PART I

# I Foundations

# 1. VALIDITY

## BACKGROUND

Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores for proposed uses that are evaluated, not the test itself. When test scores are interpreted in more than one way (e.g., both to describe a test taker's current level of the attribute being measured and to make a prediction about a future outcome), each intended interpretation must be validated. Statements about validity should refer to particular interpretations for specified uses. It is incorrect to use the unqualified phrase "the validity of the test."

Evidence of the validity of a given interpretation of test scores for a specified use is a necessary condition for the justifiable use of the test. Where sufficient evidence of validity exists, the decision as to whether to actually administer a particular test generally takes additional considerations into account. These include cost-benefit considerations, framed in different subdisciplines as utility analysis or as consideration of negative consequences of test use, and a weighing of any negative consequences against the positive consequences of test use.

Validation logically begins with an explicit statement of the proposed interpretation of test scores, along with a rationale for the relevance of the interpretation to the proposed use. The proposed interpretation includes specifying the construct the test is intended to measure. The term *construct* is used in the *Standards* to refer to the concept or characteristic that a test is designed to measure. Rarely, if ever, is there a single possible meaning that can be attached to a test score or a pattern of test responses. Thus, it is always incumbent on test developers and users to specify the construct interpretation that will be made on the basis of the score or response pattern.

Examples of constructs currently used in assessment include mathematics achievement, general cognitive ability, racial identity attitudes, depression, and self-esteem. To support test development, the proposed construct interpretation is elaborated by describing its scope and extent and by delineating the aspects of the construct that are to be represented. The detailed description provides a conceptual framework for the test, delineating the knowledge, skills, abilities, traits, interests, processes, competencies, or characteristics to be assessed. Ideally, the framework indicates how the construct as represented is to be distinguished from other constructs and how it should relate to other variables.

The conceptual framework is partially shaped by the ways in which test scores will be used. For instance, a test of mathematics achievement might be used to place a student in an appropriate program of instruction, to endorse a high school diploma, or to inform a college admissions decision. Each of these uses implies a somewhat different interpretation of the mathematics achievement test scores: that a student will benefit from a particular instructional intervention, that a student has mastered a specified curriculum, or that a student is likely to be successful with college-level work. Similarly, a test of conscientiousness might be used for psychological counseling, to inform a decision about employment, or for the basic scientific purpose of elaborating the construct of conscientiousness. Each of these potential uses shapes the specified framework and the proposed interpretation of the test's scores and also can have implications for test development and evaluation. Validation can be viewed as a process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use. The conceptual framework points to the kinds of

evidence that might be collected to evaluate the proposed interpretation in light of the purposes of testing. As validation proceeds, and new evidence regarding the interpretations that can and cannot be drawn from test scores becomes available, revisions may be needed in the test, in the conceptual framework that shapes it, and even in the construct underlying the test.

The wide variety of tests and circumstances makes it natural that some types of evidence will be especially critical in a given case, whereas other types will be less useful. Decisions about what types of evidence are important for the validation argument in each instance can be clarified by developing a set of propositions or claims that support the proposed interpretation for the particular purpose of testing. For instance, when a mathematics achievement test is used to assess readiness for an advanced course, evidence for the following propositions might be relevant: (a) that certain skills are prerequisite for the advanced course; (b) that the content domain of the test is consistent with these prerequisite skills; (c) that test scores can be generalized across relevant sets of items; (d) that test scores are not unduly influenced by ancillary variables, such as writing ability; (e) that success in the advanced course can be validly assessed; and (f) that test takers with high scores on the test will be more successful in the advanced course than test takers with low scores on the test. Examples of propositions in other testing contexts might include, for instance, the proposition that test takers with high general anxiety scores experience significant anxiety in a range of settings, the proposition that a child's score on an intelligence scale is strongly related to the child's academic performance, or the proposition that a certain pattern of scores on a neuropsychological battery indicates impairment that is characteristic of brain injury. The validation process evolves as these propositions are articulated and evidence is gathered to evaluate their soundness.

Identifying the propositions implied by a proposed test interpretation can be facilitated by considering rival hypotheses that may challenge the proposed interpretation. It is also useful to consider the perspectives of different interested parties, existing experience with similar tests and contexts, and the expected consequences of the proposed test use. A finding of unintended consequences of test use may also prompt a consideration of rival hypotheses. Plausible rival hypotheses can often be generated by considering whether a test measures less or more than its proposed construct. Such considerations are referred to as *construct underrepresentation* (or *construct deficiency*) and *construct-irrelevant variance* (or *construct contamination*), respectively.

Construct underrepresentation refers to the degree to which a test fails to capture important aspects of the construct. It implies a narrowed meaning of test scores because the test does not adequately sample some types of content, engage some psychological processes, or elicit some ways of responding that are encompassed by the intended construct. Take, for example, a test intended as a comprehensive measure of anxiety. A particular test might underrepresent the intended construct because it measures only physiological reactions and not emotional, cognitive, or situational components. As another example, a test of reading comprehension intended to measure children's ability to read and interpret stories with understanding might not contain a sufficient variety of reading passages or might ignore a common type of reading material.

Construct-irrelevance refers to the degree to which test scores are affected by processes that are extraneous to the test's intended purpose. The test scores may be systematically influenced to some extent by processes that are not part of the construct. In the case of a reading comprehension test, these might include material too far above or below the level intended to be tested, an emotional reaction to the test content, familiarity with the subject matter of the reading passages on the test, or the writing skill needed to compose a response. Depending on the detailed definition of the construct, vocabulary knowledge or reading speed might also be irrelevant components. On a test designed to measure anxiety, a response bias to underreport one's anxiety might be considered a source of construct-irrelevant variance. In the case

of a mathematics test, it might include overreliance on reading comprehension skills that English language learners may be lacking. On a test designed to measure science knowledge, test-taker internalizing of gender-based stereotypes about women in the sciences might be a source of construct-irrelevant variance.

Nearly all tests leave out elements that some potential users believe should be measured and include some elements that some potential users consider inappropriate. Validation involves careful attention to possible distortions in meaning arising from inadequate representation of the construct and also to aspects of measurement, such as test format, administration conditions, or language level, that may materially limit or qualify the interpretation of test scores for various groups of test takers. That is, the process of validation may lead to revisions in the test, in the conceptual framework of the test, or both. Interpretations drawn from the revised test would again need validation.

When propositions have been identified that would support the proposed interpretation of test scores, one can proceed with validation by obtaining empirical evidence, examining relevant literature, and/or conducting logical analyses to evaluate each of the propositions. Empirical evidence may include both local evidence, produced within the contexts where the test will be used, and evidence from similar testing applications in other settings. Use of existing evidence from similar tests and contexts can enhance the quality of the validity argument, especially when data for the test and context in question are limited.

