrelevant variance resulting from a test taker's disability is an example of this general principle. In some settings, test users are prohibited from inquiring about a test taker's disability, making the standard contingent on test taker self-report of a disability or a need for accommodation.

Standard 10.2

People who make decisions about accommodations and test modification for individuals with disabilities should be knowledgeable of existing research on the effects of the disabilities in question on test performance. Those who modify tests should also have access to psychometric expertise for so doing.

Comment: In some areas there may be little known about the effects of a particular disability on performance on a particular type of test.

Standard 10.3

Where feasible, tests that have been modified for use with individuals with disabilities should be pilot tested on individuals who have similar disabilities to investigate the appropriateness and feasibility of the modifications.

Comment: Although useful guides for modifying tests are available, they do not provide a universal substitute for trying out a modified test. Even when such tryouts are conducted on samples inadequate to produce norm data, they are useful for checking the mechanics of the modifications. In many circumstances, however, lack of ready access to individuals with similar disabilities, or an inability to postpone decision making, make this unfeasible.

Standard 10.4

If modifications are made or recommended by test developers for test takers with specific disabilities, the modifications as well as the rationale for the modifications should be described in detail in the test manual and evidence of validity should be provided whenever available. Unless evidence of validity for a given inference has been established for individuals with the specific disabilities, test developers should issue cautionary statements in manuals or supplementary materials regarding confidence in interpretations based on such test scores.

Comment: When test developers and users intend that a modified version of a test should be interpreted as comparable to an unmodified one, evidence of test score comparability should be provided.

Standard 10.5

Technical material and manuals that accompany modified tests should include a careful statement of the steps taken to modify the tests to alert users to changes that are likely to alter the validity of inferences drawn from the test score.

Comment: If empirical evidence of the nature and effects of changes resulting from modifying standard tests is lacking, it is impossible to assess the impact of significant modifications. Documentation of the procedures used to modify tests will not only aid in the administration and intetpretation of the given test but will also inform others who are modifying tests for people with spe-

cific disabilities. This standard should apply to both test developers and test users.

Standard 10.6

If a test developer recommends specific time limits for people with disabilities, empirical procedures should be used, whenever possible, to establish time limits for modified forms of timed tests rather than simply allowing test takers with disabilities a multiple of the standard time. When possible, fatigue should be investigated as a potentially important factor when time limits are extended.

Comment: Such empirical evidence is likely only in the limited settings where a sufficient number of individuals with similar disabilities are tested. Not all individuals with the same disability, however, necessarily require the same accommodation. In most cases, professional judgment based on available evidence regarding the appropriate time limits given the nature of an individual's disability will be the basis for decisions. Legal requirements may be relevant to any decision on absolute time limits.

Standard 10.7

When sample sizes permit, the validity of inferences made from test scores and the reliability of scores on tests administered to individuals with various disabilities should be investigated and reported by the agency or publisher that makes the modification. Such investigations should examine the effects of modifications made for people with various disabilities on resulting scores, as well as the effects of administering standard unmodified tests to them.

Comment: In addition to modifying tests and test administration procedures for people who have disabilities, evidence of validity for inferences drawn from these tests is needed. Validation is the only way to amass knowledge about the usefulness of modified tests for people with disabilities. The costs of obtaining validity evidence should be considered in light of the consequences of not having usable information regarding the meanings of scores for people with disabilities. This standard is feasible in the limited circumstances where a sufficient number of individuals with the same level or degree of a given disability is available.

Standard 10.8

Those responsible for decisions about test use with potential test takers who may need or may request specific accommodations should (a) possess the information necessary to make an appropriate selection of measures, (b) have current information regarding the availability of modified forms of the test in question, (c) inform individuals, when appropriate, about the existence of modified forms, and (d) make these forms available to test takers when appropriate and feasible.

Standard 10.9

When relying on norms as a basis for score interpretation in assessing individuals with disabilities, the norm group used depends upon the purpose of testing. Regular norms are appropriate when the purpose involves the test taker's functioning relative to the general population. If available, normative data from the population of individuals with the same level or degree of disability should be used when the test taker's functioning relative to individuals with similar disabilities is at issue.

Standard 10.10

Any test modifications adopted should be appropriate for the individual test taker, while maintaining all feasible standardized features. A test professional needs to consider reasonably available information about each test taker's experiences, characteristics,

and capabilities that might impact test performance, and document the grounds for the modification.

Standard 10.11

When there is credible evidence of score comparability across regular and modified administrations, no flag should be attached to a score. When such evidence is lacking, specific information about the nature of the modification should be provided, if permitted by law, to assist test users properly to interpret and act on test scores.

Comment: The inclusion of a flag on a test score where an accommodation for a disability was provided may conflict with legal and social policy goals promoting fairness in the treatment of individuals with disabilities. If a score from a modified administration is comparable to a score from a nonmodified administration, there is no need for a flag. Similarly, if a modification is provided for which there is no reasonable basis for believing that the modification would affect score comparability, there is no need for a flag. Further, reporting practices that use asterisks or other nonspecific symbols to indicate that a test's administration has been modified provide little useful information to test users. When permitted by law, if a nonstandardized administration is to be reported because evidence does not exist to support score comparability, then this report should avoid referencing the existence or nature of the test taker's disability and should instead report only the nature of the accommodation provided, such as extended time for testing, the use of a reader, or the use of a tape recorder.

Standard 10.12

In testing individuals with disabilities for diagnostic and intervention purposes, the test should not be used as the sole indicator of the test taker's functioning. Instead, multiple sources of information should be used.

Comment: For example, when assessing the intellectual functioning of persons with mental retardation, results from an individually administered intelligence test are generally supplemented with other pertinent information, such as case history, information about school functioning, and results from other cognitive tests and adaptive behavior measures. In addition, at times a multidisciplinary evaluation (e.g., physical, psychological, linguistic, neurological, etc.) may be needed to yield an accurate picture of the person's functioning.

PART III

Testing Applications

11. THE RESPONSIBILITIES OF TEST USERS

Background

Previous chapters have dealt primarily with the responsibilities of those who develop, market, evaluate, or mandate the administration of tests and the rights and obligations of test takers. Many of the standards in these chapters, and in the chapters that follow, refer to the development of tests and their use in specific settings. The present chapter includes standards of a more general nature that apply in almost all measurement contexts. In particular, attention is centered on the responsibilities of those who may be considered the *users* of tests. This group includes psychologists, educators, and other professionals who select the specific instruments or supervise test administration—on their own authority or at the behest of others. It also includes all individuals who actively participate in the interpretation and use of test results, other than the test takers themselves.

It is presumed that a legitimate educational, psychological, or employment purpose justifies the time and expense of test administration. In most settings, the user communicates this purpose to those who have a legitimate interest in the measurement process and subsequently conveys the implications of examinee performance to those entitled to receive the information. Depending on the measurement setting, this group may include individual test takers, parents and guardians, educators, employers, policymakers, the courts, or the general public.

Where administration of tests or use of test data is mandated for a specific population by governmental authorities, educational institutions, licensing boards, or employers, the developer and user of an instrument may be essentially the same. In such settings, there often is no clear separation between the professional responsibilities of those who produce the instrument and those who administer the test and interpret the results. Instruments produced by independent publishers, on the other hand, present a somewhat different picture. Typically, these tests will be used with a variety of populations and for diverse purposes.

The conscientious developer of a standardized test attempts to screen and educate potential users. Furthermore, most publishers and test sponsors work vigorously to prevent the misuse of standardized measures and the misinterpretation of individual scores and group averages. Test manuals often illustrate sound and unsound interpretations and applications. Some identify specific practices that are not appropriate and should be discouraged. Despite the best efforts of test developers, however, appropriate test use and sound interpretation of test scores are likely to remain primarily the responsibility of the test user.

Test takers, parents and guardians, legislators, policymakers, the media, the courts, and the public at large often yearn for unambiguous interpretations of test data. In particular, they often tend to attribute positive or negative results, including group differences, to a single factor or to the conditions that prevail in one social institution—most often, the home or the school. These consumers of test data frequently press for explicit rationales for decisions that are based only in part on test scores. The wise test user helps all interested parties understand that sound decisions regarding test use and score interpretation involve an element of professional judgment. It is not always obvious to the consumers that the choice of various information-gathering procedures often involves experience that is not easily quantified or verbalized. The user can help them appreciate the fact that the weighting of quantitative data, educational and occupational information, behavioral observations, anecdotal reports, and other relevant data often cannot be specified precisely.

Because of the appearance of objectivity and numerical precision, test data are sometimes allowed to totally override other sources of evidence about test takers. There are circumstances in which selection based exclusively on test scores may be appropriate. For example, this may be the case in pre-employment screening. But in educational and psychological settings, test users are well advised, and may be legally required, to consider other relevant sources of information on test takers, not just test scores. In the latter situations, the psychologist or educator familiar with the local setting and with local test takers is best qualified to integrate this diverse information effectively.

As reliance on test results has grown in recent years, greater pressure has been placed on test users to explain to the public the rationale for test-based decisions. More than ever before, test users are called upon to defend their testing practices. They do this by documenting that their test uses and score interpretations are supported by measurement authorities for the given purpose, that the inferences drawn from their instruments are validated for use with a given population, and that the results are being used in conjunction with other information, not in isolation. If these conditions are met, the test user can convincingly defend the decisions made or the administrative actions taken in which tests played a part.

It is not appropriate for these *Standards* to dictate minimal levels of test-criterion correlation, classification accuracy, or reliability for any given purpose. Such levels depend on whether decisions must be made immediately on the strength of the best available evidence, however weak, or whether decisions can be delayed until better evidence becomes available. But it is appropriate to expect the user to ascertain what the alternatives are, what the quality and consequences of these alternatives are, and whether a delay in decision making would be beneficial. Cost-benefit compromises become necessary in test use, as they often are in test development. It should be noted, however, that in some contexts legal requirements may place limits on the extent to which such compromises can be made. As with standards for the various phases of test development, when relevant standards are not met in test use, the reasons should be persuasive. The greater the potential impact on test takers, for good or ill, the greater the need to identify and satisfy the relevant standards.

In selecting a test and interpreting a test score, the test user is expected to have a clear understanding of the purposes of the testing and its probable consequences. The knowledgeable user has definite ideas on how to achieve these purposes and how to avoid bias, unfairness, and undesirable consequences. In subscribing to these *Standards*, test publishers and agencies mandating test use agree to provide information on the strengths and weaknesses of their instruments. They accept the responsibility to warn against likely misinterpretations by unsophisticated interpreters of individual scores or aggregated data. However, the ultimate responsibility for appropriate test use and interpretation lies predominantly with the test user. In assuming this responsibility, the user must become knowledgeable about a test's appropriate uses and the populations for which it is suitable. The user must also become adept, particularly in statewide and community-wide assessment programs, in communicating the implications of test results to those entitled to receive them.

In some instances, users may be obligated to collect additional evidence about a test's technical quality. For example, if performance assessments are locally scored, evidence of the degree of inter-scorer agreement may be required. Users also should be alert to the probable local consequences of test use, particularly in the case of large-scale testing programs. If the same test material is used in successive years, users should actively monitor the program to ensure that reuse has not compromised the integrity of the results.

Some of the standards that follow reiterate ideas contained in other chapters, principally chapter 5 "Test Administration, Scoring, and Reporting," chapter 7 "Fairness in Testing and Test Use," chapter 8 "Rights and Responsibilities of Test Takers," and chapter 13 "Educational Testing and Assessment." This repetition is intentional. It permits an enumeration in one chapter of the major obligations that must be assumed largely by the test administrator and user, though these responsibilities may refer to topics that are covered more fully in other chapters.

Standard 11.1

Prior to the adoption and use of a published test, the test user should study and evaluate the materials provided by the test developer. Of particular importance are those that summarize the test's purposes, specify the procedures for test administration, define the intended populations of test takers, and discuss the score interpretations for which validity and reliability data are available.

Comment: A prerequisite to sound test use is knowledge of the materials accompanying the instrument. As a minimum, these include manuals provided by the test developer. Ideally, the user should be conversant with relevant studies reported in the professional literature. The degree of reliability and validity required for sound score interpretations depends on the test's role in the assessment process and the potential impact of the process on the people involved. The test user should be aware of legal restrictions that may constrain the use of the test. On occasion, professional judgment may lead to the use of instruments for which there is little documentation of validity for the intended purpose. In these situations, the user should interpret scores cautiously and take care not to imply that the decisions or inferences are based on test results that are well-documented with respect to reliability or validity.

Standard 11.2

When a test is to be used for a purpose for which little or no documentation is available, the user is responsible for obtaining evidence of the test's validity and reliability for this purpose.

Comment: The individual who uses test scores for purposes that are not specifically recommended by the test developer is responsible for collecting the necessary validity evidence. Support for such uses may sometimes be found in the professional literature. If previous evidence is not sufficient, then additional data should be

collected. The provisions of this standard should not be construed to prohibit the generation of hypotheses from test data. For example, though some clinical tests have limited or contradictory validity evidence for common uses, clinicians generate hypotheses based appropriately on examinee responses to such tests. However, these hypotheses should be clearly labeled as tentative. Interested parties should be made aware of the potential limitations of the test scores in such situations.

Standard 11.3

Responsibility for test use should be assumed by or delegated only to those individuals who have the training, professional credentials, and experience necessary to handle this responsibility. Any special qualifications for test administration or interpretation specified in the test manual should be met.

Comment: Test users should not attempt to interpret the scores of test takers whose special needs or characteristics are outside the range of the user's qualifications. This standard has special significance in areas such as clinical testing, forensic testing, testing in special education, testing people with disabilities or limited exposure to the dominant culture, and in other such situations where potential impact is great. When the situation falls outside the user's experience, assistance should be obtained. A number of professional organizations have codes of ethics that specify the qualifications of those who administer tests and interpret scores.

Standard 11.4

The test user should have a clear rationale for the intended uses of a test or evaluation procedure in terms of its validity and contribution to the assessment and decision-making process.

Comment: Justification for the role of each instrument in selection, diagnosis, classification, and decision making should be arrived at before test administration, not afterwards. Preferably, the rationale should be available in printed materials prepared by the test publisher or by the user.

Standard 11.5

Those who have a legitimate interest in an assessment should be informed about the purposes of testing, how tests will be administered, the factors considered in scoring examinee responses, how the scores are typically used, how long the records will be retained, and to whom and under what conditions the records may be released.

Comment: This standard has greater relevance and application to educational and clinical testing than to employment testing. In most uses of tests for screening job applicants and applicants to educational programs, for licensing professionals and awarding credentials, or for measuring achievement, the purposes of testing and the uses to be made of the test scores are obvious to the examinee. Nevertheless, it is wise to communicate this information at least briefly even in these settings. In some situations, however, the rationale for the testing may be clear to relatively few test takers. In such settings, a more detailed and explicit discussion may be called for. Retention and release of records, even when such release would clearly benefit the examinee, are often governed by statutes or institutional practices. As relevant, examinees should be informed about these constraints and procedures.

Standard 11.6

Unless the circumstances clearly require that the test results be withheld, the test user is obligated to provide a timely report of the results that is understandable to the test taker and others entitled to receive this information.

Comment: The nature of score reports is often dictated by practical considerations. In some

cases only a terse printed report may be feasible. In others, it may be desirable to provide both an oral and a written report. The interpretation should vary according to the level of sophistication of the recipient. When the examinee is a young child, an explanation of the test results is typically provided to parents or guardians. Feedback in the form of a score report or interpretation is not typically provided when tests are administered for personnel selection or promotion.

Standard 11.7

Test users have the responsibility to protect the security of tests, to the extent that developers enjoin users to do so.

Comment: When tests are used for purposes of selection, licensure, or educational accountability, the need for rigorous protection of test security is obvious. On the other hand, when educational tests are not part of a high-stakes program, some publishers consider teacher review of test materials to be a legitimate tool in clarifying teacher perceptions of the skills measured by a test. Consistency and clarity in the definition of acceptable and unacceptable practices is critical in such situations. When tests are involved in litigation, inspection of the instruments should be restricted—to the extent permitted by law—to those who are legally or ethically obligated to safeguard test security.

Standard 11.8

Test users have the responsibility to respect test copyrights.

Comment: Legally and ethically, test users may not reproduce copyrighted materials for routine test use without consent of the copyright holder. These materials—in both paper and electronic form—include test items, ancillary forms such as answer sheets or profile forms, scoring templates, conversion tables of raw scores to derived scores, and tables of norms.

Standard 11.9

Test users should remind test takers and others who have access to test materials that the legal rights of test publishers, including copyrights, and the legal obligations of other participants in the testing process may prohibit the disclosure of test items without specific authorization.

Standard 11.10

Test users should be alert to the possibility of scoring errors; they should arrange for rescoring if individual scores or aggregated data suggest the need for it.

Comment: The costs of scoring error are great, particularly in high-stakes testing programs. In some cases, rescoring may be requested by the test taker. If such a test taker right is recognized in published materials, it should be respected. In educational testing programs, users should not depend entirely on test takers to alert them to the possibility of scoring errors. Monitoring scoring accuracy should be a routine responsibility of testing program administrators wherever feasible.

Standard 11.11

If the integrity of a test taker's scores is challenged, local authorities, the test developer, or the test sponsor should inform the test takers of their relevant rights, including the possibility of appeal and representation by counsel.

Comment: Proctors in entrance or licensure testing programs may report irregularities in the test process that result in challenges. University admissions officers may raise challenges when test scores are grossly inconsistent with other applicant information. Test takers should be apprised of their rights in such situations.

Standard 11.12

Test users or the sponsoring agency should explain to test takers their opportunities, if any, to retake an examination; users should also indicate whether the earlier as well as later scores will be reported to those entitled to receive the score reports.

Comment: Some testing programs permit test takers to retake an examination several times, to cancel scores, or to have scores withheld from potential recipients. If test takers have such privileges, they and score recipients should be so informed.

Standard 11.13

When test-taking strategies that are unrelated to the domain being measured are found to enhance or adversely affect test performance significantly, these strategies and their implications should be explained to all test takers before the test is administered. This may be done either in an information booklet or, if the explanation can be made briefly, along with the test directions.

Comment: Test-taking strategies, such as guessing, skipping time-consuming items, or initially skipping and then returning to difficult items as time allows, can influence test scores positively or negatively. The effects of various strategies depend on the scoring system used and aspects of item and test design such as speededness or the number of response alternatives provided in multiple-choice items. Differential use of such strategies by test takers can affect the validity and reliability of test score interpretations. The goal of test directions should be to convey information on the possible effectiveness of various strategies and, thus, to provide all test takers an equal opportunity to perform optimally. The use of such strategies by all test takers should be encouraged if their effect facilitates performance and discouraged if their effect interferes with performance.

Standard 11.14

Test users are obligated to protect the privacy of examinees and institutions that are involved in a measurement program, unless a disclosure of private information is agreed upon, or is specifically authorized by law.

Comment: Protection of the privacy of individual examinees is a well-established principle in psychological and educational measurement. In some instances, test takers and test administrators may formally agree to a lesser degree of protection than the law appears to require. In other circumstances, test users and testing agencies may adopt more stringent restrictions on the communication and sharing of test results than relevant law dictates. The more rigorous standards sometimes arise through the codes of ethics adopted by relevant professional organizations. In some testing programs the conditions for disclosure are stated to the examinee prior to testing, and taking the test can constitute agreement for the disclosure of test score information as specified. In other programs, the test taker or his/her parents or guardians must formally agree to any disclosure of test information to individuals or agencies other than those specified in the test administrator's published literature. It should be noted that the right of the public and the media to examine the aggregate test results of public school systems is guaranteed in some states.

Standard 11.15

Test users should be alert to potential misinterpretations of test scores and to possible unintended consequences of test use; users should take steps to minimize or avoid foreseeable misinterpretations and unintended negative consequences.

Comment: Well-meaning, but unsophisticated, audiences may adopt simplistic interpretations of test results or may attribute high or low scores or averages to a single causal factor.

Experienced test users can sometimes anticipate such misinterpretations and should try to prevent them. Obviously, not every unintended consequence can be anticipated. What is required is a reasonable effort to prevent negative consequences and to encourage sound interpretations.

Standard 11.16

Test users should verify periodically that their interpretations of test data continue to be appropriate, given any significant changes in their population of test takers, their modes of test administration, and their purposes in testing.

Comment: Over time, a gradual change in the demographic characteristics of an examinee population may significantly affect the inferences drawn from group averages. The accommodations made in test administration in recognition of examinee disabilities or in response to unforeseen circumstances may also affect interpretations.

Standard 11.17

In situations where the public is entitled to receive a summary of test results, test users should formulate a policy regarding timely release of the results and apply that policy consistently over time.

Comment: In school testing programs, districts commonly viewed as a coherent group may avoid controversy by adopting the same policies regarding the release of test results. If one district routinely releases aggregated data in much greater detail than another, groundless suspicions can develop that information is being suppressed in the latter district.

Standard 11.18

When test results are released to the public or to policymakers, those responsible for the release should provide and explain any supplemental information that will minimize possible misinterpretations of the data.

Comment: Preliminary briefings prior to the release of test results can give reporters for the news media an opportunity to assimilate relevant data. Misinterpretation can often be the result of the limited time reporters have to prepare media reports or inadequate presentation of information that bears on test score interpretation. It should be recognized, however, that the interests of the media are not always consistent with the intended purposes of measurement programs.

Standard 11.19

When a test user contemplates an approved change in test format, mode of administration, instructions, or the language used in administering the test, the user should have a sound rationale for concluding that validity, reliability, and appropriateness of norms will not be compromised.

Comment: In some instances, minor changes in format or mode of administration may be reasonably expected, without evidence, to have little or no effect on validity, reliability, and appropriateness of norms. In other instances, however, changes in format or administrative procedures can be assumed a priori to have significant effects. When a given modification becomes widespread, consideration should be given to validation and norming under the modified conditions.

Standard 11.20

In educational, clinical, and counseling settings, a test taker's score should not be interpreted in isolation; collateral information that may lead to alternative explanations for the examinee's test performance should be considered.