Because an interpretation for a given use typically depends on more than one proposition, strong evidence in support of one part of the interpretation in no way diminishes the need for evidence to support other parts of the interpretation. For example, when an employment test is being considered for selection, a strong predictor-criterion relationship in an employment setting is ordinarily not sufficient to justify use of the test. One should also consider the appropriateness and meaningfulness of the criterion measure, the appropriateness of the testing materials and procedures for the full range of applicants, and the consistency of the support for the proposed interpretation across groups. Professional judgment guides decisions regarding the specific forms of evidence that can best support the intended interpretation for a specified use. As in all scientific endeavors, the quality of the evidence is paramount. A few pieces of solid evidence regarding a particular proposition are better than numerous pieces of evidence of questionable quality. The determination that a given test interpretation for a specific purpose is warranted is based on professional judgment that the preponderance of the available evidence supports that interpretation. The quality and quantity of evidence sufficient to reach this judgment may differ for test uses depending on the stakes involved in the testing. A given interpretation may not be warranted either as a result of insufficient evidence in support of it or as a result of credible evidence against it.

Validation is the joint responsibility of the test developer and the test user. The test developer is responsible for furnishing relevant evidence and a rationale in support of any test score interpretations for specified uses intended by the developer. The test user is ultimately responsible for evaluating the evidence in the particular setting in which the test is to be used. When a test user proposes an interpretation or use of test scores that differs from those supported by the test developer, the responsibility for providing validity evidence in support of that interpretation for the specified use is the responsibility of the user. It should be noted that important contributions to the validity evidence may be made as other researchers report findings of investigations that are related to the meaning of scores on the test.

## Sources of Validity Evidence

The following sections outline various sources of evidence that might be used in evaluating the validity of a proposed interpretation of test scores for a particular use. These sources of evidence may illuminate different aspects of validity,

but they do not represent distinct types of validity. Validity is a unitary concept. It is the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed use. Like the 1999 *Standards*, this edition refers to types of validity evidence, rather than distinct types of validity. To emphasize this distinction, the treatment that follows does not follow historical nomenclature (i.e., the use of the terms *content validity* or *predictive validity*).

As the discussion in the prior section emphasizes, each type of evidence presented below is not required in all settings. Rather, support is needed for each proposition that underlies a proposed test interpretation for a specified use. A proposition that a test is predictive of a given criterion can be supported without evidence that the test samples a particular content domain. In contrast, a proposition that a test covers a representative sample of a particular curriculum may be supported without evidence that the test predicts a given criterion. However, a more complex set of propositions, e.g., that a test samples a specified domain and thus is predictive of a criterion reflecting a related domain, will require evidence supporting both parts of this set of propositions. Tests developers are also expected to make the case that the scores are not unduly influenced by construct-irrelevant variance (see chap. 3 for detailed treatment of issues related to construct-irrelevant variance). In general, adequate support for proposed interpretations for specific uses will require multiple sources of evidence.

The position developed above also underscores the fact that if a given test is interpreted in multiple ways for multiple uses, the propositions underlying these interpretations for different uses also are likely to differ. Support is needed for the propositions underlying each interpretation for a specific use. Evidence supporting the interpretation of scores on a mathematics achievement test for placing students in subsequent courses (i.e., evidence that the test interpretation is valid for its intended purpose) does not permit inferring validity for other purposes (e.g., promotion or teacher evaluation).

## Evidence Based on Test Content

Important validity evidence can be obtained from an analysis of the relationship between the content of a test and the construct it is intended to measure. Test content refers to the themes, wording, and format of the items, tasks, or questions on a test. Administration and scoring may also be relevant to content-based evidence. Test developers often work from a specification of the content domain. The content specification carefully describes the content in detail, often with a classification of areas of content and types of items. Evidence based on test content can include logical or empirical analyses of the adequacy with which the test content represents the content domain and of the relevance of the content domain to the proposed interpretation of test scores. Evidence based on content can also come from expert judgments of the relationship between parts of the test and the construct. For example, in developing a licensure test, the major facets that are relevant to the purpose for which the occupation is regulated can be specified, and experts in that occupation can be asked to assign test items to the categories defined by those facets. These or other experts can then judge the representativeness of the chosen set of items.

Some tests are based on systematic observations of behavior. For example, a list of the tasks constituting a job domain may be developed from observations of behavior in a job, together with judgments of subject matter experts. Expert judgments can be used to assess the relative importance, criticality, and/or frequency of the various tasks. A job sample test can then be constructed from a random or stratified sampling of tasks rated highly on these characteristics. The test can then be administered under standardized conditions in an off-the-job setting.

The appropriateness of a given content domain is related to the specific inferences to be made from test scores. Thus, when considering an available test for a purpose other than that for which it was first developed, it is especially important to evaluate the appropriateness of the original content domain for the proposed new

purpose. For example, a test given for research purposes to compare student achievement across states in a given domain may properly also cover material that receives little or no attention in the curriculum. Policy makers can then evaluate student achievement with respect to both content neglected and content addressed. On the other hand, when student mastery of a delivered curriculum is tested for purposes of informing decisions about individual students, such as promotion or graduation, the framework elaborating a content domain is appropriately limited to what students have had an opportunity to learn from the curriculum as delivered.

Evidence about content can be used, in part, to address questions about differences in the meaning or interpretation of test scores across relevant subgroups of test takers. Of particular concern is the extent to which construct underrepresentation or construct-irrelevance may give an unfair advantage or disadvantage to one or more subgroups of test takers. For example, in an employment test, the use of vocabulary more complex than needed on the job may be a source of construct-irrelevant variance for English language learners or others. Careful review of the construct and test content domain by a diverse panel of experts may point to potential sources of irrelevant difficulty (or easiness) that require further investigation.

Content-oriented evidence of validation is at the heart of the process in the educational arena known as *alignment*, which involves evaluating the correspondence between student learning standards and test content. Content-sampling issues in the alignment process include evaluating whether test content appropriately samples the domain set forward in curriculum standards, whether the cognitive demands of test items correspond to the level reflected in the student learning standards (e.g., content standards), and whether the test avoids the inclusion of features irrelevant to the standard that is the intended target of each test item.

### Evidence Based on Response Processes

Some construct interpretations involve more or less explicit assumptions about the cognitive processes engaged in by test takers. Theoretical and empirical analyses of the response processes of test takers can provide evidence concerning the fit between the construct and the detailed nature of the performance or response actually engaged in by test takers. For instance, if a test is intended to assess mathematical reasoning, it becomes important to determine whether test takers are, in fact, reasoning about the material given instead of following a standard algorithm applicable only to the specific items on the test.

Evidence based on response processes generally comes from analyses of individual responses. Questioning test takers from various groups making up the intended test-taking population about their performance strategies or responses to particular items can yield evidence that enriches the definition of a construct. Maintaining records that monitor the development of a response to a writing task, through successive written drafts or electronically monitored revisions, for instance, also provides evidence of process. Documentation of other aspects of performance, like eye movements or response times, may also be relevant to some constructs. Inferences about processes involved in performance can also be developed by analyzing the relationship among parts of the test and between the test and other variables. Wide individual differences in process can be revealing and may lead to reconsideration of certain test formats.

Evidence of response processes can contribute to answering questions about differences in meaning or interpretation of test scores across relevant subgroups of test takers. Process studies involving test takers from different subgroups can assist in determining the extent to which capabilities irrelevant or ancillary to the construct may be differentially influencing test takers' test performance.

Studies of response processes are not limited to the test taker. Assessments often rely on observers or judges to record and/or evaluate test takers' performances or products. In such cases, relevant validity evidence includes the extent to which the processes of observers or judges are consistent with the intended interpretation of scores. For instance, if judges are expected to apply particular criteria in scoring test takers' performances, it is

important to ascertain whether they are, in fact, applying the appropriate criteria and not being influenced by factors that are irrelevant to the intended interpretation (e.g., quality of handwriting is irrelevant to judging the content of an written essay). Thus, validation may include empirical studies of how observers or judges record and evaluate data along with analyses of the appropriateness of these processes to the intended interpretation or construct definition.

While evidence about response processes may be central in settings where explicit claims about response processes are made by test developers or where inferences about responses are made by test users, there are many other cases where claims about response processes are not part of the validity argument. In some cases, multiple response processes are available for solving the problems of interest, and the construct of interest is only concerned with whether the problem was solved correctly. As a simple example, there may be multiple possible routes to obtaining the correct solution to a mathematical problem.

### Evidence Based on Internal Structure

Analyses of the internal structure of a test can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based. The conceptual framework for a test may imply a single dimension of behavior, or it may posit several components that are each expected to be homogeneous, but that are also distinct from each other. For example, a measure of discomfort on a health survey might assess both physical and emotional health. The extent to which item interrelationships bear out the presumptions of the framework would be relevant to validity.

The specific types of analyses and their interpretation depend on how the test will be used. For example, if a particular application posited a series of increasingly difficult test components, empirical evidence of the extent to which response patterns conformed to this expectation would be provided. A theory that posited unidimensionality would call for evidence of item homogeneity. In this case, the number of items and item interrelationships form the basis for an estimate of score reliability, but such an index would be inappropriate for tests with a more complex internal structure.

Some studies of the internal structure of tests are designed to show whether particular items may function differently for identifiable subgroups of test takers (e.g., racial/ethnic or gender subgroups.) *Differential item functioning* occurs when different groups of test takers with similar overall ability, or similar status on an appropriate criterion, have, on average, systematically different responses to a particular item. This issue is discussed in chapter 3. However, differential item functioning is not always a flaw or weakness. Subsets of items that have a specific characteristic in common (e.g., specific content, task representation) may function differently for different groups of similarly scoring test takers. This indicates a kind of multidimensionality that may be unexpected or may conform to the test framework.

### Evidence Based on Relations to Other Variables

In many cases, the intended interpretation for a given use implies that the construct should be related to some other variables, and, as a result, analyses of the relationship of test scores to variables external to the test provide another important source of validity evidence. External variables may include measures of some criteria that the test is expected to predict, as well as relationships to other tests hypothesized to measure the same constructs, and tests measuring related or different constructs. Measures other than test scores, such as performance criteria, are often used in employment settings. Categorical variables, including group membership variables, become relevant when the theory underlying a proposed test use suggests that group differences should be present or absent if a proposed test score interpretation is to be supported. Evidence based on relationships with other variables provides evidence about the degree to which these relationships are consistent with the construct underlying the proposed test score interpretations.

**Convergent and discriminant evidence.** Relationships between test scores and other measures

intended to assess the same or similar constructs provide convergent evidence, whereas relationships between test scores and measures purportedly of different constructs provide discriminant evidence. For instance, within some theoretical frameworks, scores on a multiple-choice test of reading comprehension might be expected to relate closely (convergent evidence) to other measures of reading comprehension based on other methods, such as essay responses. Conversely, test scores might be expected to relate less closely (discriminant evidence) to measures of other skills, such as logical reasoning. Relationships among different methods of measuring the construct can be especially helpful in sharpening and elaborating score meaning and interpretation.

Evidence of relations with other variables can involve experimental as well as correlational evidence. Studies might be designed, for instance, to investigate whether scores on a measure of anxiety improve as a result of some psychological treatment or whether scores on a test of academic achievement differentiate between instructed and noninstructed groups. If performance increases due to short-term coaching are viewed as a threat to validity, it would be useful to investigate whether coached and uncoached groups perform differently.

**Test-criterion relationships.** Evidence of the relation of test scores to a relevant criterion may be expressed in various ways, but the fundamental question is always, how accurately do test scores predict criterion performance? The degree of accuracy and the score range within which accuracy is needed depends on the purpose for which the test is used.

The criterion variable is a measure of some attribute or outcome that is operationally distinct from the test. Thus, the test is not a measure of a criterion, but rather is a measure hypothesized as a potential predictor of that targeted criterion. Whether a test predicts a given criterion in a given context is a testable hypothesis. The criteria that are of interest are determined by test users, for example administrators in a school system or managers of a firm. The choice of the criterion and the measurement procedures used to obtain criterion scores are of central importance. The credibility of a test-criterion study depends on the relevance, reliability, and validity of the interpretation based on the criterion measure for a given testing application.

Historically, two designs, often called predictive and concurrent, have been distinguished for evaluating test-criterion relationships. A predictive study indicates the strength of the relationship between test scores and criterion scores that are obtained at a later time. A concurrent study obtains test scores and criterion information at about the same time. When prediction is actually contemplated, as in academic admission or employment settings, or in planning rehabilitation regimens, predictive studies can retain the temporal differences and other characteristics of the practical situation. Concurrent evidence, which avoids temporal changes, is particularly useful for psychodiagnostic tests or in investigating alternative measures of some specified construct for which an accepted measurement procedure already exists. The choice of a predictive or concurrent research strategy in a given domain is also usefully informed by prior research evidence regarding the extent to which predictive and concurrent studies in that domain yield the same or different results.

Test scores are sometimes used in allocating individuals to different treatments in a way that is advantageous for the institution and/or for the individuals. Examples would include assigning individuals to different jobs within an organization, or determining whether to place a given student in a remedial class or a regular class. In that context, evidence is needed to judge the suitability of using a test when classifying or assigning a person to one job versus another or to one treatment versus another. Support for the validity of the classification procedure is provided by showing that the test is useful in determining which persons are likely to profit differentially from one treatment or another. It is possible for tests to be highly predictive of performance for different education programs or jobs without providing the information necessary to make a comparative judgment of the efficacy of assignments or treatments. In general, decision rules for selection

or placement are also influenced by the number of persons to be accepted or the numbers that can be accommodated in alternative placement categories (see chap. 11).