Comment: It is neither necessary nor feasible to make an intensive review of every test taker's

score. In some settings there may be little or no collateral information of value. In counseling, clinical, and educational settings, however, considerable relevant information is likely to be available. Obvious alternative explanations of low scores include low motivation, limited fluency in the language of the test, unfamiliarity with cultural concepts on which test items are based, and perceptual or motor impairments. In clinical and counseling settings, the test user should not ignore how well the test taker is functioning in daily life.

Standard 11.21

Test users should not rely on computer-generated interpretations of test results unless they have the expertise to consider the appropriateness of these interpretations in individual cases.

Comment: The scoring agency has the responsibility of documenting the basis for the interpretations. The user of a computerized scoring and reporting service has the obligation to be familiar with the principles on which such interpretations were derived. The user should have the ability to evaluate a computer-based score interpretation in the light of other relevant evidence on each test taker. Automated, narrative reports are not a substitute for sound professional judgment.

Standard 11.22

When circumstances require that a test be administered in the same language to all examinees in a linguistically diverse population, the test user should investigate the validity of the score interpretations for test takers believed to have limited proficiency in the language of the test.

Comment: The achievement, abilities, and traits of examinees who do not speak the language of the test as their primary language may be seriously mismeasured by the test. The scores of test takers with severe linguistic limitations will probably be meaningless. If language proficiency is not relevant to the purposes of testing, the test user should consider excusing these individuals, without prejudice, from taking the test and substituting alternative evaluation methods. However, it is recognized that such actions may be impractical, unnecessary, or legally unacceptable in some settings.

Standard 11.23

If a test is mandated for persons of a given age or all students in a particular grade, users should identify individuals whose disabilities or linguistic background indicates the need for special accommodations in test administration and ensure that these accommodations are employed.

Comment: Appropriate accommodations depend upon the nature of the test and the needs of the test taker. The mandating authority has primary responsibility for defining the acceptable accommodations for various categories of test takers. The user must take responsibility for identifying those test takers who fall within these categories and implement the appropriate accommodations.

Standard 11.24

When a major purpose of testing is to describe the status of a local, regional, or particular examinee population, the program criteria for inclusion or exclusion of individuals should be strictly adhered to.

Comment: In census-type programs, biased results can arise from the exclusion of particular subgroups of students. Financial and other advantages may accrue either from exaggerating or from reducing the proportion of high-achieving or low-achieving students. Clearly, these are unprofessional practices.

12. PSYCHOLOGICAL TESTING AND ASSESSMENT

Background

This chapter addresses issues important to professionals who use psychological tests with their clients. Topics include test selection and administration, test interpretation, collateral information used in psychological testing, types of tests, and purposes of testing. The types of psychological tests reviewed in this chapter include cognitive and neuropsychological; adaptive, social, and problem behavior; family and couples; personality; and vocational. In addition, the chapter includes an overview of four common uses of psychological tests: diagnosis; intervention planning and outcome evaluation; legal and governmental decisions; and personal awareness, growth, and action.

Employment testing is another context in which psychological testing is used. The standards in this chapter are applicable to those employment settings in which individual in-depth assessment is conducted (e.g., an evaluation of a candidate for a senior executive position). Employment settings in which tests are designed to measure specific job-related characteristics across multiple candidates are treated in the text and standards of chapter 14.

For all professionals who use tests, knowledge of cultural background and physical capabilities that influence (a) a test taker's development, (b) the methods for obtaining and conveying information, and (c) the planning and implementation of interventions is critical. Therefore, readers are encouraged to review chapters 7, 8, 9, and 10 that discuss fairness and bias in testing, the rights and responsibilities of test takers, testing individuals of diverse linguistic backgrounds, and testing individuals with disabilities. Readers will find important additional detail on validity; reliability; test development; scaling; test administration, scoring, and reporting; and general responsibilities of test users in chapters 1, 2, 3, 4, 5, and 11, respectively.

The use of tests provides one method of collecting information within the larger framework of a psychological assessment of an individual. Typically, psychological assessments involve an interaction between a professional who is trained and experienced in testing and a client. Clients may include patients, counselees, parents, employees, employers, attorneys, students, and other responsible parties who are test takers or who use the test results contained in psychological reports.

The results from tests and inventories, used within the context of a psychological assessment, may help the professional to understand the client more fully and to develop more informed and accurate hypotheses, inferences, and decisions about a client's situation. A psychological assessment is a comprehensive examination undertaken to answer specific questions about a client's psychological functioning during a particular time interval or to predict a client's psychological functioning in the future. An assessment may include administering and scoring tests, and interpreting test scores, all within the context of the individual's personal history. Inasmuch as test scores characteristically are interpreted in the context of other information about the client, an individual psychological assessment usually also includes interviewing the client; observing client behavior; reviewing educational, psychological, and other relevant records; and integrating these findings with other information that may be provided by third parties. The tasks of a psychological assessment—collecting, evaluating, integrating, and reporting salient information relevant to those aspects of a client's functioning that are under examination—comprise a complex and sophisticated set of professional activities.

The interpretation of tests and inventories can be a valuable part of the intervention process and, if used appropriately, can provide useful information to clients as well as to other users

of the test interpretation. For example, the results of tests and inventories may be used to assess the psychological functioning of an individual; to assign diagnostic classifications; to detect neuropsychological impairment; to assess cognitive and personality strengths, vocational interests, and values; to determine developmental stages; and to evaluate treatment outcomes. Test results also may provide information used to make decisions that have a powerful and lasting impact on people's lives (e.g., vocational and educational decision making; diagnosis; treatment planning; selection decisions; intervention and outcome evaluation; parole, sentencing, civil commitment, child custody, and competency to stand trial decisions; and personal injury litigation).

Test Selection and Administration

Prior to beginning the assessment process, the test taker should understand who will have access to the test results and the written report, how test results will be shared with the test taker, and if and when decisions based on the test results will be shared with the test taker and/or a third party. The assessment process begins by clarifying, as much as is possible, the reasons for which a client is presented for assessment. Guided by these reasons or other relevant concerns, the tests, inventories, and diagnostic procedures to be used are chosen, and other sources of information needed to evaluate the client and the referral issues are identified. The professional reviews more than the name of the test in choosing a test and is guided by the validity and reliability evidence and the applicability of the normative data available in the test's accumulated research literature. In addition to being thoroughly versed in proper administrative procedure, the professional is responsible for being familiar with the validity and reliability evidence for the intended use and purposes of the tests and inventories selected and for being prepared to develop a logical analysis that supports the various facets of the assessment and the inferences made from the assessment.

Validity and reliability considerations are paramount, but the demographic characteristics (e.g., gender, age, income, sociocultural and language background, education and other socioeconomic variables) of the group for which the test was originally constructed and for which initial and subsequent normative data are available also are important test selection issues. Selecting a test with demographically appropriate normative groups relevant for the client being tested is important to the generalizability of the inferences that the professional seeks to make. Sometimes the items or tasks contained in a test are designed for a particular group and are viewed as irrelevant for another group. A test constructed for one group may be applied to other groups with appropriate qualifications that explain the test choice based on the supporting research data and on professional experience.

The selection of psychological tests and inventories, for a particular client, often is individualized. However, in some settings a predetermined battery of tests may be taken by all participants, and group interpretations may be provided. The test taker may be a child, an adolescent, or an adult. The settings in which the tests or inventories are used include (but are not limited to) preschool, elementary, middle, or secondary schools; colleges or universities; pre-employment or employment settings; mental health or outpatient clinics; hospitals; prisons; or professionals' offices.

Professionals who oversee testing and assessment are responsible for ensuring that all persons who administer and score tests have received the appropriate education and training needed to perform these tasks. In addition, they are responsible in group testing situations for ensuring that the individuals who use the test results are trained to interpret the scores properly.

When conducting psychological testing, standardized test administration procedures should be followed. When nonstandard administration procedures are needed, they are to be described and justified. Professionals

also are responsible for ensuring that testing conditions are appropriate. For example, the examiner may need to determine if the client is capable of reading at the level required, and if clients with vision, hearing, or neurological disabilities are adequately accommodated. Finally, professionals are responsible for protecting the confidentiality and security of the test results and the testing materials.

One advantage of individually administered measures is the opportunity to observe and adjust testing conditions as needed. In some circumstances, test administration may provide the opportunity for skilled examiners to carefully observe the performance of persons under standardized conditions. For example, their observations may allow them to more accurately record behaviors being assessed, to understand better the manner in which persons arrive at their answers, to identify personal strengths and weaknesses, and to make modifications in the testing process. Thus, the observations of trained professionals can be important to all aspects of test use.

Test Score Interpretation

Test scores ideally are interpreted in light of the available normative data, the psychometric properties of the test, the temporal stability of the constructs being measured, and the effect of moderator variables and demographic characteristics (e.g., gender, age, income, sexual orientation, sociocultural and language background, education, and other socioeconomic variables) on test results. The professional rarely has the resources available to personally conduct the research or to assemble representative norms needed to make accurate inferences about each individual client's current and future functioning. Therefore, the professional may rely on the research and the body of scientific knowledge available for the test that warrants appropriate inferences. Presentation and analyses of validity and reliability evidence often are not needed in a written report, but the professional strives to understand, and prepares to articulate, such evidence as the need arises.

Tests and inventories that meet high technical standards of quality are a necessary but not a sufficient condition to ensure the responsible use and interpretation of test scores. The level of competence of the professional who interprets the scores and integrates the inferences derived from psychological tests depends upon the educational and experiential qualifications of the professional. With experience, professionals learn that the challenges in psychological test score interpretation increase in magnitude along a continuum of professional judgment with brief screening inventories at one end of the continuum and comprehensive multidimensional assessments at the other. For example, the interpretations of achievement and ability test scores, personality test scores, and batteries of neuropsychological test scores represent points on a continuum that require increasing levels of specialized knowledge, judgment, and skill by an experienced professional regardless of the soundness of the technical characteristics of the tests being used. The education and experience necessary to administer group tests and/or proctor computer-administered tests generally are less stringent than are the qualifications necessary to interpret individually administered tests. The use and interpretation of individually administered tests requires completion of rigorous educational and applied training, a high degree of professional judgment, appropriate credentialing, and adherence to the professional's ethical guidelines.

When making inferences about a client's past, present, and future behaviors and other characteristics from test scores, the professional reviews the literature to develop familiarity with supporting evidence. When there is strong evidence supporting the reliability and validity of a test, including its applicability to the client being assessed, the professional's ability to draw inferences increases. Nevertheless, the professional still corroborates results from testing with additional information from a variety of sources

such as interviews and results from other tests. When an inference is based on a single study or based on several studies whose samples are not representative of the client, the professional is more cautious about the inferences. Corroborating data from the assessment's multiple sources of information—including stylistic and test-taking behaviors inferred from observations during the test—will strengthen the confidence placed in the inference. Importantly, data that are not supportive of the inference are acknowledged and either reconciled or noted as limits to the confidence placed in the inference.

An interpretation of a test taker's test scores based upon existing research examines not only the demonstrated relationship between the scores and the criterion or criteria, but also the appropriateness of the latter. The criterion and the chosen predictor test or tests are subjected to a similar examination to understand the degree to which their underlying constructs are congruent with the inferences under consideration.

Threats to the interpretability of obtained scores are minimized by clearly defining how particular psychological tests are used. These threats occur as a result of construct-irrelevant variance (i.e., aspects of the test that are not relevant to the purpose of the test scores) and construct underrepresentation (i.e., important facets relevant to the purpose of the testing, but for which the test does not account). A client's response bias is another example of a construct-irrelevant component that may significantly skew the obtained scores, possibly rendering the scores uninterpretable. In situations where response bias is anticipated, the professional may choose a test that has scales (e.g., faking good, faking bad, social desirability, percent yes, percent no) that clarify the threats to validity from the test taker's response bias. In so doing, the professional may be able to assess the degree to which test takers are acquiescing to the perceived demands of the test administrator or attempting to portray themselves as impaired by "faking bad," or well-functioning by "faking good." In interpreting the test taker's obtained response bias score(s), the evidence of validity for constructs underlying each response bias scale, each scale's internal consistency, its interrelations with other scales, and evidence of validity are considered.

For some purposes, including career counseling and neuropsychological assessment, test batteries frequently are used. Such batteries often include tests of verbal ability, numerical ability, nonverbal reasoning, mechanical reasoning, clerical speed and accuracy, spatial ability, and language usage. Some batteries also include interest and personality inventories. When psychological test batteries incorporate multiple methods and scores, patterns of test results frequently are interpreted to reflect a construct or even an interaction among constructs underlying test performances. Higher order interactions among the constructs underlying configurations of test outcomes may be postulated on the basis of test score patterns. The literature reporting evidence of reliability and validity that supports the proposed interpretations should be identifiable. If the literature is incomplete, the resulting inferences may be presented with the qualification that they are hypotheses for future verification rather than probabilistic statements that imply some known validity evidence.

Collateral Information Used in Psychological Testing and Psychological Assessment

The quality of psychological testing and psychological assessment is enhanced by obtaining credible collateral information from various third-party sources such as teachers, personal physicians, family members, and school or employment records. Psychological testing also is enhanced by using various methods to acquire information. Structured behavioral observations, checklists and ratings, interviews, and criterion- and norm-referenced measures are but a few of the methods that may be used to acquire information. The use of psychological tests also can be enhanced by acquiring information about multiple traits or attributes to help characterize a person. For example, an

evaluation of career goals may be enhanced by obtaining a history of current and prior employment as well as by administering tests to assess academic aptitude and achievement, vocational interests, work values, and personality and temperament characteristics. The availability of information on multiple traits or attributes, when acquired from various sources and through the use of various methods, enables professionals to assess more accurately an individual's psychosocial functioning and facilitates more effective decision making.

Types of Psychological Tests

For purposes of this chapter, the types of psychological tests have been divided into five categories: cognitive and neuropsychological tests; adaptive, social, and problem behavior tests; family and couples tests; personality tests; and vocational tests.

Cognitive and Neuropsychological Testing

Tests often are used to assess various classes of cognitive and neuropsychological functioning including intelligence; broad ability domains (e.g., verbal, quantitative, and spatial abilities); and more focused domains (e.g., attention, sensorimotor functions, perception, learning, memory, reasoning, executive functions, and language). Overlap may occur in the constructs that are assessed by tests of differing functions or domains. In common with other types of tests, cognitive and neuropsychological tests require a minimally sufficient level of test-taker attentional capacity.

Cognitive Ability. Measures designed to quantify cognitive abilities are among the most widely administered tests. The interpretation of cognitive ability tests is guided by the theoretical constructs used to develop the test.

Many cognitive ability tests consist of multidimensional test batteries that are designed to assess a broad range of abilities and skills. Individually administered test batteries also are required for testing for purposes such as diagnosing a cognitive disorder. Test results are used to draw inferences about a person's overall level of intellectual functioning as well as strengths and weaknesses in various cognitive abilities. Because each test in a battery examines a different function, ability, skill, or combination thereof, the test taker's performance can be understood best when scores are not combined or aggregated, but rather when each score is interpreted within the context of all other scores and other assessment data. For example, low scores on timed tests alert the examiner to slowed responding as a problem that may not be apparent if scores on different kinds of tests are combined.

Attention. Attention refers to that class of functioning that encompasses arousal, establishment and deployment of sets, sustained attention, and vigilance as constructs. Tests may measure levels of alertness, orientation, and localization; the ability to focus, shift, and maintain attention and to track one or more stimuli under various conditions; span of attention; information processing speed and choice reaction time; and short-term information storage capacity. Scores for each aspect of attention that has been examined should be reported individually so that the nature of an attention disorder can be clarified.

Motor, Sensorimotor Functions, and Lateral Preferences. Visual, auditory, somatosensory and other sensory sensitivity and discrimination can be measured by simple motor or verbal responses to selective stimulation upon command.

Perception and Perceptual Organization/Integration. This class of functioning involves reasoning and judgment as they relate to the processing and elaboration of complex sensory combinations and inputs. Tests of perception may emphasize immediate perceptual processing but also may require conceptualizations that involve some reasoning and judgmental processes. Some tests have a motor component ranging from a simple motor response to an elaborate construction. Also,

some of these tests penalize the test taker for slow performance that may be caused by something other than perceptual dysfunction.

Learning and Memory. This class of functions involves the acquisition and retention of information beyond the attentional requirements of immediate or short-term information processing and storage. These tests may measure acquisition of new information through various sensory channels and by means of assorted test formats (e.g., word lists, prose passages, geometric figures, formboards, digits, and musical melodies). Memory tests also may require retention and recall of old information (e.g., petsonal data as well as commonly learned facts and skills).

Abstract Reasoning and Categorical Thinking. Tests of reasoning and thinking vary widely. They assess the examinee's ability to infer relationships or to respond to changing environmental circumstances and to act in goal-oriented situations.

Executive Functions. This class of functions is involved in the organized performances that are necessary for the independent, purposive and effective attainment of personal goals in various cognitive processing, problem-solving and social situations. Some tests emphasize reasoned plans of action that anticipate consequences of alternative solutions, motor performance in problem-solving situations that require goal-oriented intentions, and regulation of performance for achieving a desired outcome.

Language. Language assessment typically focuses on phonology, morphology, syntax, semantics, and pragmatics. Receptive and expressive language functions may be assessed, including listening, reading, talking, and written language skills and abilities. Assessment of central language disorders focuses on functional speech and verbal comprehension measured through oral, written, or gestural modes; lexical access and elaboration; repetition of spoken language; and associative verbal fluency.

When assessing persons who are nonnative English speakers or who are bilingual or multilingual, language assessment often includes an assessment of language competence and the order of dominance among the different languages. If a multilingual person is assessed for a possible language disorder, one issue for the professional to consider is the degree to which the disorder may be due more directly to language-related qualities (e.g., phonological, morphological, syntactic, semantic, pragmatic delays; mental retardation; peripheral sensory or central neurological impairment; psychological conditions; hearing disorders) than to dominance of a non-English language.

Academic Achievement. Academic achievement tests are measures of academic knowledge and skills that a person has acquired in formal and informal learning opportunities. Two major types of academic achievement tests include general achievement batteties and diagnostic achievement tests. General achievement batteries are designed to assess a person's level of learning in multiple areas (e.g., reading, mathematics, spelling, social studies, science). Diagnostic achievement tests, on the other hand, typically focus on one particular subject area (e.g., reading) and assess important academic skills in greater detail. Test results are used to determine the test taker's strengths as well as specific difficulties and may help identify sources of the difficulties and ways to overcome them. Chapter 13 provides additional detail on academic achievement testing in educational settings.

Social, Adaptive, and Problem Behavior Testing

Measures of social, adaptive, and problem behaviors assess ability and motivation to care for one's self and to relate to others. Adaptive behaviors include a repertoire of knowledge, skills, and abilities that enable a person to meet the daily demands and expectations of the environment, such as eating, dressing, using transportation, interacting with peers, communicating with others, making purchases, managing money, maintaining a schedule, remaining in school, and maintaining a job.

Problem behaviors include behavioral adjustment difficulties that interfere with a person's effective functioning in daily life situations.

Family and Couples Testing

Family testing addresses the issues of family dynamics, cohesion, and interpersonal relations among family members including partners, parents, children, and exrended family members. Tests developed to assess families and couples are distinguished by measuring the interaction patterns of partial or whole families, requiring simultaneous focus on two or more family members in terms of their transactions. Testing with couples may address personal factors such as issues of intimacy, compatibility, shared interests, trust, and spiritual beliefs.

Personality Testing

Broadly considered, the assessment of personality rcquires a synrhcsis of aspccts of an individual's functioning that contribute to the formulation and expression of thoughts, attitudes, emotions, and behaviors. In the assessment of an individual, cognitive and emotional functioning may be considered separately, but their influences are interrelated. Fot example, a person whose petceptions are highly accurate, or who is relatively stable emotionally, may be able to control suspiciousness better than can a person whose perceptions are inaccurate or disrorted or who is emotionally unstable.

Scores on a personality test may be regarded as reflecting the underlying theotetical constructs or empirically derived scales or factors that guided the test's construction. The stimulus and response formats of personality tests vary widely. Some include a series of questions (e.g., self-report inventories) to which the test taket is tequired to choose from several well-defined options; others involve being placed in a novel situarion in which the test taker's response is not completely structured (e.g., responding to visual stimuli, telling stories, discussing pictures, or responding to other projective stimuli). The responses are scored and combined into either logically or statistically derived dimensions established by previous research.

Personality tests may be designed to focus on the assessment of normal or abnormal attitudes, feelings, traits, and related characteristics. Tests intended to measure normal personality characteristics are constructed to yield scores reflecting the degree to which a person manifests personality dimensions empirically identified and hypothesized to be present in the behavior of mosr individuals. A person's configuration of scores on these dimensions is then used to infer how the person behaves presently and how she/he may behave in new situations. Test scores outside of the expected range may be considered extreme expressions of normal traits or indicative of psychopathology. Such scores also may reflect normal functioning of the person within a culture different from that of the normative population sample.

Other personality tests arc designed specifically to measure constructs underlying abnormal functioning and psychopathology. Developers of some of these tests use previously diagnosed individuals to construct their scales and base their inferences on the association between the test's scale scores, within a given range, and the behavioral correlates of persons who scored within that range. If inferences made from scores go beyond the theory that guided the test's construction, then the inferences must be validated by collecting and analyzing additional relevant data.

Vocational Testing

Vocational testing generally includes the measurement of interests, work needs, and values, as well as consideration and assessment of related elements of career development, maturity, and indecision. The results from inventories that assess these constructs often are used for enhancing personal growth and understanding, career counseling, outplacement counseling, and vocational decision making. These interventions frequenrly take place in the context of educational settings.

However, interest inventories and measures of work values also may be used in workplace settings as part of training and development programs, for career planning, or for selection, placement, and advancement decisions.

Interest Inventories. The measurement of interests is designed to identify a person's preferences for various activities. Self-report interest inventories are widely used to assess personal preferences including likes and dislikes for various work and leisure activities, school subjects, occupations, or types of people. The resulting scores may provide insight into types and patterns of differential interests in educational curricula (e.g., college majors), in different fields of work (e.g., specific occupations), or in more general or basic areas of interests related to specific activities (e.g., sales, office practices, or mechanical activities).