Evidence about relations to other variables is also used to investigate questions of differential prediction for subgroups. For instance, a finding that the relation of test scores to a relevant criterion variable differs from one subgroup to another may imply that the meaning of the scores is not the same for members of the different groups, perhaps due to construct underrepresentation or construct-irrelevant sources of variance. However, the difference may also imply that the criterion has different meaning for different groups. The differences in test-criterion relationships can also arise from measurement error, especially when group means differ, so such differences do not necessarily indicate differences in score meaning. See the discussion of fairness in chapter 3 for more extended consideration of possible courses of action when scores have different meanings for different groups.

**Validity generalization.** An important issue in educational and employment settings is the degree to which validity evidence based on test-criterion relations can be generalized to a new situation without further study of validity in that new situation. When a test is used to predict the same or similar criteria (e.g., performance of a given job) at different times or in different places, it is typically found that observed test-criterion correlations vary substantially. In the past, this has been taken to imply that local validation studies are always required. More recently, a variety of approaches to generalizing evidence from other settings has been developed, with meta-analysis the most widely used in the published literature. In particular, meta-analyses have shown that in some domains, much of this variability may be due to statistical artifacts such as sampling fluctuations and variations across validation studies in the ranges of test scores and in the reliability of criterion measures. When these and other influences are taken into account, it may be found that the remaining variability in validity coefficients is relatively small. Thus, statistical summaries of past validation studies in similar situations may be useful in estimating test-criterion relationships in a new situation. This practice is referred to as the study of validity generalization.

In some circumstances, there is a strong basis for using validity generalization. This would be the case where the meta-analytic database is large, where the meta-analytic data adequately represent the type of situation to which one wishes to generalize, and where correction for statistical artifacts produces a clear and consistent pattern of validity evidence. In such circumstances, the informational value of a local validity study may be relatively limited if not actually misleading, especially if its sample size is small. In other circumstances, the inferential leap required for generalization may be much larger. The meta-analytic database may be small, the findings may be less consistent, or the new situation may involve features markedly different from those represented in the meta-analytic database. In such circumstances, situation-specific validity evidence will be relatively more informative. Although research on validity generalization shows that results of a single local validation study may be quite imprecise, there are situations where a single study, carefully done, with adequate sample size, provides sufficient evidence to support or reject test use in a new situation. This highlights the importance of examining carefully the comparative informational value of local versus meta-analytic studies.

In conducting studies of the generalizability of validity evidence, the prior studies that are included may vary according to several situational facets. Some of the major facets are (a) differences in the way the predictor construct is measured, (b) the type of job or curriculum involved, (c) the type of criterion measure used, (d) the type of test takers, and (e) the time period in which the study was conducted. In any particular study of validity generalization, any number of these facets might vary, and a major objective of the study is to determine empirically the extent to which variation in these facets affects the test-criterion correlations obtained.

The extent to which predictive or concurrent validity evidence can be generalized to new situations is in large measure a function of accumulated research. Although evidence of generalization can often help to support a claim of validity in a new situation, the extent of available data limits the degree to which the claim can be sustained.

The above discussion focuses on the use of cumulative databases to estimate predictor-criterion relationships. Meta-analytic techniques can also be used to summarize other forms of data relevant to other inferences one may wish to draw from test scores in a particular application, such as effects of coaching and effects of certain alterations in testing conditions for test takers with specified disabilities. Gathering evidence about how well validity findings can be generalized across groups of test takers is an important part of the validation process. When the evidence suggests that inferences from test scores can be drawn for some subgroups but not for others, pursuing options such as those discussed in chapter 3 can reduce the risk of unfair test use.

### Evidence for Validity and Consequences of Testing

Some consequences of test use follow directly from the interpretation of test scores for uses intended by the test developer. The validation process involves gathering evidence to evaluate the soundness of these proposed interpretations for their intended uses.

Other consequences may also be part of a claim that extends beyond the interpretation or use of scores intended by the test developer. For example, a test of student achievement might provide data for a system intended to identify and improve lower-performing schools. The claim that testing results, used this way, will result in improved student learning may rest on propositions about the system or intervention itself, beyond propositions based on the meaning of the test itself. Consequences may point to the need for evidence about components of the system that will go beyond the interpretation of test scores as a valid measure of student achievement.

Still other consequences are unintended, and are often negative. For example, school district or statewide educational testing on selected subjects may lead teachers to focus on those subjects at the expense of others. As another example, a test developed to measure knowledge needed for a given job may result in lower passing rates for one group than for another. Unintended consequences merit close examination. While not all consequences can be anticipated, in some cases factors such as prior experiences in other settings offer a basis for anticipating and proactively addressing unintended consequences. See chapter 12 for additional examples from educational settings. In some cases, actions to address one consequence bring about other consequences. One example involves the notion of "missed opportunities," as in the case of moving to computerized scoring of student essays to increase grading consistency, thus forgoing the educational benefits of addressing the same problem by training teachers to grade more consistently.

These types of consideration of consequences of testing are discussed further below.

**Interpretation and uses of test scores intended by test developers.** Tests are commonly administered in the expectation that some benefit will be realized from the interpretation and use of the scores intended by the test developers. A few of the many possible benefits that might be claimed are selection of efficacious therapies, placement of workers in suitable jobs, prevention of unqualified individuals from entering a profession, or improvement of classroom instructional practices. A fundamental purpose of validation is to indicate whether these specific benefits are likely to be realized. Thus, in the case of a test used in placement decisions, the validation would be informed by evidence that alternative placements, in fact, are differentially beneficial to the persons and the institution. In the case of employment testing, if a test publisher asserts that use of the test will result in reduced employee training costs, improved workforce efficiency, or some other benefit, then the validation would be informed by evidence in support of that proposition.

It is important to note that the validity of test score interpretations depends not only on the uses

of the test scores but specifically on the claims that underlie the theory of action for these uses. For example, consider a school district that wants to determine children's readiness for kindergarten, and so administers a test battery and screens out students with low scores. If higher scores do, in fact, predict higher performance on key kindergarten tasks, the claim that use of the test scores for screening results in higher performance on these key tasks is supported and the interpretation of the test scores as a predictor of kindergarten readiness would be valid. If, however, the claim were made that use of the test scores for screening would result in the greatest benefit to students, the interpretation of test scores as indicators of readiness for kindergarten might not be valid because students with low scores might actually benefit more from access to kindergarten. In this case, different evidence is needed to support different claims that might be made about the same use of the screening test (for example, evidence that students below a certain cut score benefit more from another assignment than from assignment to kindergarten). The test developer is responsible for the validation of the interpretation that the test scores assess the indicated readiness skills. The school district is responsible for the validation of the proper interpretation of the readiness test scores and for evaluation of the policy of using the readiness test for placement/admissions decisions.

**Claims made about test use that are not directly based on test score interpretations.** Claims are sometimes made for benefits of testing that go beyond the direct interpretations or uses of the test scores themselves that are specified by the test developers. Educational tests, for example, may be advocated on the grounds that their use will improve student motivation to learn or encourage changes in classroom instructional practices by holding educators accountable for valued learning outcomes. Where such claims are central to the rationale advanced for testing, the direct examination of testing consequences necessarily assumes even greater importance. Those making the claims are responsible for evaluation of the claims. In some cases, such information can be drawn from existing data collected for purposes other than test validation; in other cases new information will be needed to address the impact of the testing program.

**Consequences that are unintended.** Test score interpretation for a given use may result in unintended consequences. A key distinction is between consequences that result from a source of error in the intended test score interpretation for a given use and consequences that do not result from error in test score interpretation. Examples of each are given below.

As discussed at some length in chapter 3, one domain in which unintended negative consequences of test use are at times observed involves test score differences for groups defined in terms of race/ethnicity, gender, age, and other characteristics. In such cases, however, it is important to distinguish between evidence that is directly relevant to validity and evidence that may inform decisions about social policy but falls outside the realm of validity. For example, concerns have been raised about the effect of group differences in test scores on employment selection and promotion, the placement of children in special education classes, and the narrowing of a school's curriculum to exclude learning objectives that are not assessed. Although information about the consequences of testing may influence decisions about test use, such consequences do not, in and of themselves, detract from the validity of intended interpretations of the test scores. Rather, judgments of validity or invalidity in the light of testing consequences depend on a more searching inquiry into the sources of those consequences.

Take, as an example, a finding of different hiring rates for members of different groups as a consequence of using an employment test. If the difference is due solely to an unequal distribution of the skills the test purports to measure, and if those skills are, in fact, important contributors to job performance, then the finding of group differences per se does not imply any lack of validity for the intended interpretation. If, however, the test measured skill differences unrelated to job performance (e.g., a sophisticated reading test for

a job that required only minimal functional literacy), or if the differences were due to the test's sensitivity to some test-taker characteristic not intended to be part of the test construct, then the intended interpretation of test scores as predicting job performance in a comparable manner for all groups of applicants would be rendered invalid, even if test scores correlated positively with some measure of job performance. If a test covers most of the relevant content domain but omits some areas, the content coverage might be judged adequate for some purposes. However, if it is found that excluding some components that could readily be assessed has a noticeable impact on selection rates for groups of interest (e.g., subgroup differences are found to be smaller on excluded components than on included components), the intended interpretation of test scores as predicting job performance in a comparable manner for all groups of applicants would be rendered invalid. Thus, evidence about consequences is relevant to validity when it can be traced to a source of invalidity such as construct underrepresentation or construct-irrelevant components. Evidence about consequences that cannot be so traced is not relevant to the validity of the intended interpretations of the test scores.

As another example, consider the case where research supports an employer's use of a particular test in the personality domain (i.e., the test proves to be predictive of an aspect of subsequent job performance), but it is found that some applicants form a negative opinion of the organization due to the perception that the test invades personal privacy. Thus, there is an unintended negative consequence of test use, but one that is not due to a flaw in the intended interpretation of test scores as predicting subsequent performance. Some employers faced with this situation may conclude that this negative consequence is grounds for discontinuing test use; others may conclude that the benefits gained by screening applicants outweigh this negative consequence. As this example illustrates, a consideration of consequences can influence a decision about test use, even though the consequence is independent of the validity of the intended test score interpretation. The example also illustrates that different decision makers may make different value judgments about the impact of consequences on test use.

The fact that the validity evidence supports the intended interpretation of test scores for use in applicant screening does not mean that test use is thus required: Issues other than validity, including legal constraints, can play an important and, in some cases, a determinative role in decisions about test use. Legal constraints may also limit an employer's discretion to discard test scores from tests that have already been administered, when that decision is based on differences in scores for subgroups of different races, ethnicities, or genders.

Note that unintended consequences can also be positive. Reversing the above example of test takers who form a negative impression of an organization based on the use of a particular test, a different test may be viewed favorably by applicants, leading to a positive impression of the organization. A given test use may result in multiple consequences, some positive and some negative.

In short, decisions about test use are appropriately informed by validity evidence about intended test score interpretations for a given use, by evidence evaluating additional claims about consequences of test use that do not follow directly from test score interpretations, and by value judgments about unintended positive and negative consequences of test use.

## Integrating the Validity Evidence

A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses. It encompasses evidence gathered from new studies and evidence available from earlier reported research. The validity argument may indicate the need for refining the definition of the construct, may suggest revisions in the test or other aspects of the testing process, and may indicate areas needing further study.

It is commonly observed that the validation process never ends, as there is always additional information that can be gathered to more fully

understand a test and the inferences that can be drawn from it. In this way an inference of validity is similar to any scientific inference. However, a test interpretation for a given use rests on evidence for a set of propositions making up the validity argument, and at some point validation evidence allows for a summary judgment of the intended interpretation that is well supported and defensible. At some point the effort to provide sufficient validity evidence to support a given test interpretation for a specific use does end (at least provisionally, pending the emergence of a strong basis for questioning that judgment). Legal requirements may necessitate that the validation study be updated in light of such factors as changes in the test population or newly developed alternative testing methods.

The amount and character of evidence required to support a provisional judgment of validity often vary between areas and also within an area as research on a topic advances. For example, prevailing standards of evidence may vary with the stakes involved in the use or interpretation of the test scores. Higher stakes may entail higher standards of evidence. As another example, in areas where data collection comes at a greater cost, one may find it necessary to base interpretations on fewer data than in areas where data collection comes with less cost.

Ultimately, the validity of an intended interpretation of test scores relies on all the available evidence relevant to the technical quality of a testing system. Different components of validity evidence are described in subsequent chapters of the *Standards,* and include evidence of careful test construction; adequate score reliability; appropriate test administration and scoring; accurate score scaling, equating, and standard setting; and careful attention to fairness for all test takers, as appropriate to the test interpretation in question.

# STANDARDS FOR VALIDITY

The standards in this chapter begin with an overarching standard (numbered 1.0), which is designed to convey the central intent or primary focus of the chapter. The overarching standard may also be viewed as the guiding principle of the chapter, and is applicable to all tests and test users. All subsequent standards have been separated into three thematic clusters labeled as follows:

1. Establishing Intended Uses and Interpretations
2. Issues Regarding Samples and Settings Used in Validation
3. Specific Forms of Validity Evidence

### Standard 1.0

**Clear articulation of each intended test score interpretation for a specified use should be set forth, and appropriate validity evidence in support of each intended interpretation should be provided.**

## Cluster 1. Establishing Intended Uses and Interpretations

### Standard 1.1

**The test developer should set forth clearly how test scores are intended to be interpreted and consequently used. The population(s) for which a test is intended should be delimited clearly, and the construct or constructs that the test is intended to assess should be described clearly.**

**Comment:** Statements about validity should refer to particular interpretations and consequent uses. It is incorrect to use the unqualified phrase "the validity of the test." No test permits interpretations that are valid for all purposes or in all situations. Each recommended interpretation for a given use requires validation. The test developer should specify in clear language the population for which the test is intended, the construct it is intended to measure, the contexts in which test scores are to be employed, and the processes by which the test is to be administered and scored.