Work Values Inventories. The measurement of work values identifies a person's preferences for the various reinforcements one may obtain from work activities. Sometimes these values are identified as needs that persons seek to satisfy. Work values or needs may be categorized as intrinsic and important for the pleasure gained from the activity (e.g., independence, ability utilization, achievement) or as extrinsic and important for the rewards they bring (e.g., coworkers, supervisory relations, working conditions). The format of work values tests usually involves a self-rating of the importance of the value associated with qualities described by the items.

Measures of Career Development, Maturity, and Indecision. Additional areas of vocational assessment include measures of career development and maturity and measures of career indecision. Inventories that measure career development and maturity typically elicit client self-descriptions in response to items that inquire about the individual's knowledge of the world of work; self-appraisal of one's decision-making skills; attitudes toward careers and career choices; and the degree to which the individual already has engaged in career planning. Measures of career indecision usually are constructed and standardized to assess both the level of career indecision of a client as well as the reasons for, or antecedents of, indecision. Such career development, maturity, and indecision findings may be used with individuals and groups to guide the design and delivery of career services and to evaluate the effectiveness of career interventions.

Purposes of Psychological Testing

For purposes of this chapter, psychological test uses have been divided into four categories: testing for diagnosis; intervention planning and outcome evaluation; legal and governmental decisions; and personal awareness, growth and action. However, these categories are not always mutually exclusive.

Testing for Diagnosis

Diagnosis refers to a process that includes the collection and integration of test results with prior and current information about a person together with relevant contextual conditions to identify characteristics of healthy psychological functioning as well as psychological disorders. Disorders may manifest themselves in information obtained during the testing of an individual's cognitive, emotional, social, personality, neuropsychological, physical, perceptual, and motor attributes.

Psychodiagnosis. Psychological tests are helpful to professionals involved in the psychological diagnosis of an individual. Testing may be performed to confirm a hypothesized diagnosis or to rule out alternative diagnoses. Psychodiagnosis is complicated by the prevalence of comorbidity between diagnostic categories. For example, a client diagnosed as suffering from schizophrenia simultaneously may be diagnosed as suffering from depression. Or, a child diagnosed as having a learning disability also may be diagnosed as suffering from an attention deficit disorder. The goal of psychodiagnosis is to assist each client in receiving the appropriate interventions for the psychological or behavioral

dysfunctions that the client, or a third party, views as impairing the client's expected functioning and/or enjoyment of life. In developing treatment plans, professionals often use noncategorical diagnostic descriptions of client functioning along treatment-relevant dimensions (e.g., degree of anxiety, amount of suspiciousness, openness to interpretations, amount of insight into behaviors, and level of intellectual functioning).

The first step in evaluating a test's suitability to yield scores or information indicative of a particular diagnostic syndrome is to compare the construct that the test is intended to measure with the symptomatology described in the diagnostic criteria. This step is important because different diagnostic systems may use the same diagnostic term to describe different symptoms; even within one diagnostic system the symptoms described by the same term may differ between editions of the manual identifying the diagnostic criteria. Similarly, a test that uses a diagnostic term in its title may differ significantly from another test using a similar title or from a subscale with the same term. For example, some diagnostic systems may define depression by behavioral symptomatology (e.g., psychomotor retardation, disturbance in appetite or sleep) or by affective symptomatology (e.g., dysphoric feeling, emotional flatness) or by cognitive symptomatology (e.g., thoughts of hopelessness, morbidity) or some other symptomatology. Further, rarely are the symptoms of diagnostic categories mutually exclusive. Hence, it can be expected that a given symptom may be shared by several diagnostic categories. More knowledgeable and precisely drawn inferences relating to a diagnosis may be obtained from test scores if appropriate weight is given to the symptoms included in the diagnostic category and to the suitability of each test to assess the symptoms.

Different methods may be used to assess particular diagnostic categories. Some methods rely primarily on structured interviews using a "yes" or "no" format in which the professional is interested in the presence or absence of diagnosis-specific symptomatology. Other methods often rely principally on tests of personality or cognitive functioning and use configurations of obtained scores. These configurations of scores indicate the degree to which a client's responses are similar to those of individuals who have been determined by prior research to belong to a specific diagnostic group.

Diagnoses made with the help of test scores typically are based on empirically demonstrated relationships between the test score and the diagnostic category. Validity studies that demonstrate relationships between test scores and diagnostic categories currently are available for some diagnostic categories. Sometimes tests that do not have supporting validity studies also may be useful to the professional in arriving at a diagnosis. This also may occur, for example, when the symptoms assessed by a test are a subset of the criteria that comprise a particular diagnostic category. While it often is not feasible for individual professionals to personally conduct research into relationships between obtained scores and inferences, their familiarity with the body of the research literature that examines these relationships is important.

The professional often can enhance the diagnostic inferences derived from test scores by integrating the test results with inferences made from other sources of information regarding the client's functioning such as self-reported history or information provided by significant others or systematic observations in the natural environment or in the testing setting. In arriving at a diagnosis, a professional also looks for information that does not corroborate the diagnosis, and in those instances, places appropriate limits on the degree of confidence placed in the diagnosis. When relevant to the referral issue, the professional acknowledges alternative diagnoses that may require consideration. Particular attention is paid to all relevant available data before concluding that a client falls into a diagnostic category. Cultural sensitivity is paramount to avoid misdiagnosing and over

pathologizing culturally appropriate behavior, affect or cognition. Tests also are used to assess the appropriateness of continuing the initial diagnostic characterization, especially after a course of treatment or if the client's psychological functioning has changed over time.

Neuropsychodiagnosis. Neuropsychological testing analyzes the current psychological and behavioral status, including manifestations of neurological, neuropathological, and neurochemical changes that may arise during development or from brain injury or illness. The purposes of neuropsychological testing typically include, but are not limited to, the following: differential diagnoses between psychogenic and neurogenic sources of cognitive, perceptual, and personality dysfunction; differential diagnoses between two or more suspected etiologies of cerebral dysfunction; evaluation of impaired functioning secondary to a cerebral, cortical, or subcortical event; establishment of neuropsychological baseline measurements for monitoring progressive cerebral disease or recovery effects; comparison of pre- and post-pharmacologic, surgical, behavioral, or psychological interventions; identification of patterns of highet cortical function and dysfunction for the formulation of rehabilitation strategies and for the design of remedial procedures; and characterizing brain-behavior functions to assist the trier of fact in criminal and civil legal actions.

Testing for Intervention Planning and Outcome Evaluation

Professionals often rely on test results for assistance in planning, executing, and evaluating interventions. Therefore, their awareness of validity information that supports or does not support the relationship between test results, prescribed interventions, and desired outcome is important. Interventions may be intended to prevent the onset of one or more symptoms, to stabilize or overcome them, to ameliorate their effects, to minimize their impact, and to provide for a person's basic physical, psychological, and social needs. Intervention planning typically occurs following an evaluation of the nature and severity of a disorder and a review of personal and contextual conditions that may impact its resolution. Subsequent evaluations may occur in an effort to diagnose further the nature and severity of the disorder, to review the effects of interventions, to revise them as needed, and to meet ethical and legal standards.

Testing for Judicial and Governmental Decisions

Clients may voluntarily seek psychological testing as part of psychological assessments to assist in matters before a court or other governmental agencies. Conversely, courts or other governmental agencies sometimes require a client to submit involuntarily to a psychological or neuropsychological assessment that may involve a wide range of psychological tests. The goal of these psychological assessments is to provide important information to a third party, client's attorney, opposing attorney, judge, or administrative board about the psychological functioning of the client that has bearing on the legal issues in question. At the outset of evaluations for judicial and government decisions, it is imperative to clarify the purpose of the evaluation, who will have access to the test results and the reports, and any rights that the client may have to refuse to participate in court-ordered evaluations.

The goals of psychological testing in judicial and governmental settings are informed and constrained by the legal issues to be addressed, and a detailed understanding of their salient aspects is essential. Legal issues may arise as part of a civil proceeding (e.g., involuntary commitment, testamentary capacity, competence to stand trial, parole, child custody, personal injury, discrimination issues), a criminal proceeding (e.g., competence to stand trial, not guilty by reason of insanity, mirigating circumstances in sentencing), determination of reasonable accommodations for employees with disabilities, or an administrative proceeding or decision (e.g., license revocation, parole, worker's compensation). Each of these legal issues is

defined in law applicable to a particular legislative jurisdiction. The definition of each legal issue may be jurisdiction specific. For example, the criteria by which a person can be involuntarily committed often differ between legislative jurisdictions. Furthermore, tests initially administered for one purpose also may be used for another purpose (e.g., initially used for a civil case but later used in administrative or criminal proceedings).

Legislatures, courts, and other adminstrative bodies often define legal issues in commonly used language, not in diagnostic or other technical psychological terms. The professional is responsible for explaining the diagnostic frame of reference, including test scores and inferences made from them, in terms of the legal criteria by which the jury, judge, or administrative board will decide the legal issue. For example, a diagnosis of schizophrenia or neuropsychological impairment, which does not also include a reference to the legal criteria, neither precludes an examinee from obtaining sole custody of children in a child custody dispute nor does it necessarily acquit a person of criminal responsibility.

In instances involving legal or quasi-legal issues, it is important to assess the examinee's test-taking orientation including response bias to ensure that the legal proceedings have not affected the responses given. For example, a person seeking to obtain the greatest possible monetary award for a personal injury may be motivated to exaggerate cognitive and emotional symptoms, while persons attempting to forestall the loss of a professional license may attempt to portray themselves in the best possible light by minimizing symptoms or deficits. In forming an assessment opinion, it is necessary to interpret the test scores with informed knowledge relating to the available validity and reliability evidence. When forming such opinions, it also is necessary to integrate a client's test scores with all other sources of information that bear on current status including psychological, medical, educational, occupational, legal, and other relevant collateral records.

Some tests are intended to provide information about a client's functioning that helps clarify a given legal issue (e.g., parental functioning in a child custody case or ability to understand charges against a defendant in competency to stand trial matters). The manuals of some tests also provide demographic and actuarial data for normative groups that are representative of persons involved in the legal system. However, many tests measure constructs that are generally relevant to the legal issues even though norms specific to the judicial or governmental context may not be available. Professionals are expected to make every effort to be aware of evidence of validity and reliability that supports or does not support their inferences and to place appropriate limits on the opinions rendered. Test users who practice in judicial and government settings are expected to be aware of conflicts of interest that may lead to bias in the interpretation of test results.

Protecting the confidentiality of a client's test results and of the test instrument itself poses particular challenges for professionals involved with attorneys, judges, jurors, and other legal and quasi-legal decision makers. The test taker does have a right to expect that test results will be communicated only to persons who are legally authorized to receive them and that other information from the testing session that is not relevant to the evaluation will not be reported. It is important for the professional to be apprised of possible threats to confidentiality and test security (e.g., releasing the test questions, the examinee's responses, and raw and scaled scores on tests to another qualified professional) and to seek, if necessary, appropriate legal and professional remedies.

Testing for Personal Awareness, Growth, and Action

Tests and inventories frequently are used to provide information to help individuals to understand themselves, to identify their own strengths and weaknesses, and to otherwise clarify issues important to their own decision

making and development. For example, test results from personality inventories may help clients better understand themselves and also understand their interactions with others. Results from interest inventories and tests of ability may be useful to individuals who are making educational and career decisions. Appropriate cognitive and neuropsychological tests that have been normed and standardized for children may facilitate the monitoring of development and growth during the formative years when relevant interventions may be more efficacious for preventing potentially disabling learning disabilities from being overlooked or misdiagnosed.

Test results may be used for self-exploration, self-growth, and decision making in several ways. First, the results can provide individuals with new information that allows them to compare themselves with others or to evaluate themselves by focusing on self-descriptions and characterizations. Test results also may serve to stimulate discussions between a client and professional, to facilitate client insights, to provide directions for future considerations, to help individuals identify strengths and assets, and to provide the professional with a general framework for organizing and integrating information about an individual. Testing for personal growth may take place in training and development programs, within an educational curriculum, during psychotherapy, in rehabilitation programs as part of an educational or career planning process, or in other situations.

Summary

The application of psychological tests continues to expand in scope and depth on a course that is characterized by an increasingly diverse set of purposes, procedures, and assessment needs and challenges. Therefore, the responsible use of tests in practice requires a commitment by the professional to develop and maintain the necessary knowledge and competence to select, administer, and interpret tests and inventories as crucial elements of the psychological testing and assessment process. The standards in this chapter provide a framework for guiding the professional toward achieving relevance and effectiveness in the use of psychological tests within the boundaries or limits defined by the professional's educational, experiential and ethical foundations. Earlier chapters and standards that are relevant to psychological testing and assessment describe general aspects of test quality (chapters 1-6, chapter 11), test fairness (chapters 7-10), and test use (chapter 11). Chapter 13 discusses educational applications; chapter 14 discusses test use in the workplace, including credentialing, and the importance of collecting data that provide evidence of a test's accuracy for predicting job performance; and chapter 15 discusses test use in program evaluation and public policy.

Standard 12.1

Those who use psychological tests should confine their testing and related assessment activities to their areas of competence, as demonstrated through education, supervised training, experience, and appropriate credentialing.

Comment: The responsible use and interpretation of test scores requite appropriate levels of expetience and sound professional judgment. Competency also requites sufficient familiarity with the population from which the test taker comes to allow appropriate interaction, test selection, test administration, and test interpretation. For example, when personality tests and neuropsychological tests are administered as patt of a psychological assessment of an individual, the test scores must be understood in the context of the individual's physical and emotional state, as well as the individual's cultural, educational, occupational, and medical background, and must take into account other evidence relevant to the tests used. Test interpretation in this context requires professionally tesponsible judgment that is exercised within the boundaries of knowledge and skill afforded by the professional's education, training, and supervised experience.

Standard 12.2

Those who select tests and interpret test results should refrain from introducing biases that accommodate individuals or groups with a vested interest in decisions affected by the test interpretation.

Comment: Individuals or groups with a vested interest in the significance or meaning of the findings from psychological testing include many school personnel, attorneys, referring health professionals, employers, professional associates, and managed care organizations. In some settings a professional may have a professional relationship with multiple clients (e.g., with both the test taker and the organization requesting assessment). A professional engaged in a professional relationship with multiple clients takes care to ensure that the multiple relationships do not become a conflict of interest that would occur when the professional's judgment toward one client is unduly influenced by his or her relationship with the other client. Test selections and interpretations that favor a special external expectation or perspective by deviating from established principles of sound test interpretation are unprofessional and unethical.

Standard 12.3

Tests selected for use in individual testing should be suitable for the characteristics and background of the test taker.

Comment: Considerations for test selection should include culture, language and/or physical requirements of the test and the availability of norms and evidence of validity for a population representative of the test taker. If no normative or validity studies are available for the population at issue, test interpretations should be qualified and presented as hypotheses rather than conclusions.

Standard 12.4

If a publisher suggests that tests are to be used in combination with one another, the professional should review the evidence on which the procedures for combining tests is based and determine the rationale for the specific combination of tests and the justification of the interpretation based on the combined scores.

Comment: For example, if measures of developed abilities (e.g., achievement or specific or general abilities) or personality are packaged with interest measures to suggest a requisite combination of scores, or a neuropsychological battery is being applied, then supporting validity data for such combinations of scores should be available.

Standard 12.5

The selection of a combination of tests to address a complex diagnosis should be appropriate for the purposes of the assessment as determined by available evidence of validity. The professional's educational training and supervised experience also should be commensurate with the test user qualifications required to administer and interpret the selected tests.

Comment: For example, in a neutopsychological assessment for evidence of an injury to a particular area of the brain, it is necessary to select a combination of tests of known diagnostic sensitivity and specificity to impairments arising from trauma to various regions of the cerebral hemispheres.

Standard 12.6

When differential diagnosis is needed, the professional should choose, if possible, a test for which there is evidence of the test's ability to distinguish between the two or more diagnostic groups of concern rather than merely to distinguish abnormal cases from the general population.

Comment: Ptofessionals will find it particularly helpful if evidence of validity is in a form that enables them to determine how much confidence can be placed in inferences regarding an individual. Differences between group means and theit statistical significance provide inadequate information regarding validity for individual diagnostic purposes. Additional information might consist of confidence intervals, effect sizes, or a table showing the degree of overlap of predictor distributions among different criterion groups.

Standard 12.7

When the validity of a diagnosis is appraised by evaluating the level of agreement between test-based inferences and the diagnosis, the diagnostic terms or categories employed should be carefully defined or identified.

Standard 12.8

Professionals should ensure that persons under their supervision, who administer and score tests, are adequately trained in the settings in which the testing occurs and with the populations served.

Standard 12.9

Professionals responsible for supervising group testing programs should ensure that the individuals who interpret the test scores are properly instructed in the appropriate methods for interpreting them.

Comment: If, for example, intetest inventories are given to college students for use in academic advising, the ptofessional who supetvises the academic advisors is responsible for ensuring that the advisors know how to ptovide an examinee an apptopriate interptetation of the test results.

Standard 12.10

Prior to testing, professionals and test administrators should provide the test taker with appropriate introductory information in language understandable to the test taker. The test taker who inquires also should be advised of opportunities and circumstances, if any, for retesting.

Comment: The client should understand testing time limits, who will have access to the test results, if and when test results will be shared with the test taker, and if and when decisions based on the test results will be shared with the test taker.

Standard 12.11

Professionals and others who have access to test materials and test results should ensure

the confidentiality of the test results and testing materials consistent with legal and professional ethics requirements.

Comment: Professionals should be knowledgeable and conform to record-keeping and confidentiality guidelines required by the state or province in which they practice and the professional organizations to which they belong. Confidentiality has different meanings for the test developer, the test user, the test taker, and third parties (e.g., school, court, employer). To the extent possible, the professional who uses tests is responsible for managing the confidentiality of test information across all parties. It is important for the professional to be aware of possible threats to confidentiality and the legal and professional remedies available. Professionals also are responsible for maintaining the security of testing materials and for protecting the copyrights of all tests to the extent permitted by law.

Standard 12.12

The professional examines available norms and follows administration instructions, including calibration of technical equipment, verification of scoring accuracy and replicability, and provision of settings for testing that facilitate optimal performance of test takers. However, in those instances where realistic rather than optimal test settings will best satisfy the assessment purpose, the professional should report the reason for using such a setting and, when possible, also conduct the testing under optimal conditions to provide a comparison.

Comment: Because the normative data against which a client's performance will be evaluated were collected under the reported standard procedures, the professional needs to be aware of and take into account the effect that nonstandard procedures may have on the client's obtained score. When the professional uses tests that employ an unstructured response format, such as some projective techniques and informal behavioral ratings, the professional should follow objective scoring criteria, where available and appropriate, that are clear and minimize the need for the scorer to rely only on individual judgment. The testing may be conducted in a realistic, less than optimal, setting to determine how a client with an attentional disorder, for example, performs in a noisy or distracting environment rather than in an optimal environment that typically protects the test taker from such external threats to performance efficiency.

Standard 12.13

Those who select tests and draw inferences from test scores should be familiar with the relevant evidence of validity and reliability for tests and inventories used and should be prepared to articulate a logical analysis that supports all facets of the assessment and the inferences made from the assessment.

Comment: A presentation and analysis of validity and reliability evidence generally is not needed in a written report, because it is too cumbersome and of little interest to most report readers. However, in situations in which the selection of tests may be problematic (e.g., verbal subtests with deaf clients), a brief description of the rationale for using or not using particular measures is advisable.

When potential infetences derived from psychological test data are not supported by evidence of validity yet may hold promise for future validation, they may be described by the test developer and professional as hypotheses for further validation in test interpretation. Such interpretive remarks should be qualified to communicate to the source of the referral that such inferences do not as yet have adequately demonstrated evidence of validity and should not be the basis for a diagnostic decision or prognostic formulation.

Standard 12.14

The interpretation of test results in the assessment process should be informed when possible by an analysis of stylistic and other qualitative features of test-taking behavior that are inferred from observations during interviews and testing and from historical information.

Comment: Such features of test-taking behavior include manifestations of fatigue, momentary fluctuations in emotional state, rapport with the examiner, test taker's level of motivation, withholding or distortion of response as seen in instances of deception and malingering or in instances of pseudoneurological conditions, and unusual response or general adaptation to the testing environment.

Standard 12.15

Those who use computer-generated interpretations of test data should evaluate the quality of the interpretations and, when possible, the relevance and appropriateness of the norms upon which the interpretations are based.

Comment: Efforts to reduce a complex set of data into computer-generated interpretations of a given construct may yield grossly misleading or simplified analyses of meanings of test scores, that in turn may lead to faulty diagnostic and prognostic decisions as well as mislead the trier of fact in judicial and government settings.

Standard 12.16

Test interpretations should not imply that empirical evidence exists for a relationship among particular test results, prescribed interventions, and desired outcomes, unless empirical evidence is available for populations similar to those representative of the examinee.

Standard 12.17

Criterion-related evidence of validity should be available when recommendations or decisions are presented by the professional as having an actuarial basis.

Standard 12.18

The interpretation of test or test battery results generally should be based upon multiple sources of convergent test and collateral data and an understanding of the normative, empirical, and theoretical foundations as well as the limitations of such tests.

Comment: A given pattern of test performances represents a cross-sectional view of the individual being assessed within a particular context (i.e., medical, psychosocial, educational, vocational, cultural, ethnic, gender, familial, genetic, and behavioral). The interpretation of findings derived from a complex battery of tests in such contexts requires appropriate education, supervised experience, and an appreciation of procedural, theoretical, and empirical limitations of the tests.

Standard 12.19

The interpretation of test scores or patterns of test battery results should take cognizance of the many factors that may influence a particular testing outcome. Where appropriate, a description and analysis of the alternative hypotheses or explanations that may have contributed to the pattern of results should be included in the report.

Comment: Many factors (e.g., unusual testing conditions, motivation, educational level, employment status, lateral sensorimotor usage preferences, health, or disability status) may influence individual testing results. When such factors are known to introduce construct-irrelevant variance in component test scores, those factors should be considered during test score interpretations.

Standard 12.20

Except for some judicial or governmental referrals, or in some employment testing situations when the client is the employer, professionals should share test results and interpretations with the test taker. Such information should be expressed in language that the test taker, or when appropriate the test taker's legal representative, can understand.