### Standard 1.2

**A rationale should be presented for each intended interpretation of test scores for a given use, together with a summary of the evidence and theory bearing on the intended interpretation.**

**Comment:** The rationale should indicate what propositions are necessary to investigate the intended interpretation. The summary should combine logical analysis with empirical evidence to provide support for the test rationale. Evidence may come from studies conducted locally, in the setting where the test is to be used; from specific prior studies; or from comprehensive statistical syntheses of available studies meeting clearly specified study quality criteria. No type of evidence is inherently preferable to others; rather, the quality and relevance of the evidence to the intended test score interpretation for a given use determine the value of a particular kind of evidence. A presentation of empirical evidence on any point should give due weight to all relevant findings in the scientific literature, including those inconsistent with the intended interpretation or use. Test developers have the responsibility to provide support for their own recommendations, but test users bear ultimate responsibility for evaluating the quality of the validity evidence provided and its relevance to the local situation.

### Standard 1.3

**If validity for some common or likely interpretation for a given use has not been evaluated, or if such an interpretation is inconsistent with available evidence, that fact should be made clear and potential users should be strongly cautioned about making unsupported interpretations.**

**Comment:** If past experience suggests that a test is likely to be used inappropriately for certain

kinds of decisions or certain kinds of test takers, specific warnings against such uses should be given. Professional judgment is required to evaluate the extent to which existing validity evidence supports a given test use.

### Standard 1.4

**If a test score is interpreted for a given use in a way that has not been validated, it is incumbent on the user to justify the new interpretation for that use, providing a rationale and collecting new evidence, if necessary.**

**Comment:** Professional judgment is required to evaluate the extent to which existing validity evidence applies in the new situation or to the new group of test takers and to determine what new evidence may be needed. The amount and kinds of new evidence required may be influenced by experience with similar prior test uses or interpretations and by the amount, quality, and relevance of existing data.

A test that has been altered or administered in ways that change the construct underlying the test for use with subgroups of the population requires evidence of the validity of the interpretation made on the basis of the modified test (see chap. 3). For example, if a test is adapted for use with individuals with a particular disability in a way that changes the underlying construct, the modified test should have its own evidence of validity for the intended interpretation.

### Standard 1.5

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

**Comment:** If it is asserted, for example, that interpreting and using scores on a given test for employee selection will result in reduced employee errors or training costs, evidence in support of that assertion should be provided. A given claim may be supported by logical or theoretical argument as well as empirical data. Appropriate weight should be given to findings in the scientific literature that may be inconsistent with the stated expectation.

### Standard 1.6

**When a test use is recommended on the grounds that testing or the testing program itself will result in some indirect benefit, in addition to the utility of information from interpretation of the test scores themselves, the recommender should make explicit the rationale for anticipating the indirect benefit. Logical or theoretical arguments and empirical evidence for the indirect benefit should be provided. Appropriate weight should be given to any contradictory findings in the scientific literature, including findings suggesting important indirect outcomes other than those predicted.**

**Comment:** For example, certain educational testing programs have been advocated on the grounds that they would have a salutary influence on classroom instructional practices or would clarify students' understanding of the kind or level of achievement they were expected to attain. To the extent that such claims enter into the justification for a testing program, they become part of the argument for test use. Evidence for such claims should be examined—in conjunction with evidence about the validity of intended test score interpretation and evidence about unintended negative consequences of test use—in making an overall decision about test use. Due weight should be given to evidence against such predictions, for example, evidence that under some conditions educational testing may have a negative effect on classroom instruction.

### Standard 1.7

**If test performance, or a decision made therefrom, is claimed to be essentially unaffected by practice and coaching, then the propensity for test performance to change with these forms of instruction should be documented.**

**Comment:** Materials to aid in score interpretation should summarize evidence indicating the degree to which improvement with practice or coaching can be expected. Also, materials written for test takers should provide practical guidance about the value of test preparation activities, including coaching.

## Cluster 2. Issues Regarding Samples and Settings Used in Validation

### Standard 1.8

**The composition of any sample of test takers from which validity evidence is obtained should be described in as much detail as is practical and permissible, including major relevant sociodemographic and developmental characteristics.**

**Comment:** Statistical findings can be influenced by factors affecting the sample on which the results are based. When the sample is intended to represent a population, that population should be described, and attention should be drawn to any systematic factors that may limit the representativeness of the sample. Factors that might reasonably be expected to affect the results include self-selection, attrition, linguistic ability, disability status, and exclusion criteria, among others. If the participants in a validity study are patients, for example, then the diagnoses of the patients are important, as well as other characteristics, such as the severity of the diagnosed conditions. For tests used in employment settings, the employment status (e.g., applicants versus current job holders), the general level of experience and educational background, and the gender and ethnic composition of the sample may be relevant information. For tests used in credentialing, the status of those providing information (e.g., candidates for a credential versus already-credentialed individuals) is important for interpreting the resulting data. For tests used in educational settings, relevant information may include educational background, developmental level, community characteristics, or school admissions policies, as well as the gender and ethnic composition of the sample. Sometimes legal restrictions about privacy preclude obtaining or disclosing such population information or limit the level of particularity at which such data may be disclosed. The specific privacy laws, if any, governing the type of data should be considered, in order to ensure that any description of a population does not have the potential to identify an individual in a manner inconsistent with such standards. The extent of missing data, if any, and the methods for handling missing data (e.g., use of imputation procedures) should be described.

### Standard 1.9

**When a validation rests in part on the opinions or decisions of expert judges, observers, or raters, procedures for selecting such experts and for eliciting judgments or ratings should be fully described. The qualifications and experience of the judges should be presented. The description of procedures should include any training and instructions provided, should indicate whether participants reached their decisions independently, and should report the level of agreement reached. If participants interacted with one another or exchanged information, the procedures through which they may have influenced one another should be set forth.**

**Comment:** Systematic collection of judgments or opinions may occur at many points in test construction (e.g., eliciting expert judgments of content appropriateness or adequate content representation), in the formulation of rules or standards for score interpretation (e.g., in setting cut scores), or in test scoring (e.g., rating of essay responses). Whenever such procedures are employed, the quality of the resulting judgments is important to the validation. Level of agreement should be specified clearly (e.g., whether percent agreement refers to agreement prior to or after a consensus discussion, and whether the criterion for agreement is exact agreement of ratings or agreement within a certain number of scale points.) The basis for specifying certain types of individuals (e.g., experienced teachers, experienced

job incumbents, supervisors) as appropriate experts for the judgment or rating task should be articulated. It may be entirely appropriate to have experts work together to reach consensus, but it would not then be appropriate to treat their respective judgments as statistically independent. Different judges may be used for different purposes (e.g., one set may rate items for cultural sensitivity while another may rate for reading level) or for different portions of a test.