Comment: For example, in rehabilitation settings, where clients typically are required to participate actively in intervention programs, sharing of such information, expressed in terms that can be understood readily by the client and family members, may facilitate the effectiveness of intervention.

13. EDUCATIONAL TESTING AND ASSESSMENT

Background

This chaprer concerns testing in formal educational settings from kindergatten thtough postgraduate rraining. Results of rests administered to studenrs are used to make judgments, for example, about the status, progress, or accomplishments of individuals or groups. Tests that provide information about individual performance are used to (a) evaluate a student's overall achievement and growth in a content domain, (b) diagnose student strengths and weaknesses in and across content domains, (c) plan educational interventions and to design individualized instructional plans, (d) place students in appropriate educational programs, (e) select applicants into programs with limited enrollment, and (f) cerrify individual achievemenr or qualifications. Tests that provide information about the status, ptogress, ot accomplishments of groups such as schools, school districts, or states are used (a) to judge and monitor the quality of educational programs for all or for particular subsers of individuals, and (b) to infer the success of policies and interventions that have been selecred for evaluation. These testing purposes are typically mandated by institurions such as schools and colleges and by governing bodies of public and privately administered educational programs.

In this chapter, three broad areas of educational resring are considered rhat encompass one or more of the above purposes: (a) routine school, district, state, or other system-wide resting programs; (b) tesring for selecrion in higher education; and (c) individualized and special needs resting. While the second and rhird areas refer to relatively specific purposes of testing, system-wide testing programs can encompass mulriple individual and group purposes. For each of these areas, the chapter elaborates on the specific purposes and domains encompassed and raises specific issues of technical quality and fairness in testing that may not be addressed or emphasized in the preceding chapters. This chapter does not explicitly address issues related to tests consrrucred and administered by teachers for their own classroom use or provided by publishets of instructional materials. While many aspects of the *Standards*, particularly those in the areas of validity, reliability, test development, and fairness, are relevant to such tests, this documenr is not intended for tests used by teachers for their own classroom purposes.

Issues in Educational Testing

This chapter first considers some cross-cutting issues: the distinctions among types of tests, the design or use of tests to serve multiple purposes including the measurement of change, and the "stakes" associated with different purposes for testing in education.

Distinctions Among Types of Tests and Assessments

Tests used in educational settings range from tests consisring of traditional item formats such as multiple-choice items to performance assessments including scorable portfolios. Every test, regardless of its fotmat, measures test-taker performance in a specified domain. Performance assessments, however, attempt to emulate the conrext or conditions in which the intended knowledge or skills are actually applied. As discussed in chapter 3, they are diverse in nature and can be producr-based as well as behavior-based. The execution of the tasks posed in these tests often involves relatively extended rime periods, ranging from a few minutes ro a class period or more to several hours or days. Examples of such performances might include solving problems using manipulablc marerials, making complex inferences after collecring information, or explaining orally or in writing

the rationale for a particular course of government action under given economic conditions. The performance task may be undertaken by a single individual or a team of students. Performance assessments may require increased testing time to provide sufficient domain sampling for reasonable estimates of individual attainment and for making generalizations to the broader domain. Extended time periods, collaboration, and the use of ancillary materials pose great challenges to the standardization of administration and scoring of some performance assessments. This is particularly true when test takers define their own tasks or when they select their own work products for evaluation. When this is the case, test takers need to be aware of the basis for scoring as well as the nature of the criteria that will be applied. Further, performance assessments often require complex procedures and training to increase the accuracy of judgments made by those evaluating student performance (see chapter 3).

An individual portfolio may be used as another type of performance assessment. Scorable portfolios are systematic collections of educational products typically collected over time and possibly amended over time. The particular purpose of the portfolio determines whether it will include representative products, the best work of the student, or indicators of progress. The purpose also dictates who will be responsible for compiling the contents of the portfolio—the examiner, the student, or both parties working together. The more standardized the contents and procedures of administration, the easier it is to establish comparability of portfolio-based scores. Establishing comparability requires portfolios to be constructed according to test specifications and standards, and the development of objective procedures to judge their quality. The test specifications for portfolios may indicate that students are to make certain decisions about the nature of the work to be included. For example, in constructing an art portfolio, students may select the media that best represent their work. Establishing comparability also requires specifications regarding the kinds of assistance the student may have received during portfolio preparation. It is particularly difficult to compare the performance of students whose portfolios may vary in content. All performance assessments, including scorable portfolios, are judged by the same standards of technical quality as traditional tests of achievement.

Electronic media are often used both to present testing material and to record and score test takers' responses. These tests may be administered in schools, in special laboratory settings, or in external testing centers. Examples include simple enhancements of text by audio-taped instructions to facilitate student understanding, computer-based tests traditionally given in papet-and-pencil format, computer-adaptive tests, and newer, interactive multimedia testing situations where attributes of performance assessments are supported by computer. Some computer-based tests also may have the capacity to capture aspects of students' processes as they solve test items. They may, for example, monitor time spent on items, solutions tried and tejected, or editing sequences for texts. Electronic media also make it possible to provide test administration conditions designed to assist students with particular needs, such as those with different language backgrounds, attention problems, or physical disabilities. Computers can also help identify the contributions of individuals to a group task completed by a team or in geographically remote locations on a network.

Computer-based tests are evaluated by the same technical quality standards as other tests administered through more traditional means. It is especially important that test takets be familiarized with the media of the test so that any unfamiliarity with computers or strategies does not lead to inferences based on construct-irrelevant variance. Furthermore, it is important to describe scoring algorithms, expert models upon which they may be based, and technical data supporting their use in any documentation accompanying the testing system. It is important, however, to assute that the docu-

mentation does not jeopardize the security of the items that could adversely affect the validity of score interpretations. Some computer-based tests may also generate recommendations for instructional practices based on test results. Describing the basis for these recommendations assists the user in evaluating their applicability in a given situation.

Multiple Purposes and Measuring Change

Many tests are designed or used to serve multiple purposes in education. For example, a rest may be used to monitor individual student achievement as well as to evaluate the quality of educational programs at the school or district level. As another example, a test may be used to evaluate an individual's performance relative to the perfotmance of one or more reference populations as well as to evaluate the level of the individual's competence in some defined domain (see chaprers 3 and 4). The evidence needed for the technical quality of one purpose, however, will differ from the evidence needed fot another purpose. Consequently, it is important to evaluate the evidence of technical quality fot each purpose of testing.

Test results may be used to infer the growth ot progress as well as the status of individuals or groups of students, such as when tests are expected to reveal the effects of instruction, of changes in educational policy, or of other interventions. In such cases, the test's ability to detect change is essential. If differences in scores are reported, the technical quality of the differences needs attention. More generally, whenever inferences about growth or progress are made, it is important to evaluate the validity of those inferences.

Stakes of Testing

The importance of the results of testing programs for individuals, institutions, or groups is often referted to as the *stakes* of the testing program. At the individual level, when significant educational paths or choices of an individual are directly affected by test perfotmance, such as whether a student is promoted or retained at a grade level, graduated, or admitted or placed into a desired program, the test use is said to have high stakes. A low-srakes test, on the other hand, is one administered for informational purposes or for highly tentative judgments such as when test results provide feedback to students, teachets, and patents on student progress during an academic period. Testing programs for institutions can have high stakes when aggregate performance of a sample or of the entire population of test takers is used to infer the quality of service provided, and decisions are made about institutional status, rewards, or sanctions based on test results. For example, the quality of reading curriculum and instruction may be judged on the basis of test results because test scores can indicate the rate of student progress or the levels of attainment reached by groups of students. Even when test results are reported in the aggregare and intended for a low-stakes purpose such as monitoring the educational system, the public release of data can taise the stakes for particular schools or districts. Judgments about program quality, personnel, and educational ptograms might be made and policy decisions might be affected, even though the tests wete not intended or designed for those purposes.

The higher the stakes associated with a given test use, the more important it is that test-based inferences are supported with strong evidence of technical quality. In particular, when the srakes for an individual are high, and important decisions depend substantially on tesr performance, the test needs to exhibit higher standards of technical quality for its avowed purposes than might be expected of tests used for lower-stakes purposes (see chapters 1, 2, and 7 for a mote thorough discussion on validity, reliability, and bias in testing, respectively). Although it is never possible ro achieve perfect accutacy in describing an individual's performance, efforts need to be made to minimize errors in estimating individual scores or in classifying individuals in pass/fail or admit/reject categories.

Further, enhancing validity for high-stakes purposes, whether individual or institutional, typically entails collecting sound collateral information both to assist in understanding the factors that contributed to test results and to provide corroborating evidence that supports inferences based on test results. These issues will be addressed more fully as they relate to the three areas of testing described below.

School, District, State, or Other System-Wide Testing Programs

As indicated previously, system-wide testing programs can span multiple purposes. At the individual level, tests are used for low-stakes purposes, such as monitoring and providing feedback on student progress, and for more high-stakes purposes, such as certifying students' acquisition of particular knowledge and skills for promotion, placement into special instructional programs, or graduation. At the school, district, state, or other aggregate level, a common purpose of tests is to evaluate the progress made by groups of students or to monitor the long-term effectiveness of the overall educational system. Educational testing programs may also permit comparisons among the performance of various groups of students in different programs or in diverse settings for the purpose of making an evaluation of those learning environments. Chapter 15 provides a more thorough discussion on program evaluation.

In these contexts, educational tests are designed to measure certain aspects of students' knowledge and skills as reflected in curriculum goals and standards. There may be considerable variation in the breadth and depth of the knowledge and skills that are measured by such tests. Some educational tests focus on the test takers' general ability or knowledge in a particular content area, such as their understanding of mathematics or science. Other tests focus on test takers' specific knowledge of a topic in detail, such as trigonometry. Still others emphasize specific skills or procedures, such as the ability to write persuasively or to design, conduct, and interpret the results of a scientific experiment. Tests may address other cognitive aspects of test takers' development, such as their ability to work with others to solve problems or their self-reported habits and attitudes, as well as noncognitive aspects, such as students' ability to perform particular physical tasks. In most cases, valid interpretation of the results requires that evidence of the fit between the test domain and the relevant curriculum goals or standards be ascertained.

Testing programs may involve the use of tests designed to represent a set of general educational standards as determined for instance by the state, district, or relevant educational professional organization. Such tests are conceptually similar to criterion-referenced tests, in that a set of content standards is developed that is intended to provide broad specifications for student performance by delimiting the content and general skills to be measured. Subsequently, descriptive or empirical targets or levels of achievement are developed and referred to as performance standards. These performance standards are intended to define further the knowledge and skills required of students for each of the different categories of proficiency.

This type of testing may involve the development of a new test to assess the relevant content and skills or the selection of an existing test that can be referenced to the standards. Whether a test is designed or selected, valid interpretation of the results in light of the standards entails assessment of the degree of fit between the test domain and contents and the descriptive statements of standards or goals. This involves a process of mapping or referencing the content and skills of the test to those of the standards to be sure that gaps or imbalances do not occur. The curriculum goals or standards may be sufficiently broad to encompass many different ways for students to demonstrate their status, accomplishments, or

progress. Moreover, some goals or standards may not lend themselves to conventional test formats. These are cases in which the test may result in construct underrepresentation that refers to the extent to which a test fails to capture important aspects of what it is intended to measure. Chapter 1 provides a more thorough discussion of construct underrepresentation. In these cases, interpretation of test results in light of goals or standards is enhanced by an understanding of what is not covered as well as what is covered by the test. Sometimes, additional commercial or locally developed tests are administered within a particular jurisdiction, and attempts are made to link these existing tests to the proficiency levels reported for the new test or to provide other evidence of comparability. It is important to provide logical and empirical validity evidence of any reported links. For example, evidence can be collected to determine the extent to which the existing test can provide information about the proficiency of individual students and groups of students in the particular content areas and skills addressed by the standards. The validity of such links is problematic to the extent that the tests measure different content (see chapter 4 for a discussion on issues in equating and linking tests).

When inferences are to be drawn about the performance of groups of students, practical considerations and the format of the test (e.g., performance assessment) often dictate that different subgroups of students within each unit respond to different sets of tasks or items, a procedure referred to as matrix sampling. This matrix sampling approach allows for a test to better represent the breadth of the target domain without increasing the testing time for each test taker. Group-level results are most useful when testing programs and student populations remain sufficiently stable to provide information about trends over time. When a testing program is designed for group-level reporting and employs matrix sampling, reporting individual scores generally is not appropriate.

When interpreting and using scores about individuals or groups of students, consideration of relevant collateral information can enhance the validity of the interpretation, by providing corroborating evidence or evidence that helps explain student performance. Test results can be influenced by multiple factors, including institutional and individual factors such as the quality of education provided, students' exposure to education (e.g., through regular school attendance), and students' motivation to perform well on the test.

As the stakes of testing increase for individual students, the importance of considering additional evidence to document the validity of score interpretations and the fairness in testing increases accordingly. The validity of individual interpretations can be enhanced by taking into account other relevant information about individual students before making important decisions. It is important to consider the soundness and relevance of any collateral information or evidence used in conjunction with test scores for making educational decisions. Further, fairness in testing can be enhanced through careful consideration of conditions that affect students' opportunities to demonstrate their capabilities. For example, when tests are used for promotion and graduation, the fairness of individual interpretations can be enhanced by (a) providing students with multiple opportunities to demonstrate their capabilities through repeated testing with alternate forms or through other construct-equivalent means, (b) ensuring students have had adequate notice of skills and content to be tested along with other appropriate test preparation material, (c) providing students with curriculum and instruction that affords them the opportunity to learn the content and skills that are tested, and (d) providing students with equal access to any specific preparation for test taking (e.g., test-taking strategies). Chapter 7 provides a more thorough discussion on fairness in testing.

Collateral information can also enhance interpretation and decisions at the institutional

level. For instance, changes in test scores from year to year may not only reflect changes in the capabilities of students but also changes in the student population (e.g., successive cohorts of students). Differences in scores across ethnic groups may be confounded with differences in socioeconomic status of the communities in which they live and, hence, the educational resources to which students have access. Differences in scores from school to school may similarly reflect differences in resources and activities such as the qualification of teachers or the number of advanced course offerings. While local empirical evidence of the influence of these factors may not be readily available, consideration of evidence from similar contexts available in published literature can enhance the quality of the interpretation and use of current results.

Because public participation is an integral part of educational governance, policymakers, professional educators, and members of the public are concerned with the nature of educational tests, the domains that the tests are intended to measure, the choices in test design, adoption, and implementation, and the issues associated with valid interpretation and uses of test results. It is important that test results be reported in a way that all stakeholders can understand, that enables sound interpretations, and that decreases the chance of misinterpretations and inappropriate decisions.

Large-scale testing is increasingly viewed as a tool of educational policy. From this perspective, tests used for program evaluation, such as some state tests that are aligned to the state's own curriculum standards, are not used solely as measures of school outcomes (see chapter 15 for a more thorough discussion on the use of tests for program evaluation). They are also viewed as a means to influence curriculum and instruction, to hold teachers and school administrators accountable, to increase student motivation, and to communicate performance expectations to students, to teachers, and to the public. If such goals are set forth as part of the rationale for a testing program, the validity of the testing program needs to be examined with respect to these goals. Beyond any intended policy goals, it is important to consider potential unintended effects that may result from large-scale testing programs. Concerns have been raised, for instance, about narrowing the curriculum to focus only on the objectives tested, restricting the range of instructional approaches to correspond to the testing format, increasing the number of dropouts among students who do not pass the test, and encouraging other instructional or administrative practices that may raise test scores without affecting the quality of education. It is important for those who mandate tests to consider and monitor their consequences and to identify and minimize the potential of negative consequences.

Selection in Higher Education

It is widely recognized that tests are used in the selection of applicants for admission to particular educational programs, especially admissions to colleges, universities, and professional schools. Selection criteria may vary within an institution by academic specialization. In addition to scores from selection tests, many other sources of evidence are used in making selection decisions, including past academic records, transcripts, and grade-point average or rank in class. Scores on tests used to certify students for high school graduation may be used in the college admissions process. Other measures used by some institutions are samples of previous work by students, lists of academic and service accomplishments, letters of recommendation, and student-composed statements evaluated for the appropriateness of the goals and experience of the student or for writing proficiency.

Two major points may be made about the role of tests in the admissions process. Often, scores are used in combination with other sources of information. Some of these supple-

mental sources of evidence may not be reliably assessed or may lack comparability from applicant to applicant. For this reason, it is important that studies be conducted examining the relationships among test scores, data from other sources of information, and college performance. Second, the public and policymakers are to be cautious about the widespread use of reports of college admission test scores to infer the effectiveness of middle school and high school as well as to compare schools or states. Admissions tests, whether they are intended to measure achievement or ability, are not directly linked to a particular instructional curriculum and, therefore, are not appropriate for detecting changes in middle school or high school performance. Because of differential motivational factors and other demographic variables found across and within pre-collegiate programs, self-selection precludes general comparisons of test scores across demographic groups. Therefore, self-selection also precludes comparisons of test scores among the full ranges of pre-collegiate programs.

Individualized and Special Needs Testing

Individually administered tests are used by school psychologists and other professionals in schools and other related settings to facilitate the learning and development of students who may have special educational needs (see chapter 12). Some of these services are reserved for those students who have gifted capabilities as well as for those students who may have relatively minor academic difficulties (e.g., such as those requiring remedial reading). Other services are reserved for students who display behavioral, emotional, physical, and/or more severe learning difficulties. Services may be provided to students who are in regular classroom settings as well as to students who need more specialized instruction outside of the regular classroom. The ultimate purpose of these services is to assure all students are placed into appropriate educational programs.

Individually administered tests can serve a number of purposes, including screening, diagnostic classification, intervention planning, and program evaluation. For screening purposes, tests are administered to identify students who might differ significantly from their peers and might require additional assessment. For example, screening tests may be used to identify young children who show signs of developmental disorders and to signal the need for further evaluation. For diagnostic purposes, tests may be used to clarify the types and extent of an individual's difficulties or problems in light of well-established criteria. Test results provide an important basis for determining whether the student meets eligibility requirements for special education and other related services and, if so, the specific types of services that the student needs. Test results may be used for intervention purposes in establishing behavior and learning goals and objectives for the student, planning instructional strategies that should be used, and specifying the appropriate setting in which the special services are to be delivered (e.g., regular classroom, resource room, full-time special class, etc.). Subsequent to the student's placement in special services, tests may be administered to monitor the progress of the student toward prescribed learning goals and objectives. Test results may be used also to evaluate the effectiveness of instruction to determine whether the special services need to be continued, modified, or discontinued.

Many types of tests are used in individualized and special needs testing. These include tests of cognitive abilities, academic achievement, learning processes, visual and auditory memory, speech and language, vision and hearing, and behavior and personality. These tests are used typically in conjunction with other assessment methods such as interviews, behavioral observation, and review of records. Each of these may provide useful data for mak-

ing appropriate decisions about a student. In addition, procedures that aim to link assessment closely to intervention may be used, including behavioral assessments, assessments of learning environments, curriculum-based tests, and portfolios. Regardless of the qualities being assessed and types of data collection methods employed, assessment data used in making special education decisions are evaluated in terms of validity, reliability, and relevance to the specific needs of the students. They must also be judged in terms of their usefulness for designing appropriate educational programs for students who have special needs.

The amount and complexity of the assessment data required for making various decisions about a student will vary depending on the purpose of testing, the needs of the student, and other information already available about the student (e.g., current scores on a relevant test may be on file for some students but not for others). In general, testing for screening and program evaluation purposes typically involves the use of one or two tests rather than comprehensive test batteries. For determining eligibility and designing intervention, testing and assessment is more comprehensive and may involve multiple procedures and sources. Moreover, in-depth analyses and interpretation of the data are necessary.

In special education, tests are selected, administered, and interpreted by school psychologists, school counselors, regular and special educators, speech pathologists, and physical therapists, among other professionals. The validity of inferences will be enhanced if test users possess adequate knowledge of the principles of measurement and evaluation. However, this diverse group of test users may differ in their levels of technical expertise in measurement and degree of professional training in assessment procedures. It is important that professional evaluators administer and interpret only those tests with which they have training and competence, in order to prevent misuse of tests.

State and federal law generally requires that students who are referred for possible special education services be screened for eligibility. The screening or initial assessment may in turn call for a more comprehensive evaluation. But the large numbers of students to be tested, the high cost of special education programs, and the limits of time create pressures on special education assessment practices. Assessment usually must be completed within a specific number of working days after referral, and, in most instances, the school district is responsible for funding special services recommended by the child study team. Occasionally, administrators might be inclined to use less expensive, less time-consuming, or more readily available testing procedures than a professional evaluator believes are warranted. An example would be the inappropriate use of available, but less adequately trained, staff to evaluate students. There also might be pressures to minimize or overlook problems that require expensive services. These conditions are likely to adversely affect the validity of the interpretation of test results. Adherence to professional standards governing test use in conducting special education assessments is important, in the face of pressures to use more expedient procedures. The responsible use of tests by school personnel can improve the opportunities for promoting the development and learning of all children.

Standard 13.1

When educational testing programs are mandated by school, district, state, or other authorities, the ways in which test results are intended to be used should be clearly described. It is the responsibility of those who mandate the use of tests to monitor their impact and to identify and minimize potential negative consequences. Consequences resulting from the uses of the test, both intended and unintended, should also be examined by the test user.

Comment: Mandated testing programs are often justified in terms of their potential benefits for teaching and learning. Concerns have been raised about the potential negative impact of mandated testing programs, particularly when they result directly in important decisions for individuals or institutions. Frequent concerns include narrowing the curriculum to focus only on the objectives tested, increasing the number of dropouts among students who do not pass the test, or encouraging other instructional or administrative practices simply designed to raise test scores rather than to affect the quality of education.

Standard 13.2

In educational settings, when a test is designed or used to serve multiple purposes, evidence of the test's technical quality should be provided for each purpose.

Comment: In educational testing, it has become common practice to use the same test for multiple purposes (e.g., monitoring achievement of individual students, providing information to assist in instructional planning for individuals or groups of students, evaluating schools or districts). No test will serve all purposes equally well. Choices in test development and evaluation that enhance validity for one purpose may diminish validity for other purposes. Different purposes require somewhat different kinds of technical evidence, and appropriate evidence of technical quality for each purpose should be provided by the test developer. If the test user wishes to use the test for a purpose not supported by the available evidence, it is incumbent on the user to provide the necessary additional evidence (see chapter 1).

Standard 13.3

When a test is used as an indicator of achievement in an instructional domain or with respect to specified curriculum standards, evidence of the extent to which the test samples the range of knowledge and elicits the processes reflected in the target domain should be provided. Both tested and target domains should be described in sufficient detail so their relationship can be evaluated. The analyses should make explicit those aspects of the target domain that the test represents as well as those aspects that it fails to represent.

Comment: Increasingly, tests are being developed to monitor progress of individuals and groups toward local, state, or professional curriculum standards. Rarely can a single test cover the full range of performances reflected in the curriculum standards. To assure appropriate interpretations of test scores as indicators of performance on these standards, it is essential to document and evaluate both the relevance of the test to the standards and the extent to which the test represents the standards. When existing tests are selected by a school, district, or state to represent local curricula, it is incumbent on the user to provide the necessary evidence of the congruency of the curriculum domain and the test content. Further, conducting studies of the cognitive strategies and skills employed by test takers or studies of the

relationships between test scores and other performance indicators relevant to the broader domain enables evaluation of the extent to which generalizations to the broader domain are supported. This information should be made available to all those who use the test and interpret the test scores.

Standard 13.4

Local norms should be developed when necessary to support test users' intended interpretations.

Comment: Comparison of examinees' scores to local as well as mote broadly teptesentative norm groups can be informative. Thus, sample size permitting, local norms are often useful in conjunction with published norms, especially if the local population differs markedly from the population on which published norms ate based. In some cases, local norms may be used exclusively.

Standard 13.5

When test results substantially contribute to making decisions about student promotion or graduation, there should be evidence that the test adequately covers only the specific or generalized content and skills that students have had an opportunity to learn.

Comment: Students, parents, and educational staff should be informed of the domains on which the students will be tested, the nature of the item types, and the standards for mastery. Reasonable efforts should be made to document the provision of instruction on tested content and skills, even though it may not be possible or feasible to determine the specific content of instruction for every student. Chapter 7 provides a more thorough discussion of the difficulties that arise with this conception of fairness in testing.

Standard 13.6

Students who must demonstrate mastery of certain skills or knowledge before being promoted or granted a diploma should have a reasonable number of opportunities to succeed on equivalent forms of the test or be provided with construct-equivalent testing alternatives of equal difficulty to demonstrate the skills or knowledge. In most circumstances, when students are provided with multiple opportunities to demonstrate mastery, the time interval between the opportunities should allow for students to have the opportunity to obtain the relevant instructional experiences.

Comment: The number of opportunities and time between each testing opportunity will vary with the specific circumstances of the setting. Further, some students may benefit from a diffetent testing approach to demonstrate their achievement. Care must be taken that evidence of construct equivalence of alternative approaches is provided as well as the equivalence of cut scores defining passing expectations.

Standard 13.7

In educational settings, a decision or characterization that will have major impact on a student should not be made on the basis of a single test score. Other relevant information should be taken into account if it will enhance the overall validity of the decision.

Comment: As an example, when the purpose of testing is to identify individuals with special needs, including students who would benefit from gifted and talented programs, a screening for eligibility or an initial assessment should be conducted. The screening or initial assessment may in turn call for more comprehensive evaluation. The comprehensive assessment should involve the use of

multiple measures, and data should be collected from multiple sources. Any assessment data used in making decisions are evaluated in terms of validity, reliability, and relevance to the specific needs of the students. It is important that in addition to test scores, other relevant information (e.g., school record, classroom observation, parent report) is taken into account by the professionals making the decision.

Standard 13.8

When an individual student's scores from different tests are compared, any educational decision based on this comparison should take into account the extent of overlap between the two constructs and the reliability or standard error of the difference score.

Comment: When difference scores between two tests are used to aid in making educational decisions, it is important that the two tests are standardized and, if appropriate, normed on the same population at about the same time. In addition, the reliability and standard error of the difference scores between the two tests are affected by the relationship between the constructs measured by the tests as well as the standard errors of measurement of the scores of the two tests. In the case of comparing ability with achievement test scores, the overlapping nature of the two constructs may render the reliability of the difference scores lower than test users normally would assume. If the ability and/or achievement tests involve a significant amount of measurement error, this will also reduce the confidence one may place on the difference scores. All these factors affect the reliability of difference scores between tests and should be considered by professional evaluators in using difference scores as a basis for making important decisions about a student. This standard is also relevant when comparing scores from different components of the same test such as multiple aptitude test batteries and selection tests.

Standard 13.9

When test scores are intended to be used as part of the process for making decisions for educational placement, promotion, or implementation of prescribed educational plans, empirical evidence documenting the relationship among particular test scores, the instructional programs, and desired student outcomes should be provided. When adequate empirical evidence is not available, users should be cautioned to weigh the test results accordingly in light of other relevant information about the student.

Comment: The validity of test scores for placement or promotion decisions rests, in part, upon evidence about whether students, in fact, benefit from the differential instruction. Similarly, in special education, when test scores are used in the development of specific educational objectives and instructional strategies, evidence is needed to show that the prescribed instruction enhances students' learning. When there is limited evidence about the relationship among test results, instructional plans, and student achievement outcomes, test developers and users should stress the tentative nature of the test-based recommendations and encourage teachers and other decision makers to consider the usefulness of test scores in light of other relevant information about the students.

Standard 13.10

Those responsible for educational testing programs should ensure that the individuals who administer and score the test(s) are proficient in the appropriate test administration procedures and scoring procedures and that they understand the importance of adhering to the directions provided by the test developer.

Standard 13.11

In educational settings, test users should ensure that any test preparation activities and materials provided to students will not adversely affect the validity of test score inferences.

Comment: In most educational testing contexts, the goal is to use a sample of test items to make inferences to a broader domain. When inappropriate test preparation activities occur, such as teaching items that are equivalent to those on the test, the validity of test score inferences is adversely affected. The appropriateness of test preparation activities and materials can be evaluated, for example, by determining the extent to which they reflect the specific test items and the extent to which test scores are artificially raised without actually increasing students' level of achievement.

Standard 13.12

In educational settings, those who supervise others in test selection, administration, and interpretation should have received education and training in testing necessary to ensure familiarity with the evidence for validity and reliability for tests used in the educational setting and to be prepared to articulate or to ensure that others articulate a logical explanation of the relationship among the tests used, the purposes they serve, and the interpretations of the test scores.

Standard 13.13

Those responsible for educational testing programs should ensure that the individuals who interpret the test results to make decisions within the school context are qualified to do so or are assisted by and consult with persons who are so qualified.

Comment: When testing programs are used as a strategy for guiding instruction, teachers expected to make inferences about instructional needs may need assistance in interpreting test results for this purpose. If the tests are normed locally, statewide, or nationally, teachers and administrators need to be proficient in interpreting the norm-referenced test scores.

The interpretation of some test scores is sufficiently complex to require that the user have relevant psychological training and experience or be assisted by and consult with persons who have such training and experience. Examples of such tests include individually administered intelligence tests, personality inventories, projective techniques, and neuropsychological tests.

Standard 13.14

In educational settings, score reports should be accompanied by a clear statement of the degree of measurement error associated with each score or classification level and information on how to interpret the scores.

Comment: This information should be communicated in a way that is accessible to persons receiving the score report. For instance, the degree of uncertainty might be indicated by a likely range of scores or by the probability of misclassification.

Standard 13.15

In educational settings, reports of group differences in test scores should be accompanied by relevant contextual information, where possible, to enable meaningful interpretation of these differences. Where appropriate contextual information is not available, users should be cautioned against misinterpretation.

Comment: Observed differences in test scores between groups (e.g., classified by gender, race/ethnicity, school/district, geographical region) can be influenced, fot example, by differences in course-taking patterns, in curriculum, in teacher's qualifications, or in parental educational level. Differences in performance of cohorts of students across time may be influenced by changes in the population of students tested or changes in learning opportunities for students. Users should be advised to consider the appropriate contextual information and cautioned against misinterpretation.

Standard 13.16

In educational settings, whenever a test score is reported, the date of test administration should be reported. This information and the age of any norms used for interpretation should be considered by test users in making inferences.

Comment: When a test score is used for a particular purpose, the date of the test score should be taken into consideration in determining its worth or appropriateness for making inferences about a student. Depending on the parricular domain measured, the validity of score inferences may be questionable as time ptogresses. For instance, a reading score from a test administered 6 months ago to an elementary school-aged student may no longer reflect the student's current reading level. Thus, a test score should not be used if it has been determined that undue time has passed since rhe time of data collection and that the score no longer can be considered a valid indicator of a student's current level of proficiency.

Standard 13.17

When change or gain scores are used, such scores should be defined and their technical qualities should be reported.

Comment: The use of change or gain scores presumes the same test or equivalent forms of the rest were used and that the test has (or the forms have) not been materially altered between administrations. The standard error of the difference between scores on the pretest and posttest, the regression of posttest scores on pretest scores, or relevant data from other reliable methods for examining change, such as those based on structural equation modeling, should be reported.

Standard 13.18

Documentation of design, models, scoring algorithms, and methods for scoring and classifying should be provided for tests administered and scored using multimedia or computers. Construct-irrelevant variance pertinent to computer-based testing and the use of other media in testing, such as the test taker's familiarity with technology and the test format, should be addressed in their design and use.

Comment: It is important to assure that the documentation does not jeopardize the security of the items that could adversely affect the validity of score interpretations. Computer and multimedia testing need to be held to the same requirements of technical quality as are other tests.

Standard 13.19

In educational settings, when average or summary scores for groups of students are reported, they should be supplemented with additional information about the sample size and shape or dispersion of score distributions.

Comment: Score reports should be designed to communicate clearly and effectively to their intended audiences. In most cases, reports that go beyond average score comparisons are helpful in furthering thoughtful use

and interpretation of test scores. Depending on the intended purpose and audience of the score report, additional information might take the form of standard deviations or other common measures of score variability, or of selected percentile points for each distribution. Alternatively, benchmark score levels might be established and then, for each group or region, the proportions of test takers attaining each specified level could be reported. Such benchmarks might be defined, for example, as selected percentiles of the pooled distribution for all groups or regions. Other distributional summaries of reporting formats may also be useful. The goal of more detailed reporting must be balanced against goals of clarity and conciseness in communicating test scores.

14. TESTING IN EMPLOYMENT AND CREDENTIALING

Background

Employment testing is carried out by otganizations for purposes of employee selection, promotion, or placement. *Selection* generally refers to decisions about which individuals will enter the organization; *placement* refers to decisions as to how to assign individuals to positions within the work force; and *promotion* refers to decisions about which individuals within the organization will advance. What all three have in common is a focus on the prediction of future job behaviors, with the goal of influencing organizational outcomes such as efficiency, growth, productivity, and employee motivation and satisfaction.

Testing used in the processes of licensure and certification, which will here generically be called credentialing, focuses on the applicant's current skill or competency in a specified domain. In many occupations, individuals must be licensed by governmental agencies in order to engage in the particular occupation. In other occupations, professional societies or other organizations assume responsibility for credentialing. Although licensure is typically a credential for entry into an occupation, credentialing programs may exist at varying levels, from novice to expert in a given field. Certification is usually sought voluntarily, although occupations differ in the degree to which obtaining certification influences employability or advancement. Testing is commonly only a part of a credentialing process, which may also include other requirements, such as education or supervised experiences. The *Standards* apply to the use of tests in the broader credentialing process.

Testing is also carried out in work organizations for a variety of purposes other than employment decision making and credentialing. Testing to detect psychopathology can take place, as in the case of an employee exhibiting behavioral problems at work. Testing as a tool for personal growth can be part of training and development programs, in which instruments measuring personality characteristics, interests, values, preferences, and work styles are commonly used with the goal of providing self-insight to employees. Testing can also take place in the context of program evaluation, as in the case of an experimental study of the effectiveness of a training program, where tests may be administered as pre- and post-measures. The focus of this chapter, though, is on the use of testing in employment and credentialing. Many issues relevant to such testing are discussed in other chapters: technical matters in chapters 1-6, fairness issues in chapters 7-10, general issues of test use in chapter 11, and individualized assessment of job candidates in chapter 12.

Employment Testing

The Influence of Context on Test Use

Employment testing involves using test information to aid in personnel decision making. Both the content and the context of employment testing varies widely. Content may cover various domains of knowledge, skills, abilities, traits, dispositions, and values. The context in which tests are used also varies widely. Some contextual features represent choices made by the employing organization; others represent constraints that must be accommodated by the employing organization. Decisions about the design, evaluation, and implementation of a testing system are specific to the context in which the system is to be used. Important contextual features include the following:

Internal vs. external candidate pool. In some instances, such as promotional settings, the candidates to be tested are already employed by the organization. In others, applications are sought from outside the

organization. In others, a mix of internal and external candidates is sought.

Untrained vs. specialized jobs. In some instances, untrained individuals are selected either because the job does not require specialized knowledge or skill or because the organization plans to offer training after the point of hire. In other instances, trained or experienced workers are sought with the expectation that they can immediately step into a specialized job. Thus, the same job may require very different selection systems depending on whether trained or untrained individuals will be hired or promoted.

Short-term vs. long-term focus. In some instances, the goal of the selection system is to predict performance immediately upon or shortly after hire. In other instances, the concern is with longer-term performance, as in the case of predictions as to whether candidates will successfully complete a multiyear overseas job assignment. Concerns about changing job tasks and job requirements also can lead to a focus on characteristics projected to be necessary for performance on the target job in the future, even if not a part of the job as currently constituted.

Screen in vs. screen out. In some instances, the goal of the selection system is to screen in individuals who will perform well on one set of behavioral or outcome criteria of interest to the organization. In others, the goal is to screen out individuals for whom the risk of pathological, deviant, or criminal behavior on the job is deemed too high. A testing system well suited to one objective may be completely inappropriate for another. That an individual is evaluated as a low risk for engaging in pathological behavior does not imply a prediction that the individual will exhibit high levels of job performance. That a test is predictive of one criterion does not support the inference of linkages to other criteria of interest as well.

Mechanical vs. judgmental decision making. In some instances, test information is used in a mechanical, standardized fashion. This is the case when scores on a test battery are combined by formula and candidates are selected in strict top-down rank order, or when only candidates above specific cut scores are eligible to continue to subsequent stages of a selection system. In other instances, information from a test is judgmentally integrated with information from other tests and with nontest information to form an overall assessment of the candidate.

Ongoing vs. one-time use of a test. In some instances, a test may be used for an extended period of time in an organization, permitting the accumulation of data and experience about the test in that context. In other instances, concerns about test security are such that repeated use is infeasible, and a new test is required for each test administration. For example, a work-sample test for lifeguards, requiring retrieving a mannequin from the bottom of a pool, is not compromised if candidates possess detailed knowledge of the test in advance. In contrast, a written job knowledge test may be severely compromised if some candidates have access to the test in advance. The key question is whether advance knowledge of test content changes the constructs measured by the test.

Fixed applicant pool vs. continuous flow. In some instances, an applicant pool can be assembled prior to beginning the selection process, as in the case of a policy that all candidates applying before a specific date will be considered. In other cases, there is a continuous flow of applicants about whom employment decisions need to be made on an ongoing basis. A ranking of candidates is possible in the case of the fixed pool; in the case of a continuous flow, a decision may need to be made about each candidate independent of information about other candidates.

Small vs. large sample size. Large sample sizes are sometimes available for jobs with many incumbents, in situations in which multiple similar jobs can be pooled, or in situa-

tions in which organizations with similar jobs collaborate in selection system development. In other situations, sample sizes are small; at the extteme is the case of the single-incumbent job. Sample size affects the degree to which different lines of evidence can be dtawn on in examining validity for the intended inference to be drawn from the test. For example, relying on the local setting for empirical linkages between test and criterion scores is not technically feasible with small sample sizes.

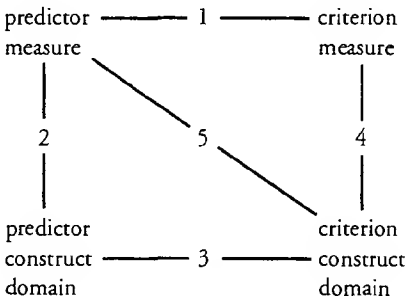
Size of applicant pool, relative to the number of job openings. The size of an applicant pool can constrain the type of testing system that is feasible. For desirable jobs, very large numbers of candidates may vie fot a small number of jobs. Under such scenarios, short screening tests may be used to reduce the pool to a size fot which the administration of mote time-consuming and expensive tests is practicable. Large applicant pools may also pose test security concerns, limiting the organization to testing methods that permit simultaneous test administration to all candidates.

Thus, test use by employers is conditioned by contextual features such as those in the foregoing list. Knowledge of these features plays an important part in the professional judgment that will influence both the type of testing system that will be developed and the strategy that will be used to evaluate critically the validity of the inference(s) drawn using the testing system.

The Validation Process in Employment Testing

The fundamental inference to be drawn from test scores in most applications of testing in employment settings is one of prediction: the test uset wishes to make an infetence from test results to some future job behavior or job outcome. Even when the validation strategy used does not involve empirical predictor-criterion linkages, as in the case of reliance on validity evidence based on test content, there is an implied criterion. Thus, while different sttategies of gathering evidence may be used, the infetence to be supported is that scores on the test can be used to predict subsequent job behavior. The validation process in employment settings involves the gathering and evaluation of evidence relevant to sustaining or challenging this inference. As detailed below, a variety of validation strategies can be used to support this inference.

It thus follows that establishing this predictive inference requires that attention be paid to two domains: that of the test (the predictor) and that of the job behavior or outcome of interest (the criterion). Evaluating the use of a test for an employment decision can be viewed as testing the hypothesis of a linkage between these domains. Operationally, there ate many ways of testing this hypothesis. This is illustrated by the following diagram:



The diagram differentiates between a predictor construct domain and a predictor measure and between a criterion construct domain and a criterion measure. A *predictor construct domain* is defined by specifying the set of behaviors that will be included under a particulat construct label (e.g., verbal reasoning, typing speed, conscientiousness). Similarly, a *criterion construct domain* specifies the set of job behaviors or job outcomes that will be included under a particular construct label (e.g., performance of core job tasks, teamwork, attendance, sales volume, overall job performance). Ptedictor and ctiterion measures are attempts at opetationalizing these domains.

The diagram enumerates a number of inferences commonly of interest. The first is the inference that scores on a predictor measure are related to scores on a criterion measure. This inference is tested through empirical examination of relationships between the two measures. The second and fourth are conceptually similar: both examine the inference that an operarional measure can be interpreted as representing an individual's standing on the construct domain of interest. Logical analysis, expert judgment, and convergence with or divergence from conceptually similar or different measures are among the forms of evidence that can be examined in testing these linkages. The third is the inference of a relationship between the predictor construct domain and the criterion construct domain. This linkage is established on the basis of theoretical and logical analysis. It commonly draws on systematic evaluation of job content and expert judgment as to the individual characteristics linked to successful job performance. The fifth represents the linkage between the predictor measure and the criterion construct domain.

Some predictor measures are designed explicitly as samples of the criterion construct domain of interest, and, thus, isomorphism between the measure and the construct domain constitutes direct evidence for linkage 5. Establishing linkage 5 in this fashion is the hallmark of approaches that rely heavily on what these *Standards* refer to as "validity evidence based on test content," referred to as content validity in prior conceptualizations of the validation process. Tests in which candidates for lifeguard positions perform rescue operations or in which candidates for word processor positions type and edit text exemplify this approach.

A prerequisite to the use of a predictor measure for personnel selection is that the linkage between the predictor measure and the criterion construct domain be established. As the diagram illustrates, there are multiple strategies for establishing this crucial linkage. One strategy is direct, via linkage 5; a second involves pairing linkage 1 and linkage 4; and a third involves pairing linkage 2 and linkage 3.

When the test is designed as a sample of the criterion construct domain, this linkage can be established directly via linkage 5. Another strategy for linking a predictor measure and the criterion construct domain focuses on linkages 1 and 4: pairing an empirical link between the predictor and criterion measures with evidence of the adequacy with which the criterion measure represents the criterion construct domain. The empirical link between the predictor measure and the criterion measure is part of what these *Standards* refer to as "validity evidence based on relationships to other variables," referred to as criterion-related validity in prior conceptualizations of the validation process. The empirical link of the test and the criterion measure must be supplemented by evidence of the relevance of the criterion measure to the criterion construct domain to complete the linkage between the test and the criterion construct domain. Evidence of the relevance of the criterion measure to the criterion construct domain is commonly based on job analysis, though in some cases the link between the domain and the measure is so direct that relevance is apparent without job analysis (e.g., when the criterion construct of interest is absenteeism or turnover). Note that this strategy does not necessarily rely on a well-developed predictor construct domain. Predictor measures such as empirically keyed biodata measures are constructed on the basis of empirical links between test item responses and the criterion measure of interest. Such measures may, in some instances, be developed without a fully established a priori conception of the predictor construct domain; the basis for their use is the direct empirical link between test responses and a relevant criterion measure.

Yet another strategy for linking predictor scores and the criterion construct domain focuses on pairing evidence of the adequacy with which the predictor measure represents the predictor construct domain (linkage 2)

with evidence of the linkage between the predictor construct domain and the criterion construct domain (linkage 3). As noted above, there is no single direct route to establishing these linkages. They involve lines of evidence subsumed under "construct validity" in prior conceptualizations of the validation process. A combination of lines of evidence, such as expert judgment of the characteristics predictive of job success, inferences drawn from an analysis of critical incidents of effective and ineffective job performance, and interview and observation methods, may support inferences about the predictor constructs linked to the criterion construct domain. Measures of these predictor constructs may then be selected or developed, and the linkage between the predictor measure and the predictor construct domain can be established with various lines of evidence for linkage 2 discussed above.