### Standard 1.10

**When validity evidence includes statistical analyses of test results, either alone or together with data on other variables, the conditions under which the data were collected should be described in enough detail that users can judge the relevance of the statistical findings to local conditions. Attention should be drawn to any features of a validation data collection that are likely to differ from typical operational testing conditions and that could plausibly influence test performance.**

**Comment:** Such conditions might include (but would not be limited to) the following: test-taker motivation or prior preparation, the range of test scores over test takers, the time allowed for test takers to respond or other administrative conditions, the mode of test administration (e.g., unproctored online testing versus proctored on-site testing), examiner training or other examiner characteristics, the time intervals separating collection of data on different measures, or conditions that may have changed since the validity evidence was obtained.

## Cluster 3. Specific Forms of Validity Evidence

### (a) Content-Oriented Evidence

### Standard 1.11

**When the rationale for test score interpretation for a given use rests in part on the appropriateness of test content, the procedures followed in specifying and generating test content should be described and justified with reference to the intended population to be tested and the construct the test is intended to measure or the domain it is intended to represent. If the definition of the content sampled incorporates criteria such as importance, frequency, or criticality, these criteria should also be clearly explained and justified.**

**Comment:** For example, test developers might provide a logical structure that maps the items on the test to the content domain, illustrating the relevance of each item and the adequacy with which the set of items represents the content domain. Areas of the content domain that are not included among the test items could be indicated as well. The match of test content to the targeted domain in terms of cognitive complexity and the accessibility of the test content to all members of the intended population are also important considerations.

### (b) Evidence Regarding Cognitive Processes

### Standard 1.12

**If the rationale for score interpretation for a given use depends on premises about the psychological processes or cognitive operations of test takers, then theoretical or empirical evidence in support of those premises should be provided. When statements about the processes employed by observers or scorers are part of the argument for validity, similar information should be provided.**

**Comment:** If the test specification delineates the processes to be assessed, then evidence is needed that the test items do, in fact, tap the intended processes.

### (c) Evidence Regarding Internal Structure

### Standard 1.13

**If the rationale for a test score interpretation for a given use depends on premises about the rela-**

**tionships among test items or among parts of the test, evidence concerning the internal structure of the test should be provided.**

**Comment:** It might be claimed, for example, that a test is essentially unidimensional. Such a claim could be supported by a multivariate statistical analysis, such as a factor analysis, showing that the score variability attributable to one major dimension was much greater than the score variability attributable to any other identified dimension, or showing that a single factor adequately accounts for the covariation among test items. When a test provides more than one score, the interrelationships of those scores should be shown to be consistent with the construct(s) being assessed.

### Standard 1.14

**When interpretation of subscores, score differences, or profiles is suggested, the rationale and relevant evidence in support of such interpretation should be provided. Where composite scores are developed, the basis and rationale for arriving at the composites should be given.**

**Comment:** When a test provides more than one score, the distinctiveness and reliability of the separate scores should be demonstrated, and the interrelationships of those scores should be shown to be consistent with the construct(s) being assessed. Moreover, evidence for the validity of interpretations of two or more separate scores would not necessarily justify a statistical or substantive interpretation of the difference between them. Rather, the rationale and supporting evidence must pertain directly to the specific score, score combination, or score pattern to be interpreted for a given use. When subscores from one test or scores from different tests are combined into a composite, the basis for combining scores and for how scores are combined (e.g., differential weighting versus simple summation) should be specified.

### Standard 1.15

**When interpretation of performance on specific items, or small subsets of items, is suggested, the rationale and relevant evidence in support of such interpretation should be provided. When interpretation of individual item responses is likely but is not recommended by the developer, the user should be warned against making such interpretations.**

**Comment:** Users should be given sufficient guidance to enable them to judge the degree of confidence warranted for any interpretation for a use recommended by the test developer. Test manuals and score reports should discourage overinterpretation of information that may be subject to considerable error. This is especially important if interpretation of performance on isolated items, small subsets of items, or subtest scores is suggested.

## (d) Evidence Regarding Relationships With Conceptually Related Constructs

### Standard 1.16

**When validity evidence includes empirical analyses of responses to test items together with data on other variables, the rationale for selecting the additional variables should be provided. Where appropriate and feasible, evidence concerning the constructs represented by other variables, as well as their technical properties, should be presented or cited. Attention should be drawn to any likely sources of dependence (or lack of independence) among variables other than dependencies among the construct(s) they represent.**

**Comment:** The patterns of association between and among scores on the test under study and other variables should be consistent with theoretical expectations. The additional variables might be demographic characteristics, indicators of treatment conditions, or scores on other measures. They might include intended measures of the same construct or of different constructs. The reliability of scores from such other measures and the validity of intended interpretations of scores from these measures are an important part of the validity evidence for the test under study. If such variables include composite scores, the manner in which

the composites were constructed should be explained (e.g., transformation or standardization of the variables, and weighting of the variables). In addition to considering the properties of each variable in isolation, it is important to guard against faulty interpretations arising from spurious sources of dependency among measures, including correlated errors or shared variance due to common methods of measurement or common elements.

### (e) Evidence Regarding Relationships With Criteria

#### Standard 1.17

**When validation relies on evidence that test scores are related to one or more criterion variables, information about the suitability and technical quality of the criteria should be reported.**

**Comment:** The description of each criterion variable should include evidence concerning its reliability, the extent to which it represents the intended construct (e.g., task performance on the job), and the extent to which it is likely to be influenced by extraneous sources of variance. Special attention should be given to sources that previous research suggests may introduce extraneous variance that might bias the criterion for or against identifiable groups.

#### Standard 1.18

**When it is asserted that a certain level of test performance predicts adequate or inadequate criterion performance, information about the levels of criterion performance associated with given levels of test scores should be provided.**

**Comment:** For purposes of linking specific test scores with specific levels of criterion performance, regression equations are more useful than correlation coefficients, which are generally insufficient to fully describe patterns of association between tests and other variables. Means, standard deviations, and other statistical summaries are needed, as well as information about the distribution of criterion performances conditional upon a given test score. In the case of categorical rather than continuous variables, techniques appropriate to such data should be used (e.g., the use of logistic regression in the case of a dichotomous criterion). Evidence about the overall association between variables should be supplemented by information about the form of that association and about the variability of that association in different ranges of test scores. Note that data collections employing test takers selected for their extreme scores on one or more measures (extreme groups) typically cannot provide adequate information about the association.