Thus multiple sources of data and multiple lines of evidence can be drawn on to evaluate the linkage between a predictor measure and the criterion construct domain of interest. There is not a single correct or even a preferred method of inquiry for establishing this linkage. Rather, the test user must consider the specifics of the testing situation and apply professional judgment in developing a strategy for testing the hypothesis of a linkage between the predictor measure and the criterion domain.

For many testing applications, there is a considerable cumulative body of research that speaks to some, if not all, of the inferences discussed above. A meta-analytic integration of this research can form an integral part of the strategy for linking test information to the construct domain of interest. The value of collecting local validation data varies with the magnitude, relevance, and consistency of research findings using similar predictor measures and similar criterion construct domains for similar jobs. In some cases, a small and inconsistent cumulative research record may lead to a validation strategy that relies heavily on local data; in others, a large, consistent research base may make investing resources in additional local data collection unnecessary.

Bases for Evaluating Test Use

While a primary goal of employment testing is the accurate prediction of subsequent job behaviors or job outcomes, it is important to recognize that there are limits to the degree to which such criteria can be predicted. Perfect prediction is an unattainable goal. First, behavior in work settings is also influenced by a wide variety of organizational and extra-organizational factors, including supervisor and peer coaching, formal and informal training, changes in job design, changes in organizational structures and systems, and changing family responsibilities, among others. Second, behavior in work settings is influenced by a wide variety of individual characteristics, including knowledge, skills, abilities, personality, and work attitudes, among others. Thus any single characteristic will be only an imperfect predictor, and even complex selection systems focus on the set of constructs deemed most critical for the job, rather than on all characteristics that can influence job behavior. Third, some measurement error always occurs even in well-developed test and criterion measures.

Thus, testing systems cannot be judged against a standard of perfect prediction but rather in terms of comparisons with available alternative selection methods. Professional judgment, informed by knowledge of the research literature about the degree of predictive accuracy relative to available alternatives, influences decisions about test use.

Decisions about test use are often influenced by additional considerations including utility (i.e., cost-benefit) evaluation, value judgments about the relative importance of selecting for one criterion domain vs. others, concerns about applicant reactions to test content and process, the availability and appropriateness of alternative selection methods, statutory or regulatory requirements governing test use, and social issues such as workforce

diversity. Organizational values necessarily come into play in making decisions about test use; organizations with comparable evidence supporting an intended inference drawn from test scores may thus reach different conclusions about whether to use any particular test.

Testing in Professional and Occupational Credentialing

Tests are widely used in the credentialing of persons for many occupations and professions. Licensing requiremenrs are imposed by state and local governments to ensure that those licensed possess knowledge and skills in sufficient degree to perform important occupational activities safely and effectively. Certification plays a similar role in many occupations not regulated by governments and is often a necessary precursor to advancement in many occupations. Certification has also become widely used to indicate that a person has certain specific skills (e.g., operation of specialized auto repair equipment) or knowledge (e.g., estate planning), which may be only a part of their occupational duties. Licensure and certification, as well as registry and other warrants of expertise, will here generically be called credentialing.

Tests used in credentialing are intended to provide the public, including employers and government agencies, with a dependable mechanism for identifying practirioners who have met particular standards. The standards are srrict, but not so stringenr as to unduly restrain rhe right of qualified individuals to offer their services to the public. Credentialing also serves to protect the profession by excluding persons who are deemed ro be not qualified to do the work of the occupation. Qualifications for credentials typically include educational requirements, some amount of supervised experience, and other specific criteria, as well as artainment of a passing score on one or more examinations. Tests are used in credentialing in a broad spectrum of professions and occupations, including medicine, law, psychology, teaching, architecture, real estate, and cosmetology. In some of these, such as actuarial science, clinical neuropsychology, and medical specialties, tesrs are also used to certify advanced levels of expertise. Relicensure or recertification is also required in some occupations and professions.

Tests used in credentialing are designed to determine whether the essential knowledge and skills of a specified domain have been mastered by the candidate. The focus of performance standards is on levels of knowledge and performance necessary for safe and appropriate practice. Test design generally starts with an adequate definition of the occupation or specialty, so that persons can be clearly identified as engaging in rhe activity. Then, the nature and requirements of the occupation, in its current form, are delineated. Often, a thorough analysis is conducted of the work performed by people in the profession or occupation to document the tasks and abilities that are essential to practice. A wide variety of empirical approaches is used, including delineation, critical incidence techniques, job analysis, training needs assessments, or practice studies and surveys of practicing professionals. Panels of respected experts in the field often work in collaboration with qualified specialists in tesring to define test specifications, including the knowledge and skills needed for safe, effective performance, and an appropriate way of assessing that performance. Forms of testing may include rraditional mulriple-choice tests, wrirren essays, and oral examinations. More elaborate performance tasks, somerimes using computer-based simulation, are also used in assessing such practice components as, for example, patient diagnosis or treatment planning. Hands-on performance tasks may also be used (e.g., operating a boom crane or filling a tooth) while being observed by one or more examiners.

Credentialing tests may cover a number of related but distinct areas. Designing the testing

program includes deciding what areas are to be covered, whether one or a series of tests is to be used, and how multiple test scores are to be combined to reach an overall decision. In some cases high scores on some tests are permitted to offset low scores on other tests, so that additive combination is appropriate. In other cases, an acceptable performance level is required on each test in an examination series.

Validation of credentialing tests depends mainly on content-related evidence, often in the form of judgments that the test adequately represents the content domain of the occupation or specialty being considered. Such evidence may be supplemented with other forms of evidence external to the test. Criterion-related evidence is of limited applicability in licensure settings because criterion measures are generally not available for those who are not granted a license.

Defining the minimum level of knowledge and skill required for licensure or certification is one of the most important and difficult tasks facing those responsible for credentialing. Verifying the appropriateness of the cut score or scores on the tests is a critical element in validity. The validity of the inference drawn from the test depends on whether the standard for passing makes a valid distinction between adequate and inadequate performance. Often, panels of experts are used to specify the level of performance that should be required. Standards must be high enough to protect the public, as well as the practitioner, but not so high as to be unreasonably limiting. Verifying the appropriateness of the cut score or scores on a test used for licensure or certification is a critical element of the validity of test results.

Legislative bodies sometimes attempt to legislate a cut score, such as a score of 70%. Arbitrary numerical specifications of cut scores are unhelpful for two reasons. First, without detailed information about the test, job requirements, and their relationship, sound standard setting is impossible. Second, without detailed information about the format of the test and the difficulty of items, such numerical specifications have little meaning.

Tests for credentialing need to be precise in the vicinity of the passing, or cut, score. They may not need to be precise for those who clearly pass or clearly fail. Sometimes a test used in credentialing is designed to be precise only in the vicinity of the cut score. Computer-based mastery tests may include a procedure to end the testing when a decision about the candidate's performance can be clearly made or when a maximum time limit is reached. This may result in a shorter test for candidates whose performance clearly exceeds or falls far below the minimum performance required for a passing score. The test taker may be told only whether the decision was pass or fail. Because such mastery tests are not designed to indicate how badly the candidate failed, or how well the candidate passed, providing scores that are much higher or lower than the cut score could be misleading. Nevertheless, candidates who fail are likely to profit from information about the areas in which their performance was especially weak. When feedback to candidates about how well or how poorly they performed is intended, precision throughout the score range is needed.

Practice in professions and occupations often changes over time. Evolving legal restrictions, progress in scientific fields, and refinements in techniques can result in a need for changes in test content. When change is substantial, it becomes necessary to revise the definition of the job, and the test content, to reflect changing circumstances. When major revisions are made in the test, the cut score that identifies required test performance is also reestablished.

Because credentialing is an ongoing process, with tests given on a regular schedule, new versions of the test are often needed. From a technical perspective, all versions of a test should be prepared to the same specifications and represent the same content.

Alternate test forms should have comparable score scales so that scores can retain their meaning. Various methods of jointly calibrating alternate forms can be used to assure that the srandard for passing represents the same level of performance on all forms. It may be nored rhat release of past test forms may compromise the quality of test form comparability.

Some credentialing groups consider it necessary, as a practical matter, to adjust their criteria yearly in order to regulate the number of accredited candidates entering the profession. This questionable procedure raises serious problems for the technical quality of the test scores. Adjusting the cut score annually implies higher standards in some years than in others, which, although open and straightforward, is difficult to justify on the grounds of quality of performance. Adjusting the score scale so that a certain number or proportion reach the passing score, while less obvious to the candidates, is technically inappropriate because it changes the meaning of the scores ftom year to year. Passing a credentialing examination should signify that the candidate meets the knowledge and skill standards set by the ctedentialing body, independent of the availability of work.

Issues of cheating and test security are of special importance for testing practices in credentialing. Issues of test security are covered in chapters 5 and 11. Issues of cheating by test takers are covered in chapter 8. Issues concerning the technical quality of tests are found in chapters 1-6, and issues of fairness in chapters 7-10.

Standard 14.1

Prior to development and implementation of an employment test, a clear statement of the objective of testing should be made. The subsequent validation effort should be designed to determine how well the objective has been achieved.

Comment: The objectives of employment tests can vary considerably. Some aim to screen out those least suited for the job in question, while others are designed to identify those best suited for the job. Tests also vary in the aspects of job behaviot they ate intended to predict, which may include quantity or quality of work output, tenure, counterptoductive behavior, and teamwork, among others.

Standard 14.2

When a test is used to predict a criterion, the decision to conduct local empirical studies of predictor-criterion relationships and interpretation of the results of local studies of predictor-criterion relationships should be grounded in knowledge of relevant research.

Comment: The cumulative literature on the relationship between a particular type of predictor and type of criterion may be sufficiently large and consistent to support the predictor-criterion relationship without additional research. In some settings, the cumulative research literature may be so substantial and so consistent that a dissimilar finding in a local study should be viewed with caution unless the local study is exceptionally sound. Local studies are of greatest value in settings where the cumulative research literature is sparse (e.g., due to the novelty of the predictor and/or criterion used), where the cumulative record is inconsistent, or where the cumulative literature does not include studies similar to the local setting (e.g., a test with a

large cumulative literature dealing exclusively with production jobs, and a local setting involving managerial jobs).

Standard 14.3

Reliance on local evidence of empirically determined predictor-criterion relationships as a validation strategy is contingent on a determination of technical feasibility.

Comment: Meaningful evidence of predictor-criterion relationships is conditional on a number of features, including (a) the job being relatively stable, rather than in a period of rapid evolution; (b) the availability of a relevant and reliable criterion measure; (c) the availability of a sample reasonably representative of the population of interest; and (d) an adequate sample size for estimating the strength of the predictor-criterion relationship.

Standard 14.4

When empirical evidence of predictor-criterion relationships is part of the pattern of evidence used to support test use, the criterion measure(s) used should reflect the criterion construct domain of interest to the organization. All criteria used should represent important work behaviors or work outputs, on the job or in job-relevant training, as indicated by an appropriate review of information about the job.

Comment: When criteria are constructed to represent job activities or behaviors (e.g., supervisory ratings of subordinates on important job dimensions), systematic collection of information about the job informs the development of the criterion measures, though there is no clear choice among the many available job analysis methods. There is not a clear need for job analysis to support criterion use when measures such as absenteeism or turnover are the criteria of interest.

Standard 14.5

Individuals conducting and interpreting empirical studies of predictor-criterion relationships should identify contaminants and artifacts that may have influenced study findings, such as error of measurement, range restriction, and the effects of missing data. Evidence of the presence or absence of such features, and of actions taken to remove or control their influence, should be retained and made available as needed.

Comment: Error of measurement in the criterion and restriction in the variability of predictor or criterion scores systematically reduce estimates of the relationship between predictor measures and the criterion construct domain, and procedures for correction for the effects of these artifacts are available. When these procedures are applied, both corrected and uncorrected values should be presented, along with the rationale for the correction procedures chosen. Statistical significance tests for uncorrected correlations should not be used with corrected correlations. Other features to be considered include issues such as missing data for some variables for some individuals, decisions about the retention or removal of extreme data points, the effects of capitalization on chance in selecting predictors from a larger set on the basis of strength of predictor-criterion relationships, and the possibility of spurious predictor-criterion relationships, as in the case of collecting criterion ratings from supervisors who know selection test scores.

Standard 14.6

Evidence of predictor-criterion relationships in a current local situation should not be inferred from a single previous validation study unless the previous study of the predictor-criterion relationship was done under favorable conditions (i.e., with a large sample size and a relevant criterion) and if the current situation corresponds closely to the previous situation.

Comment: Close correspondence means that the job requirements or underlying psychological constructs are substantially the same (as is determined by a job analysis), and that the predictor is substantially the same.

Standard 14.7

If tests are to be used to make job classification decisions (e.g., the pattern of predictor scores will be used to make differential job assignments), evidence that scores are linked to different levels or likelihoods of success among jobs or job groups is needed.

Standard 14.8

Evidence of validity based on test content requires a thorough and explicit definition of the content domain of interest. For selection, classification, and promotion, the characterization of the domain should be based on job analysis.

Comment: In general, the job content domain should be described in terms of job tasks or worker knowledge, skills, abilities, and other personal characteristics that are clearly operationally defined so that they can be linked to test content, and for which job demands are not expected to change substantially over a specified period of time. Knowledge, skills, and abilities included in the content domain should be those the applicant should already possess when being considered for the job in question.

Standard 14.9

When evidence of validity based on test content is a primary source of validity evidence in support of the use of a test in selection or promotion, a close link between test content and job content should be demonstrated.

Comment: For example, if the test content samples job tasks with considerable fidelity (e.g., actual job samples such as machine operation) or, in the judgment of experts, correctly simulates job task content (e.g., certain assessment center exercises), or samples specific job knowledge required for successful job performance (e.g., information necessary to exhibit certain skills), then content-related evidence can be offered as the principal form of evidence of validity. If the link between the test content and the job content is not clear and direct, other lines of validity evidence take on greater importance.

Standard 14.10

When evidence of validity based on test content is presented, the rationale for defining and describing a specific job content domain in a particular way (e.g., in terms of tasks to be performed or knowledge, skills, abilities, or other personal characteristics) should be stated clearly.

Comment: When evidence of validity based on test content is presented for a job or class of jobs, the evidence should include a description of the major job characteristics that a test is meant to sample, including the relative frequency, importance, or criticality of the elements.

Standard 14.11

If evidence based on test content is a primary source of validity evidence supporting the use of a test for selection into a particular job, a similar inference should be made about the test in a new situation only if the critical job content factors are substantially the same (as is determined by a job analysis), the reading level of the test material does not exceed that appropriate for the new job, and there are no discernible features of the new situation that would substantially change the original meaning of the test material.

Standard 14.12

When the use of a given test for personnel selection relies on relationships between a predictor construct domain that the test represents and a criterion construct domain, two links need to be established. First, there should be evidence for the relationship between the test and the predictor construct domain, and second, there should be evidence for the relationship between the predictor construct domain and major factors of the criterion construct domain.

Comment: There should be a clear conceptual rationale for these linkages. Both the predictor construct domain and the criterion construct domain to which it is to be linked should be defined carefully. There is no single route to establishing these linkages. Evidence in support of linkages between the two construct domains can include patterns of findings in the research literature and systematic evaluation of job content to identify predictor constructs linked to the criterion domain. The bases for judgments linking the predictor and criterion construct domains should be articulated.

Standard 14.13

When decision makers integrate information from multiple tests or integrate test and nontest information, the role played by each test in the decision process should be clearly explicated, and the use of each test or test composite should be supported by validity evidence.

Comment: A decision maker may integrate test scores with interview data, reference checks, and many other sources of information in making employment decisions. The inferences drawn from test scores should be limited to those for which validity evidence is available. For example, viewing a high test score as indicating overall job suitability, and thus precluding the need for reference checks, would be an inappropriate inference from a test measuring a single narrow, albeit relevant, domain, such as job knowledge. In other circumstances, decision makers integrate scores across multiple tests, or across multiple scales within a given test.

Standard 14.14

The content domain to be covered by a credentialing test should be defined clearly and justified in terms of the importance of the content for credential-worthy performance in an occupation or profession. A rationale should be provided to support a claim that the knowledge or skills being assessed are required for credential-worthy performance in an occupation and are consistent with the purpose for which the licensing or certification program was instituted.

Comment: Some form of job or practice analysis provides the primary basis for defining the content domain. If the same examination is used in the licensure or certification of people employed in a variety of settings and specialties, a number of different job settings may need to be analyzed. Although the job analysis techniques may be similar to those used in employment testing, the emphasis for licensure is limited appropriately to knowledge and skills necessary for effective practice. The knowledge and skills contained in a core curriculum designed to train people for the job or occupation may be relevant, especially if the curriculum has been designed to be consistent with empirical job or practice analyses. In tests used for licensure, skills that may be important to success but are not directly related to the purpose of licensure (e.g., protecting the public) should not be included. For example, in real estate, marketing skills may be important for success as a broker, and assessment of these skills might have utility for agencies selecting brokers for

employment. However, lack of these skills may not present a threat to the public and would appropriately be excluded from consideration for a licensing examination. The fact that successful practitioners possess certain knowledge or skills is relevant but not persuasive. Such information needs to be coupled with an analysis of the purpose of a licensing program and the reasons that the knowledge or skill is required in an occupation or profession.

Standard 14.15

Estimates of the reliability of test-based credentialing decisions should be provided.

Comment: The standards for decision reliability described in chapter 2 are applicable to tests used for licensure and certification. Other types of reliability estimates and associated standard errors of measurement may also be useful, but the reliability of the decision of whether or not to certify is of primary importance.

Standard 14.16

Rules and procedures used to combine scores on multiple assessments to determine the overall outcome of a credentialing test should be reported to test takers, preferably before the test is administered.

Comment: In some cases, candidates may be required to score above a specified minimum on each of several tests. In other cases, the pass-fail decision may be based solely on a total composite score. While candidates may be told that tests will be combined into a composite, the specific weights given to various components may not be known in advance (e.g., to achieve equal effective weights, nominal weights will depend on the variance of the components).

Standard 14.17

The level of performance required for passing a credentialing test should depend on the knowledge and skills necessary for acceptable performance in the occupation or profession and should not be adjusted to regulate the number or proportion of persons passing the test.

Comment: The number or proportion of persons granted credentials should be adjusted, if necessary, on some basis other than modifications to either the passing score or the passing level. The cut score should be determined by a careful analysis and judgment of acceptable performance. When there are alternate forms of the test, the cut score should be carefully equated so that it has the same meaning for all forms.

15. TESTING IN PROGRAM EVALUATION AND PUBLIC POLICY

Background

Tests are widely used in program evaluation and in public policy decision making. Program evaluation is the set of procedures used to make judgments about the client's need for a program, the way it is implemented, its effectiveness, and its value. Policy studies are somewhat broader than program evaluations and refer to studies that contribute to judgments about plans, principles, or procedures enacted to achieve broad public goals. There is no sharp distinction between policy studies and program evaluations, and in many instances there is substantial overlap between the two types of investigations. Test results are often one important source of evidence for the initiation, continuation, modification, termination, or expansion of various programs and policies.

Intetpretation of test scores in program evaluation and policy studies usually entails the complex analysis of a number of variables. For example, some programs are mandated for a broad population; others target only certain subgroups. Some are designed to affecr attitudes, while others are intended to have a more direct impact on behaviot. It is important that the participants included in any study at least meet the specified criteria for the program or policy under review so that appropriate intetpretation of test results will be possible. Test results will reflect not only the effecrs of rules for participant selection and the impact of participarion in different programs ot treatments, but also the characteristics of those tested. Relevant background informarion abour clients or students may be obtained in order to strengthen the inferences derived from the rest results. Valid interpretations may depend upon additional considerations that have nothing to do with the appropriateness of the test or its technical quality, including study design, administrative feasibility, and the quality of other available data. It is not the intent of this chapter to deal with these varied considerations in any substantial way. In order to develop defensible conclusions, however, investigators conducting program evaluations and policy studies are encouraged to supplement test results with data from other sources. These include information about program characteristics, delivery, costs, client backgrounds, degree of participation, and evidence of side effects. Because test results lend important weight ro evaluation and policy studies, it is critical that any tests used in these investigations be sensitive to the questions of the study and appropriate for the test takers.

It is important to evaluate any proposed test in terms of its relevance to the goals of the program or policy and/or to the particular question its use will address. It is relatively rare for a test to be designed specifically for program evaluation or policy study purposes. Typically, the instruments used in such studies were originally developed for purposes other than program or policy evaluation. In addition, because of cost or convenience, certain tests may be adopted for use in a program evaluation or policy study even though they may have been developed for a somewhat diffetent population of respondents. Some tests may be selected for use in program evaluation or policy studies because the tests are well known and thought to be especially credible ro the clients or the public consumer. Even though certain tesrs may be more familiar to the public or may be less rime-consuming or less expensive ro use than an instrument developed specifically for the evaluation, they may be nonetheless inappropriate for use as criterion measures to determine rhe need for or to evaluate the effects of particular interventions.

As government agencies and other insritutions move to improve their own routine data collection capability, fewer special studies are

conducted to evaluate programs and policies. Instead, evaluations and policy studies may depend upon a special analysis of data previously collected for other purposes. In these cases, the investigators may reanalyze test data already obtained and analyzed for another purpose in order to make inferences about program or policy effectiveness. This procedure is called *secondary data analysis*. In some circumstances, it may be difficult to assure a good match between the existing test and the intervention or the policy under examination. Moreover, it may be difficult to reconstruct in detail the conditions under which the data were originally collected. Secondary data analysis also requires consideration of whether adequate informed consent was obtained from subjects in the original data collection to allow secondary analysis to occur without obtaining additional consent. In selecting (or developing) a test or in deciding to use existing data in evaluation and policy studies, careful investigators attempt to balance the purpose of the test, its likelihood to be sensitive to the intervention under study, the credibility of the test to interested parties, and the costs of its administration. Otherwise, test results may lead to inappropriate interpretations about the progress, impact, and overall value of programs and policies under review.

Program Evaluation

Tests may be used in program evaluations to provide information on the status of clients or students before, during, or following an intervention, as well as to provide information on appropriate comparison groups. Whereas understanding the performance of an individual student or client is often the goal of many testing activities, program evaluation targets the performance of, or impact on, groups. Tests are used in program evaluations in a variety of fields, such as social services, education, health services, and military and employment training. The term *program*, broadly interpreted, describes interventions that range from large-scale state or national programs with provisions for local flexibility to small-scale, more experimental projects. In many cases, evaluation is mandated by the agency or funding source for the program, and the intervention is evaluated by judging its effectiveness in meeting stated goals. Some examples of programs that might use test results as part of their evaluation data include psychotherapeutic services, military training programs and job placement programs, school curricula, or services for individuals with special needs.

Test results, along with other information, may be used to compare competing interventions, such as alternative reading curricula or different psychotherapeutic interventions, or to describe the long-term pattern of effects for one or more groups. It is often important to assess a program for its differential effectiveness in meeting the needs of subgroups (such as different ethnic or gender groups within the target population). Even though the performance of groups is of primary interest in program evaluation, the analysis of individuals' histories and test performances may provide additional useful information to aid in the interpretation of test results.

Because of administrative realities, such as cost constraints and response burden, methodological refinements may be adopted to increase the efficiency of testing. One strategy is to obtain a sample of participants to be evaluated from the larger set of those exposed to a program or policy. When there is a sufficient number of clients affected by the program or policy to be evaluated, and when there is a desire to limit the time spent on testing, evaluators can create multiple forms of shorter tests from a larger pool of items. By constructing a number of different test forms consisting of relatively few items and assigning these test forms to different subsamples of test takers (a procedure known as matrix sampling), a larger number of items can be included in the study than could reasonably be administered to any

single test taker. When it is desirable to represent a domain with a large number of test items, this approach is often used. However, individual scores are not usually created or interpreted when matrix sampling is employed. Because procedures for sampling individuals or test items may vary in a number of ways, adequate analysis and interpretation of test results for any study depend upon a clear description of how samples were formed and the manner in which test results were aggregated.

Policy Uses of Tests

As noted previously, tests are also used in policy analyses, and the distinction between program evaluation and policy uses of tests is often a matter of degree. Programs are expected to share particular goals, procedures, and resources. Policy is a broader term, applying to plans, principles, procedures, or programs enacted to achieve particular goals in different settings. Programs provide direct services or interventions. Policies may be constructed to achieve their goals by direct or indirect means. Indeed, one direct approach used to achieve a policy goal might include the funding of specific programs. Other examples of direct policy approaches might involve the provision of training resources to improve performance in particular health-service occupations, or the enactment of new recertification requirements for accountants. Studies of the need for or impact of both of these policies could in part depend upon the analyses of test results. To illustrate in more depth, to meet the general policy objective of containing the costs of health care, direct policies might include giving incentives to clients to participate in fitness programs and the development of patient education programs. Tests could measure the understandings and attitudes of participants about the relationship of fitness to the prevention of illness. Another policy example, using a more indirect approach, is to encourage educators to create more effective programs for children from low-income families. As an approach, a state's educational authorities might require the separate reporting of test scores for children in high-poverty areas. Large differences in group performance would be expected to attract the attention of the public and to place greater pressure on the schools to improve the performance of particular groups of children.

In decentralized governments, policy implementation may be left to local authorities and may be interpreted in a number of different ways. As a result, it may be difficult to select or develop a single test or outcome measure that will be sensitive to the range of different activities or tactics used to implement a given policy. For that reason, policy studies may often use more than one test or outcome measure to provide a more adequate picture of the range of effects.

Issues in Program and Policy Evaluation

Test results are sometimes used as one way to inspire program administrators as well as to infer institutional effectiveness. This use of tests, including the public reporting of results, is thought to encourage an institution to improve its services for its clients. For example, consistently poor achievement test results may trigger special management attention for public schools in some locales. The interpretation of test results is especially complex when tests are used both as an institutional policy mechanism and as a measure of effectiveness. For example, a policy or program may be based on the assumption that providing clear goals and general specifications of test content (such as the type of topics, constructs and cognitive domains, and responses included in the test) may be a reasonable strategy to communicate new expectations to educators. Yet, the desire to influence test or evaluation results to show acceptable institutional performance could lead to inappropriate testing practices, such as

teaching the test items in advance, modifying test administration procedures, discouraging certain students or clients from participating in the testing sessions, or focusing exclusively on test-taking procedures. These practices might occur instead of those aimed at helping the test taker learn the domains measured by the test. Because results derived from such practices might lead to spuriously high estimates of impact and might reflect the negative side effects of this particular policy, diligent investigators may estimate the impact of such consequences in order to interpret the test results appropriately. Looking at possible inappropriate consequences of tests as well as their benefits will better assess policy claims that particular types of testing programs lead to improved performance.

On the other hand, policy studies and program evaluations often do not make available reports of results to the test takers and may give no clear reasons to the test taker for participating in the testing procedure. For example, when matrix sampling is used for program evaluation, it may not be feasible to provide such reports. If little effort is made to motivate the test taker to regard the test seriously (for instance, if the purpose of the test is not explained to the test taker), it is possible that test takers might have little reason to try to perform well on the test. Obtained test results then might well underrepresent the impact of the program, institution, or policy because of poor motivation on the part of the test taker. When there is a suspicion that the test might not have been taken seriously, motivation of test takers may be explored by collecting additional information, using observation or interview methods. The issues of inappropriate preparation or unmotivated performance are examples that raise basic questions about the validity of interpretations of test results. In every case, it is important to consider the potential impact of the testing process itself, including test administration and reporting practices, on the test taker.

Public policy decisions are rarely based solely on the results of empirical studies, even when the studies have been well done. The more expansive and indirect the policy, the more likely will it be that other considerations will come into play, such as the political and economic impact of abandoning, changing, or retaining the policy, or the reaction to offering rewards or sanctions to institutions. In a political climate, tests used in policy settings may be subjected to intense and detailed scrutiny. When results do not support a favored position, attempts may be made to discount the appropriateness of the testing procedure, construct, or interpretation.

It is important that all tests used in public evaluation or policy contexts meet the standards described in earlier chaptets. As described in chapter 8, tests are to be administered by trained personnel. It is also essential that assistance be provided to those tesponsible for interpreting study results to practitioners, to the lay public, and to the media. Careful communication of the study's goals, procedures, findings, and limitations increases the chances that the public's interpretations will be accurate and useful.

Additional Considerations

This chapter and its associated standards are directed to users of tests in program evaluation and policy studies and to the conditions under which those studies are usually conducted. Other standards documents that are relevant to this chapter include *The Program Evaluation Standards: How to Assess Evaluations of Educational Programs,* prepared by the Joint Committee on Standards for Educational Evaluation (2nd ed., Thousand Oaks, CA: Sage Publications, 1994), and the *Code of Fair Testing Practices in Education,* prepared by the Joint Committee on Testing Ptactices (Washington, DC: Joint Committee on Testing Practices, 1988).

Standard 15.1

When the same test is designed or used to serve multiple purposes, evidence of technical quality for each purpose should be provided.

Comment: In educational resting, for example, it has become common practice to use the same test for multiple purposes (e.g., monitoring achievement of individual students, providing information to assist in instructional planning for individuals or groups of students, evaluating schools or districts). No test will serve all purposes equally well. Choices in test development and evaluation that enhance validity for one purpose may diminish validity for other purposes. Different purposes require somewhat diffetent kinds of technical evidence, and appropriate evidence of technical quality for each purpose should be provided by the test developer. If the test uset wishes to use the test for a purpose not supported by the available evidence, it is incumbent on the user to provide the necessary additional evidence.

Standard 15.2

Evidence should be provided of the suitability of a test for use in evaluation or policy studies, including the relevance of the test to the goals of the program or policy under study and the suitability of the test for the populations involved.

Comment: Faulty inferences may be made when test scores are not sensitive to the features of a particular intervention. For instance, a test designed for selecrion may be ineffecrive as a measure of rhe effects of an intervention. It is also important to employ tesrs that are appropriate for the age and background of test takers.

Standard 15.3

When change or gain scores are used, the definition of such scores should be made explicit, and their technical qualities should be reported.

Comment: The use of change or gain scores presumes rhat the same test or equivalent forms of the test were used and that the test (or forms) have not been materially alteted between administrations. The standard error of the difference between scores on pretests and posttests, the regression of posttest scores on pretest scores, or relevant data from other reliable methods for examining change, such as those based on structural equation modeling, should be reported.

Standard 15.4

In program evaluation or policy studies, investigators should complement test results with information from other sources to generate defensible conclusions based on the interpretation of test results.

Comment: Descriptions or analyses of such variables as client selection criteria, services, clients, setting, and resources are often needed to provide a comprehensive picture of the program or policy under review and to aid in the interpretation of test results. Performance on indicators other than tests is almost always useful and in many cases is essential. Examples of other information include attrition rates or patterns of participation. Another source of information might be ro derermine rhe degree of motivarion of the test takers. When individual scores are not reported ro test takers, it is important to determine whether the examinees took the test experience seriously.

Standard 15.5

Agencies using tests to conduct program evaluations or policy studies, or to monitor outcomes, should clearly describe the population the program or policy is intended to serve and should document the extent to which the sample of test takers is representative of that population.

Comment: For example, a clinic with a diverse client population using testing to assess the outcome of a particular treatment may routinely report the extent of participation by subgroups of clients, for instance, those of diverse ethnic backgrounds or for whom English is a second language.

Standard 15.6

When matrix sampling procedures are used for program evaluation or population descriptions, rules for sampling items and test takers should be provided, and reliability analyses must take the sampling scheme into account.

Standard 15.7

When educational testing programs are mandated by school, district, state, or other authorities, the ways in which test results are intended to be used should be clearly described. It is the responsibility of those who mandate the use of tests to identify and monitor their impact and to minimize potential negative consequences. Consequences resulting from the uses of the test, both intended and unintended, should also be examined by the test user.

Comment: Mandated testing programs are often justified in terms of their potential benefits for teaching and learning. Concerns have been raised about the potential negative impact of mandated testing programs, particularly when they affect important decisions for individuals or institutions. To the extent possible, students, parents, and staff should be informed of the domains on which the students will be tested, the nature of the item types, and the standards for mastery. Effort should be made to document the provision of instruction in tested content and skills, even though it may not be possible or feasible to determine the specific content of instruction for every student. An example of negative impact is the use of strategies to raise performance artificially.

Standard 15.8

When it is clearly stated or implied that a recommended test use will result in a specific outcome, the basis for expecting that outcome should be presented, together with relevant evidence.

Comment: A given claim for the benefits of test use, such as improving students' achievement, may be supported by logical or theoretical argument as well as empirical data. Due weight should be given to findings in the scientific literature that may be inconsistent with the stated claim.

Standard 15.9

The integrity of test results should be maintained by eliminating practices designed to raise test scores without improving performance on the construct or domain measured by the test.

Comment: Such practices may include teaching test items in advance, modifying test administration procedures, and discouraging or excluding certain test takers from taking the test. These practices can lead to spuriously high scores that do not reflect performance on the underlying construct or domain of interest.

Standard 15.10

Those who have a legitimate interest in an assessment should be informed about the purposes of testing, how tests will be administered and scored, how long records will be retained, and to whom and under what conditions the records may be released.

Comment: Those with a legitimate interest may include the test takers, their parents or guardians, or personnel who may be affected by results (teachers, program staff).

Standard 15.11

When test results are released to the public or to policymakers, those responsible for the release should provide and explain any supplemental information that will minimize possible misinterpretations of the data.

Comment: The context and limitations of the study should be described, with particular attention given to methods of causal inferences.

Standard 15.12

Reports of group differences in average test scores should be accompanied by relevant contextual information, where possible, to enable meaningful interpretation of these differences. Where appropriate contextual information is not available, users should be cautioned against misinterpretation.

Comment: Observed differences in average test scores between groups (e.g., classified by gender, race/ethnicity, or geographical region) can be influenced, for example, by differences in life experiences, training experience, effort, instructor quality, or level and type of parental support. In education, differences in group performance across time may be influenced by changes in the population of those tested or changes in their experiences. Users should be advised to consider the appropriate contextual information and be cautioned against misinterpretation.

Standard 15.13

Those who mandate testing programs should ensure that the individuals who interpret the test results to make decisions within the school or program context are qualified to assume this responsibility and proficient in the appropriate methods for interpreting test results.

Comment: When testing programs are used as a strategy for guiding interventions or instruction, professionals expected to make inferences leading to program improvement may need assistance in interpreting test results for this purpose.

The interpretation of some test scores is sufficiently complex to require that the user have relevant psychological training and experience. Examples of such tests include individually administered intelligence tests, personality inventories, projective techniques, and neuropsychological tests.

GLOSSARY

This glossary provides definitions of terms as used in this text. For many of the terms, multiple definitions can be found in the literature; also, technical usage may differ from common usage.

---

ability/trait parameter In item response theory (IRT), a theoretical value indicating the level of a test taker on the ability or trait measured by the test; analogous to the concept of true score in classical test theory.

ability testing The use of standardized tests to evaluate the current performance of a person in some defined domain of cognitive, psychomotor, or physical functioning.

absolute score interpretation The meaning of a test score for an individual or an average score for a defined group, indicating an individual's or group's level of performance in some defined criterion domain. By contrast, see *relative score interpretation.*

accommodation See *test modification.*

acculturation The process whereby individuals from one culture adopt the characteristics and values of another culture with which they have come in contact.

achievement levels/proficiency levels Descriptions of a test taker's competency in a particular area of knowledge or skill, usually defined as ordered categories on a continuum, often labeled from "basic" to "advanced," or "novice" to "expert," that constitute broad ranges for classifying performance. See *cut score.*

achievement testing A test to evaluate the extent of knowledge or skill attained by a test taker in a content domain in which the test taker had received instruction.

adaptive testing A sequential form of individual testing in which successive items, or sets of items, in the test are chosen based primarily on their psychometric properties and content, in relation to the test taker's responses to previous items.

adjusted validity/reliability coefficient A validity or reliability coefficient—most often, a product-moment correlation—that has been adjusted to offset the effects of differences in score variability, criterion variability, or the unreliability of test and/or criterion. See *restriction of range or variability.*

age equivalent The chronological age in a defined population for which a given score is the median (middle) score. Thus, if children 10 years and 6 months of age have a median score of 17 on a test, the score 17 is said to have an age equivalent of 10-6 for that population. See *grade equivalent.*

alternate forms Two or more versions of a test that are considered interchangeable, in that they measure the same constructs in the same ways, are intended for the same purposes, and are administered using the same directions. *Alternate forms* is a generic term used to refer to any of three categories. *Parallel forms* have equal raw score means, equal standard deviations, equal error structures, and equal correlations with other measures for any given population. *Equivalent forms* do not have the statistical similarity of parallel forms, but the dissimilarities in raw score statistics are compensated for in the conversions to derived scores or in form-specific norm tables. *Comparable forms* are highly similar in content, but the degree of statistical similarity has not been demonstrated. See *linkage.*

analytic scoring A method of scoring in which each critical dimension of performance

is judged and scored separately, and the resultant values are combined for an overall score. In some instances, scores on the separate dimensions may also be used in interpreting performance. See *holistic scoring.*

anchor test A common set of items administered with each of two or more different forms of a test for the purpose of equating the scores obtained on these forms.

assessment Any systematic method of obtaining information from tests and other sources, used to draw inferences about characteristics of people, objects, or programs.

attention assessment The process of collecting data and making an appraisal of a person's ability to focus on the relevant stimuli in a situation. The assessment may be directed at mechanisms involved in arousal, sustained attention, selective attention and vigilance, or limitation in the capacity to attend to incoming information.

automated narrative report See *computer-prepared test interpretation.*

back translation A translation of a test, which is itself a translation from an original test, back into the language of the original test. The degree to which a back translation matches the original test indicates the accuracy of the original translation.

battery A set of tests usually administered as a unit. The scores on the several tests usually are scaled so that they can readily be compared or used in combination for decision making.

bias In a statistical context, a systematic error in a test score. In discussing test fairness, bias may refer to construct underrepresentation or construct-irrelevant components of test scores that differentially affect the performance of different groups of test takers. See *predictive bias, construct underrepresentation, construct irrelevance.*

bilingual The characteristic of being relatively proficient in two languages.

calibration 1. In linking test score scales, the process of setting the test score scale, including mean, standard deviation, and possibly shape of score distribution, so that scores on a scale have the same relative meaning as scores on a related scale. **2.** In item response theory, the process of determining the parameters of the response function for an item.

certification A voluntary process, often national in scope, by which individuals who have been certified have demonstrated some level of knowledge and skill in an occupation. See *licensing, credentialing.*

classical test theory A psychometric theory based on the view that an individual's observed score on a test is the sum of a true score component for the test taker, plus an independent measurement error component.

classification accuracy The degree to which neither false positive nor false negative categorizations and diagnoses occur when a test is used to classify an individual or event. See *sensitivity* and *specificity.*

coaching Planned short-term instructional activities in which prospective test takers participate prior to the test administration for the primary purpose of improving their test scores. Coaching typically includes simple practice, instruction on test-taking strategies, and related activities. Activities that approximate the instruction provided by regular school curricula or training programs are not typically referred to as coaching.

coefficient alpha An internal consistency reliability coefficient based on the number

of parts into which the test is partitioned (e.g., items, subtests, or raters), the interrelationships of the parts, and the total test score variance. Also called *Cronbach's alpha* and, for dichotomous items, *KR 20.*

cognitive assessment The process of systematically gathering test scores and related data in order to make judgments about an individual's ability to perform various mental activities involved in the processing, acquisition, retention, conceptualization, and organization of sensory, perceptual, verbal, spatial, and psychomotor information.

composite score A score that combines several scores according to a specified formula.

computer-administered test A test administered by a computer. Questions appear on a computer-produced display, and the test taker answers by using a keyboard, "mouse" or other similar response device.

computer-based mastery test An adaptive test administered by computer that indicates whether or not the test taker has mastered a certain domain. The test is not designed to provide scores indicating degree of mastery, but only whether the test performance was above or below some specified level. Thus a *computer-based mastery test* is not simply a *mastery test* given by computer. See *mastery test.*

computer-based test See *computer-administered test.*

computer-generated test interpretation See *computer-prepared test interpretation.*

computer-prepared test interpretation A programmed, computer-prepared interpretation of an examinee's test results, based on empirical data and/or expert judgment.

computerized adaptive test An adaptive test administered by computer. See *adaptive testing.*

conditional measurement error variance The variance of measurement errors that affect the scores of examinees at a specified test score level; the square of the conditional standard error of measurement.

conditional standard error of measurement The standard deviation of measurement errors that affect the scores of examinees at a specified test score level.

confidence interval An interval between two values on a score scale within which, with specified probability, a score or parameter of interest lies. The term is also used in these standards to designate Bayesian credibility intervals that define the probability that the unknown parameter falls in the specified interval.

configural scoring rule A rule for scoring a set of two or more elements (such as items or subtests) in which the score depends on a particular pattern of responses to the elements.

construct The concept or the characteristic that a test is designed to measure.

construct domain The set of interrelated attributes (e.g., behaviors, attitudes, values) that are included under a construct's label. A test typically samples from this construct domain.

construct equivalence 1. The extent to which the construct measured by one test is essentially the same as the construct measured by another test. 2. The degree to which a construct measured by a test in one cultural or linguistic group is comparable to the construct measured by the same test in a different cultural or linguistic group.

construct irrelevance The extent to which test scores are influenced by factors that are irrelevant to the construct that the test is

intended to measure. Such extraneous factors distort the meaning of test scores from what is implied in the proposed interpretation.

construct underrepresentation The extent to which a test fails to capture important aspects of the construct that the test is intended to measure. In this situation, the meaning of test scores is narrower than the proposed interpretation implies.

construct validity A term used to indicate that the test scores are to be interpreted as indicating the test taker's standing on the psychological construct measured by the test. A construct is a theoretical variable inferred from multiple types of evidence, which might include the interrelations of the test scores with other variables, internal test structure, observations of response processes, as well as the content of the test. In the current standards, all test scores are viewed as measures of some construct, so the phrase is redundant with validity. The validity argument establishes the construct validity of a test. See *construct, validity argument.*

constructed response item An exercise for which examinees must create their own responses or products rather than choose a response from an enumerated set. Short-answer items require a few words or a number as an answer, whereas extended-response items require at least a few sentences.

content domain The set of behaviors, knowledge, skills, abilities, attitudes or other characteristics to be measured by a test, represented in a detailed specification, and often organized into categories by which items are classified.

content standard A statement of a broad goal describing expectations for students in a subject matter at a particular grade or at the completion of a level of schooling.

content validity A term used in the 1974 *Standards* to refer to a *kind* or *aspect* of validity that was "required when the test user wishes to estimate how an individual performs in the universe of situations the test is intended to represent" (p. 28). In the 1985 *Standards*, the term was changed to *content-related evidence* emphasizing that it referred to one type of evidence within a unitary conception of validity. In the current *Standards*, this type of evidence is characterized as "evidence based on test content."

convergent evidence Evidence based on the relationship between test scores and other measures of the same construct.

credentialing Granting to a person, by some authority, a credential, such as a certificate, license, or diploma, that signifies an acceptable level of performance in some domain of knowledge or activity.

criterion domain The construct domain of a variable used as a criterion. See *construct domain.*

criterion-referenced score interpretation See *criterion-referenced test.*

criterion-referenced test A test that allows its users to make score interpretations in relation to a functional performance level, as distinguished from those interpretations that are made in relation to the performance of others. Examples of criterion-referenced interpretations include comparison to cut scores, interpretations based on expectancy tables, and domain-referenced score interpretations.

cross-validation A procedure in which a scoring system or set of weights for predicting performance, derived from one sample, is applied to a second sample in order to investigate the stability of prediction of the scoring system or weights.

cut score A specified point on a score scale, such that scores at or above that point are interpreted or acted upon differently from scores below that point. See *performance standard.*

derived score A score to which raw scores are converted by numerical transformation (e.g., conversion of raw scores to percentile ranks or standard scores).

diagnostic and intervention decisions Decisions based upon inferences derived from psychological test scores as part of an assessment of an individual that lead to placing the individual in one or more categories. See also *intervention planning.*

differential item functioning A statistical property of a test item in which different groups of test takers who have the same total test score have different average item scores or, in some cases, different rates of choosing various item options. Also known as DIF.