#### Standard 1.19

**If test scores are used in conjunction with other variables to predict some outcome or criterion, analyses based on statistical models of the predictor-criterion relationship should include those additional relevant variables along with the test scores.**

**Comment:** In general, if several predictors of some criterion are available, the optimum combination of predictors cannot be determined solely from separate, pairwise examinations of the criterion variable with each separate predictor in turn, due to intercorrelation among predictors. It is often informative to estimate the increment in predictive accuracy that may be expected when each variable, including the test score, is introduced in addition to all other available variables. As empirically derived weights for combining predictors can capitalize on chance factors in a given sample, analyses involving multiple predictors should be verified by cross-validation or equivalent analysis whenever feasible, and the precision of estimated regression coefficients or other indices should be reported. Cross-validation procedures include formula estimates of validity in subsequent samples and empirical approaches such as deriving weights in one portion of a sample and applying them to an independent subsample.

### Standard 1.20

**When effect size measures (e.g., correlations between test scores and criterion measures, standardized mean test score differences between subgroups) are used to draw inferences that go beyond describing the sample or samples on which data have been collected, indices of the degree of uncertainty associated with these measures (e.g., standard errors, confidence intervals, or significance tests) should be reported.**

**Comment:** Effect size measures are usefully paired with indices reflecting their sampling error to make meaningful evaluation possible. There are various possible measures of effect size, each applicable to different settings. In the presentation of indices of uncertainty, standard errors or confidence intervals provide more information and thus are preferred in place of, or as supplements to, significance testing.

### Standard 1.21

**When statistical adjustments, such as those for restriction of range or attenuation, are made, both adjusted and unadjusted coefficients, as well as the specific procedure used, and all statistics used in the adjustment, should be reported. Estimates of the construct-criterion relationship that remove the effects of measurement error on the test should be clearly reported as adjusted estimates.**

**Comment:** The correlation between two variables, such as test scores and criterion measures, depends on the range of values on each variable. For example, the test scores and the criterion values of a selected subset of test takers (e.g., job applicants who have been selected for hire) will typically have a smaller range than the scores of all test takers (e.g., the entire applicant pool.) Statistical methods are available for adjusting the correlation to reflect the population of interest rather than the sample available. Such adjustments are often appropriate, as when results are compared across various situations. The correlation between two variables is also affected by measurement error, and methods are available for adjusting the correlation to estimate the strength of the correlation net of the effects of measurement error in either or both variables. Reporting of an adjusted correlation should be accompanied by a statement of the method and the statistics used in making the adjustment.

### Standard 1.22

**When a meta-analysis is used as evidence of the strength of a test-criterion relationship, the test and the criterion variables in the local situation should be comparable with those in the studies summarized. If relevant research includes credible evidence that any other specific features of the testing application may influence the strength of the test-criterion relationship, the correspondence between those features in the local situation and in the meta-analysis should be reported. Any significant disparities that might limit the applicability of the meta-analytic findings to the local situation should be noted explicitly.**

**Comment:** The meta-analysis should incorporate all available studies meeting explicitly stated inclusion criteria. Meta-analytic evidence used in test validation typically is based on a number of tests measuring the same or very similar constructs and criterion measures that likewise measure the same or similar constructs. A meta-analytic study may also be limited to multiple studies of a single test and a single criterion. For each study included in the analysis, the test-criterion relationship is expressed in some common metric, often as an effect size. The strength of the test-criterion relationship may be moderated by features of the situation in which the test and criterion measures were obtained (e.g., types of jobs, characteristics of test takers, time interval separating collection of test and criterion measures, year or decade in which the data were collected). If test-criterion relationships vary according to such moderator variables, then the meta-analysis should report separate estimated effect-size distributions conditional upon levels of these moderator variables when the number of studies available for analysis permits doing so. This might be accomplished,

for example, by reporting separate distributions for subsets of studies or by estimating the magnitudes of the influences of situational features on effect sizes.

This standard addresses the responsibilities of the individual who is drawing on meta-analytic evidence to support a test score interpretation for a given use. In some instances, that individual may also be the one conducting the meta-analysis; in other instances, existing meta-analyses are relied on. In the latter instance, the individual drawing on meta-analytic evidence does not have control over how the meta-analysis was conducted or reported, and must evaluate the soundness of the meta-analysis for the setting in question.

### Standard 1.23

**Any meta-analytic evidence used to support an intended test score interpretation for a given use should be clearly described, including methodological choices in identifying and coding studies, correcting for artifacts, and examining potential moderator variables. Assumptions made in correcting for artifacts such as criterion unreliability and range restriction should be presented, and the consequences of these assumptions made clear.**

**Comment:** The description should include documented information about each study used as input to the meta-analysis, thus permitting evaluation by an independent party. Note also that meta-analysis inevitably involves judgments regarding a number of methodological choices. The bases for these judgments should be articulated. In the case of choices involving some degree of uncertainty, such as artifact corrections based on assumed values, the uncertainty should be acknowledged and the degree to which conclusions about validity hinge on these assumptions should be examined and reported.

As in the case of Standard 1.22, the individual who is drawing on meta-analytic evidence to support a test score interpretation for a given use may or may not also be the one conducting the meta-analysis. As Standard 1.22 addresses the reporting of meta-analytic evidence, the individual drawing on existing meta-analytic evidence must evaluate the soundness of the meta-analysis for the setting in question.

### Standard 1.24

**If a test is recommended for use in assigning persons to alternative treatments, and if outcomes from those treatments can reasonably be compared on a common criterion, then, whenever feasible, supporting evidence of differential outcomes should be provided.**

**Comment:** If a test is used for classification into alternative occupational, therapeutic, or educational programs, it is not sufficient just to show that the test predicts treatment outcomes. Support for the validity of the classification procedure is provided by showing that the test is useful in determining which persons are likely to profit differentially from one treatment or another. Treatment categories may have to be combined to assemble sufficient cases for statistical analysis. It is recognized, however, that such research may not be feasible, because ethical and legal constraints on differential assignments may forbid control groups.

## (f) Evidence Based on Consequences of Tests

### Standard 1.25

**When unintended consequences result from test use, an attempt should be made to investigate whether such consequences arise from the test's sensitivity to characteristics other than those it is intended to assess or from the test's failure to fully represent the intended construct.**

**Comment:** The validity of test score interpretations may be limited by construct-irrelevant components or construct underrepresentation. When unintended consequences appear to stem, at least in part, from the use of one or more tests, it is especially important to check that these consequences do not arise from construct-

irrelevant components or construct underrepresentation. For example, although group differences, in and of themselves, do not call into question the validity of a proposed interpretation, they may increase the salience of plausible rival hypotheses that should be evaluated as part of the validation effort. A finding of unintended consequences may also lead to reconsideration of the appropriateness of the construct in question. Ensuring that unintended consequences are evaluated is the responsibility of those making the decision whether to use a particular test, although legal constraints may limit the test user's discretion to discard the results of a previously administered test, when that decision is based on differences in scores for subgroups of different races, ethnicities, or genders. These issues are discussed further in chapter 3.