discriminant evidence Evidence based on the relationship between test scores and measures of different constructs.

documentation The body of literature (e.g., test manuals, manual supplements, research reports, publications, user's guides, etc.) made available by publishers and test authors to support test use.

domain sampling The process of selecting test items to represent a specified universe of performance.

empirical evidence Evidence based on some form of data, as opposed to that based on logic or theory. As used here, the term does not specify the type of evidence; this is in contrast to some settings where the term is equated with criterion-related evidence of validity.

equated forms Two or more test forms constructed to cover the same explicit content, to conform to the same statistical specifications, and to be administered under identical procedures (*alternate forms*); through statistical adjustments, the scores on the alternate forms share a common scale.

equating Putting two or more essentially parallel tests on a common scale. See *alternate forms.*

equivalent forms See *alternate forms.*

error of measurement The difference between an observed score and the corresponding true score or proficiency. See *standard error of measurement* and *true score.*

factor **1.** Any variable, real or hypothetical, that is an aspect of a concept or construct. **2.** In measurement theory, a statistical dimension defined by a factor analysis. See *factor analysis.*

factor analysis Any of several statistical methods of describing the interrelationships of a set of variables by statistically deriving new variables, called factors, that are fewer in number than the original set of variables.

factorial structure **1.** The set of factors obtained in a factor analysis. **2.** Technically, the correlation of each factor with each of the original variables from which the factors are derived.

fairness In testing, the principle that every test taker should be assessed in an equitable way. See chapter 7.

false negative In classification, diagnosis, or selection, an error in which an individual is assessed or predicted not to meet the criteria for inclusion in a particular group but in truth does (or would) meet these criteria. See *sensitivity* and *specificity.*

false positive In classification, diagnosis, or selection, an error in which an individual is assessed or predicted to meet the criteria for inclusion in a particular group but in truth does not (or would not) meet these criteria. See *sensitivity* and *specificity.*

field test A test administration used to check the adequacy of testing procedures, generally including test administration, test responding, test scoring, and test reporting. A field test is generally more extensive than a pilot test. See *pilot test.*

flag An indicator attached to a test score, a test item, or other entity to indicate a special status. A flagged test score generally signifies a score obtained in a modified, nonstandatd test administration. A flagged test item generally signifies an item with undesirable characteristics, such as excessive differential item functioning.

functional equivalence In evaluating test translations, the degree to which similar activities or behaviors have the same functions in different cultural or linguistic groups.

gain score In testing, the difference between two scores obtained by a test taket on the same test or two equated tests taken on different occasions, often before and after some treatment.

generalizability coefficient A reliability index encompassing one or more independent sources of error. It is formed as the ratio of (a) the sum of vatiances that are considered components of test score variance in the setting under study to (b) the foregoing sum plus the weighted sum of variances attributable to various error sources in this setting. Such indices, which atise from the application of generalizability theory, are typically interpteted in the same manner as reliability coefficients. See *generalizability theory.*

generalizability theory An extension of classical reliability theory and methodology in which the magnitudes of errors from specified sources are estimated through the use of one or another experimental design, and the application of the statistical techniques of the analysis of variance. The analysis indicates the genetalizability of scores beyond the specific sample of items, persons, and observational conditions that were studied.

grade equivalent The school grade level for a given population for which a given score is the median score in that population. See *age equivalent.*

high-stakes test A test used to provide results that have important, direct consequences for examinees, programs, or institutions involved in the testing.

holistic scoring A method of obtaining a scote on a test, ot a test item, based on a judgment of overall petformance using specified criteria. See *analytic scoring.*

informed consent The agreement of a person, or that person's legal representative, for some procedure to be performed on or by the individual, such as taking a test or completing a questionnaire. The agreement, which is usually written, is made after the nature, possible effects, and use of the procedure has been explained.

intelligence test A psychological or educational test designed to measure an individual's level of cognitive functioning in accord with some recognized theory of intelligence.

internal consistency coefficient An index of the reliability of test scores derived from the statistical interrelationships of responses among item responses or scores on separate parts of a test.

internal structure In test analysis, the factorial structure of item responses or subscales of a test. See *factorial structure.*

inter-rater agreement The consistency with which two or more judges rate the work or performance of test takers; sometimes referred to as *inter-rater reliability.*

intervention planning The activity of a practitioner that involves the development of a treatment protocol.

inventory A questionnaire or checklist, usually in the form of a self-report, that elicits information about an individual's personal opinions, interests, attitudes, preferences, personality characteristics, motivations, and typical reactions to situations and problems.

item A statement, question, exercise, or task on a test for which the test taker is to select or construct a response, or perform a task. See *item prompt.*

item characteristic curve A mathematical function relating the probability of a certain item response, usually a correct response, to the level of the attribute measured by the item. Also called *item response curve*, or *item response function,* or *icc.*

item pool The aggregate of items from which a test or test scale's items are selected during test development, or the total set of items from which a particular test is selected for a test taker during adaptive testing.

item prompt The question, stimulus, or instructions that direct the efforts of examinees in formulating their responses to a constructed-response exercise.

item response theory (IRT) A mathematical model of the relationship between performance on a test item and the test taker's level of performance on a scale of the ability, trait, or proficiency being measured, usually denoted as θ. In the case of items scored 0 / 1 (incorrect/correct response) the model describes the relationship between θ and the item mean score (P) for test takers at level θ, over the range of permissible values of θ. In most applications, the mathematical function relating P to θ is assumed to be a logistic function that closely resembles the cumulative normal distribution.

job analysis A general term referring to the investigation of positions or job classes to obtain descriptive information about job duties and tasks, responsibilities, necessary worker characteristics (e.g. knowledge, skills, and abilities), working conditions, and/or other aspects of the work.

job performance measurement The measurement of an incumbent's performance of a job. This may include a job sample test, an assessment of job knowledge, and possibly ratings of the incumbent's actual performance on the job.

job sample test A test of the ability of an individual to perform the tasks of which the job is comprised.

licensing The granting, usually by a government agency, of an authorization or legal permission to practice an occupation or profession. See also *certification, credentialing.*

linkage The result of placing two ot more tests on the same scale, so that scores can be used interchangeably. Several linking methods are used: See *equating, calibration, moderation,* and *projection,* and *alternate forms.*

literature In this document, a term denoting accessible reports of research, such as books, articles published in professional journals, technical reports, and accessible versions of papers presented at professional meetings.

local evidence Evidence (usually related to reliability or validity) collected for a specific set of test takers in a single institution or at a specific location.

local norms Norms by which test scores are referred to a specific, limited *reference population* of particular interest to the test user (e.g., locale, organization, or institution); local norms are not intended as representative of populations beyond that setting.

local setting The organization or institution where a test is used.

low-stakes test A test used to provide results that have only minor or indirect consequences for examinees, programs, or institutions involved in the testing.

mandated tests Tests that are administered because of a mandate from an external authority.

mastery test 1. A criterion-referenced test designed to indicate the extent to which the test taker has mastered some domain of knowledge or skill. Mastery is generally indicated by attaining a passing score or cut score. **2.** In some technical use, a test designed to indicate whether a test taker has or has not attained a prescribed level of mastery of a domain. See *cut score, computer-based mastery test.*

matrix sampling A measurement format in which a large set of test items is organized into a number of relatively short item sets, each of which is randomly assigned to a subsample of test takers, thereby avoiding the need to administer all items to all examinees in a program evaluation.

meta-analysis A statistical method of research in which the results from several independent, comparable studies are combined to determine the size of an overall effect or the degree of relationship between two variables.

moderation In test linking, the term moderation, used without a modifier, usually signifies statistical moderation, which is the adjustment of the score scale of one test, usually by setting the mean and standard deviation of one set of test scores to be equal to the mean and standard deviation of another distribution of test scores.

moderator variable In regression analysis, a variable that serves to explain, at least in part, the correlation of two other variables.

modification See *test modification.*

neuropsychodiagnosis Classification or description of inferred central nervous system status on the basis of neuropsychological assessment.

neuropsychological assessment A specialized type of psychological assessment of normal or pathological processes affecting the central nervous system and the resulting psychological and behavioral functions or dysfunctions.

norm-referenced test interpretation A score interpretation based on a comparison of a test taker's performance to the performance of other people in a specified *reference population.* See *criterion-referenced test.*

normalized standard score A derived test score in which a numerical transformation has been chosen so that the score distribution closely approximates a normal distribution, for some specific population.

norms Statistics or tabular data that summarize the distribution of test performance for one or more specified groups, such as test takers of various ages or grades. Norms are usually designed to represent some larger population, such as test takers throughout the country. The group of examinees represented by the norms is referred to as the *reference population.*

operational use The actual use of a test, after inirial test development has been completed, to inform an interpretation, decision, or acrion based, in part, upon test scores.

outcome evaluation An evaluation of the efficacy of an intervention.

parallel forms See *alternate forms*.

percentile The score on a test below which a given percentage of scores fall.

percentile rank Most commonly, the percentage of scores in a specified distribution thar fall below the point at which a given score lies. Sometimes the percentage is defined to include scores that fall at the point; sometimes the percentage is defined to include half of the scores at the point.

performance assessments Product- and behavior-based measurements based on settings designed to emulate real-life contexts or conditions in which specific knowledge or skills are actually applied.

performance standard 1. An objective definition of a certain level of performance in some domain in rerms of a cut score or a range of scores on the score scale of a test measuring proficiency in that domain. 2. A statement or description of a set of operarional rasks exemplifying a level of performance associated with a more general content standard; the srarement may be used to guide judgments about the location of a cut score on a score scale. The rerm often implies a desired level of performance. See *cut score*.

personality inventory An invenrory that measures one or more characreristics thar are regarded generally as psychological attributes or interpersonal proclivities or skills.

pilot test A test administered to a sample of test takers to try out some aspects of rhe rest or test items, such as instructions, time limits, item response formats, or item response options. See *field test*.

policy The principles, plan, or procedures established by an agency, institution, organization, or government, generally with the intent of reaching a long-term goal.

portfolio In assessment, a systematic collection of educational or work products that have been compiled or accumulated over time, according to a specific set of principles.

precision of measurement A general term that refers to a measure's sensitivity to measurement error. See *standard error of measurement, error of measurement*.

practice analysis A general term referring to the investigation of a certain work position, or profession, to obtain descriptive information about the activities and responsibilities of the position and about the knowledge, skills, and abilities needed to engage in the work of the position. The concept is essentially the same as a job analysis but is generally preferred for professional occupations involving a great deal of individual decision making. See *job analysis*.

predictive bias The systematic under- or overpredicrion of criterion performance for people belonging to groups differentiated by characrer-istics not relevant to criterion performance.

predictive validity A term used in the 1974 *Standards* ro refer to a type of "criterion-relared validity" that applies "when one wishes ro infer from a test score an individual's most probable sranding on some other variable called a crire-rion" (p. 26). In the 1985 *Standards*, the rerm *criterion-related validity* was changed to *criteri-on-related evidence,* emphasizing that ir referred

to one type of evidence within a unitary conception of validity. The current document refers to "evidence based on relations to other variables" that include "test-criterion relationships." Predictive evidence indicates how accurately test data can predict criterion scores that are obtained at a later time.

program evaluation The collection and synthesis of systematic evidence about the use, operation, and effects of some planned set of procedures.

program norms See *user norms.*

projection In test scaling, a method of linking in which scores on one test (X) are used to predict scores on another test (Y). The projected Y score is the average Y score for all persons with a given X score. Like regression, the projection of test Y onto test X is different from the projection of test X onto test Y. See *linkage.*

proposed interpretation A summary, or a set of illustrations, of the intended meaning of test scores, based on the construct(s) or concept(s) the test is designed to measure.

protocol A record of events. A test protocol will usually consist of the test record and test scores.

psychodiagnosis Formalization or classification of functional mental health status based on psychological assessment. See *neuropsychodiagnosis.*

psychological assessment A comprehensive examination of psychological functioning that involves collecting, evaluating, and integrating test results and collateral information, and reporting information about an individual. Various methods may be used to acquire information during a psychological assessment: administering, scoring and interpreting tests and inventories; behavioral observation; client and third-party interviews; analysis of prior educational, occupational, medical, and psychological records.

psychological testing Any procedure that involves the use of tests or inventories to assess particular psychological characteristics of an individual.

random error An unsystematic error; a quantity (often observed indirectly) that appears to have no relationship to any other variable.

random sample See *sample.*

raw score The unadjusted score on a test, often determined by counting the number of correct answers, but more generally a sum or other combination of item scores. In item response theory, the estimate of test taker proficiency, usually symbolized $\hat{\theta}$, is analogous to a raw score although, unlike a raw score, its scaling is not arbitrary.

reference population The population of test takers represented by test norms. The sample on which the test norms are based must permit accurate estimation of the test score distribution for the reference population. The reference population may be defined in terms of examinee age, grade, or clinical status at time of testing, or other characteristics.

relative score interpretation The meaning of the test score for an individual, or the average score for a definable group, derived from the rank of the score or average within one or more reference distributions of scores. See *absolute score interpretation.*

reliability The degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable, and repeatable for an individual test taker; the degree to which scores are free of errors of measurement for a given group. See *generalizability theory.*

reliability coefficient A unit-free indicator that reflects the degree to which scotes are free of measurement error. The indicator resembles (or is) a product-moment correlation. In classical test theory, the term represents the ratio of true score variance to observed score variance fot a particular examinee population. The conditions under which the coefficient is estimated may involve variation in test forms, measurement occasions, raters, scorers, or clinicians, and may entail multiple examinee products or performances. These and othet variations in conditions give rise to qualifying adjectives, such as alternate-form reliability, internal consistency reliability, test-retest reliability, etc. See *generalizability theory.*

response bias A test taker's tendency to respond in a particular way or style to items on a test (i.e., acquiescence, social desirability, the tendency to choose 'true' on a true-false test) that yields systematic, construct-irrelevant error in test scores.

response process A component, usually hypothetical, of a cognitive account of some behavior, such as making an item response.

response protocol A record of the responses given by a test taker to a particulat test.

restriction of range or variability Reduction in the observed score variance of an examinee sample, compared to the variance of the entire examinee population, as a consequence of constraints on the process of sampling examinees. See *adjusted validity/reliability coefficient.*

rubric See *scoring rubric.*

sample A selection of a specified number of entities called sampling units (test takers, items, etc.) from a larger specified set of possible entities, called the population. A random sample is a selection according to a random process, with the selection of each entity in no way dependent on the selection of other entities. A stratified random sample is a set of random samples, each of a specified size, from several different sets, which are viewed as strata of the population.

scale **1**. The system of numbers, and their units, by which a value is reported on some dimension of measurement. Length can be reported in the English system of feet and inches or in the metric system of meters and centimeters. **2**. In testing, *scale* sometimes refers to the set of items or subtests used in the measurement and is distinguished from a test in the type of characteristic being measured. One speaks of a test of verbal ability, but a scale of extroversion-introversion.

scale score See *derived score.*

scaling The process of creating a scale or a scale score. Scaling may enhance test score interpretation by placing scores from different tests or test forms onto a common scale or by producing scale scores designed to support criterion-refetenced or notm-referenced scote interpretations. See *scale.*

score Any specific number resulting from the assessment of an individual; a generic term applied for convenience to such diverse measures as test scores, estimates of latent variables, production counts, absence recotds, course grades, ratings, and so forth.

scoring formula The formula by which the raw score on a test is obtained. The simplest scoring formula is "raw score equals number correct." Other formulas differentially weight item responses. For example, in an attempt to correct for guessing or nonresponse, zero weights may be assigned to nonresponses and negative weights to incorrect responses.

scoring rubric The established criteria, including rules, principles, and illustrations, used in scoring responses to individual items and clusters of items. The term usually refers to the scoring procedures for assessment tasks that do not provide enumerated responses from which test takers make a choice. Scoring rubrics vary in the degree of judgment entailed, in the number of distinct score levels defined, in the latitude given scorers for assigning intermediate or fractional score values, and in other ways.

screening test A test that is used to make broad categorizations of examinees as a first step in selection decisions or diagnostic processes.

security (of a test) See *test security*.

selection A purpose for testing that results in the acceptance or rejection of applicants for a particular educational or employment opportunity.

sensitivity In classification of disorders, the proportion of cases in which a disorder is detected when it is in fact present.

Spearman-Brown formula A formula derived within classical test theory that projects the reliability of a shortened or lengthened test from the reliability of a test of specified length.

specificity In classification of disorders, the proportion of cases for which a diagnosis of disorder is rejected when rejection is warranted.

speededness A test characteristic, dictated by the test's time limits, that results in a test taker's score being dependent on the rate at which work is performed as well as the correctness of the responses. The term is not used to describe tests of speed. Speededness is often an undesirable characteristic.

split-halves reliability coefficient An internal consistency coefficient obtained by using half the items on the test to yield one score and the other half of the items to yield a second, independent score. The correlation between the scores on these two half-tests, adjusted via the Spearman-Brown formula, provides an estimate of the alternate-form reliability of the total test.

stability The extent to which scores on a test are essentially invariant over time. Stability is an aspect of reliability and is assessed by correlating the test scores of a group of individuals with scores on the same test, or an equated test, taken by the same group at a later time.

standard error of measurement The standard deviation of an individual's observed scores from repeated administrations of a test (or parallel forms of a test) under identical conditions. Because such data cannot generally be collected, the standard error of measurement is usually estimated from group data. See *error of measurement*.

standard score A type of derived score such that the distribution of these scores for a specified population has convenient, known values for the mean and standard deviation. The term is sometimes used to signify a mean of 0.0 and a standard deviation of 1.0. See *derived score*.

standardization **1**. In test administration, maintaining a constant testing environment and conducting the test according to detailed rules and specifications, so that testing conditions are the same for all test takers. **2**. In test development, establishing scoring norms based on the test performance of a representative sample of individuals with which the test is intended to be used. **3**. In statistical analysis, transforming a variable so that its standard deviation is 1.0 for some specified population or sample. See *standard score*.

standards-based assessment Assessments intended to represent systematically described content and performance standards.

stratified coefficient alpha A modification of coefficient alpha that renders it appropriate for a multi-factor test by defining the total score as the composite of scores on single-factor part-tests.

stratified sample See *sample.*

systematic error A consistent score component (often observed indirectly), not related to the test performance. See *bias.*

technical manual A publication prepared by test authors and publishers to provide technical and psychometric information on a test.

test An evaluative device or procedure in which a sample of an examinee's behavior in a specified domain is obtained and subsequently evaluated and scored using a standardized process.

test developer The person(s) or agency responsible for the construction of a test and for the documentation regarding its technical quality for an intended purpose.

test development The process through which a test is planned, constructed, evaluated, and modified, including consideration of content, format, administration, scoring, item properties, scaling, and technical quality for its intended purpose.

test documents Publications such as test manuals, technical manuals, user's guides, specimen sets, and directions for test administrators and scorers that provide information for evaluating the appropriateness and technical adequacy of a test for its intended purpose.

test information function A mathematical function relating each level of an ability or latent trait, as defined under item response theory (IRT), to the reciprocal of the corresponding conditional measurement error variance.

test manual A publication prepared by test developers and publishers to provide information on test administration, scoring, and interpretation and to provide technical data on test characteristics. See *user's guide.*

test modification Changes made in the content, format, and/or administration procedure of a test in order to accommodate test takers who are unable to take the original test under standard test conditions.

test security Limiting access to the specific content of a test to those who need to know it for test development, test scoring, and test evaluation. In particular, test items on secure tests are not published; unauthorized copying is forbidden by any test taker or anyone otherwise associated with the test. A secure test is not for publication in any form, in any venue.

test specifications A detailed description for a test, often called a test blueprint, that specifies the number or proportion of items that assess each content and process/skill area; the format of items, responses, and scoring rubrics and procedures; and the desired psychometric properties of the items and test such as the distribution of item difficulty and discrimination indices.

test user The person(s) or agency responsible for the choice and administration of a test, for the interpretation of test scores produced in a given context, and for any decisions or actions that are based, in part, on test scores.

test-retest reliability A reliability coefficient obtained by administering the same test a second time to the same group after a time interval and correlating the two sets of scores.

timed tests A test administered to a test taker who is allotted a strictly prescribed amount of time to respond to the test.

top-down A method of selecting the best applicants according to some numerical scale of suitability. Often, "best" is taken to mean "highest scoring on some test."

translational equivalence The degree to which the translated version of a test is equivalent to the original test. Translational equivalence is typically examined in terms of the language used, the scores produced, and the constructs measured by the translated version and the original test. See *back translation.*

true score In classical test theory, the average of the scores that would be earned by an individual on an unlimited number of perfectly parallel forms of the same test. In item response theory, the error-free value of test taker proficiency, usually symbolized by θ.

unidimensional Having only one dimension, or only one latent variable.

user norms Descriptive statistics (including percentile ranks) for a sample of test takers that does not represent a well-defined reference population, for example, all persons tested during a certain period of time, or a set of self-selected test takers. Also called program norms. See *norms.*

user's guide A publication prepared by the test authors and publishers to provide information on a test's purpose, appropriate uses, proper administration, scoring procedures, normative data, interpretation of results, and case studies. See *test manual.*

validation The process through which the validity of the proposed interpretation of test scores is investigated.

validity The degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test.

validity argument An explicit scientific justification of the degree to which accumulated evidence and theory support the proposed interpretation(s) of test scores.

validity generalization Applying validity evidence obtained in one or more situations to other similar situations on the basis of simultaneous estimation, meta-analysis, or synthetic validation arguments.

variance components In testing, variances accruing from the separate constituent sources that are assumed to contribute to the overall variance of observed scores. Such variances, estimated by methods of the analysis of variance, often reflect situation, location, time, test form, rater, and related effects.

vocational assessment A specialized type of psychological assessment designed to generate hypotheses and inferences about interests, work needs and values, career development, vocational maturity, and indecision.

weighted scoring A method of scoring a test in which the number of points awarded for a correct (or diagnostically relevant) response is not the same for all items in the test. In some cases, the scoring formula awards more points for one response to an item than for another.

INDEX

Numbers in this index refer to specific standard(s).


ISBN-13: 978-0-935302-25-7
90000
9 780935 302257