11

Validity

Suppose you have taken an online test designed to measure where you fall on the introversion/extroversion continuum. You receive the results and are surprised to learn that your score pegs you as an extrovert, as you had always thought of yourself as an introvert. You show the results to your friends and find that they, too, have always seen you as an introvert. This causes you and your friends to question whether the test is really measuring introversion/extroversion and whether the interpretation that you are an introvert is, in fact, justifiable. In psychometric terms, you and your friends question the *validity* of the test. Validity is arguably the most important quality of a test because it has to do with the fundamental measurement issue of what our measurement instruments are really measuring. This may seem a straightforward question, but measurement specialists have long been engaged in discussions about how validity should be defined, how it should be assessed, and what aspects of the testing process should be included under the heading of validity. Although these discussions have resulted in something approaching consensus in some areas, other issues continue to be hotly debated in psychometric circles.

In the following sections, I describe the "traditional" forms of validity evidence: *content*, *criterion-related*, and *construct*, as these are the focus of many of the recent debates. In doing so, I put these into historical context, briefly explaining how the different conceptualizations came about. I then discuss current conceptualizations of validity, with an emphasis on how these differ from the traditional views. In the final section, I turn to the types of validity evidence emphasized in the most recent edition (2014) of the *Standards for Educational and Psychological Measurement* (referred to hereafter as the *Standards*), which is a joint publication of the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME), and is widely considered to be one of the most authoritative sources on measurement standards in the social sciences.

VALIDITY DEFINED

Currently, no definition of *validity* is accepted by all players in the validity debate. The following definition is taken from the *Standards:* "Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA, APA, & NCME, p. 11). Validity thus has to do with the underlying rationale for our interpretations and uses of test scores. In other words, validity has to do with both the *meaning* of test scores and how we use them. As such, validity is justifiably "the most fundamental consideration in developing tests and evaluating tests," as stated in the *Standards* (p. 11). Returning to my earlier example, I might conclude, based on the extroversion test score described previously, that I am much more extroverted than I had ever realized. However, making such an inference from an online test of dubious origin may not be justified. Psychometrically speaking, it may be that the test simply does not yield scores that allow for such an inference. Suppose a company decided to hire salespeople on the basis of high scores on this extroversion test, with the rationale that extroverted people should make the best salespeople. Would doing so achieve the company's goal of obtaining the best salespeople? Probably not, if the scores do not really indicate a person's level of extroversion (and ignoring for now the fact that extroverted people may not actually make the best salespeople). These examples are meant to illustrate the types of "interpretations of test scores for proposed uses of tests" alluded to in the *Standards'* validity definition. But what about commonly used tests, such as those designed to measure one's intelligence, college aptitude, perceived level of pain, or other attributes? Do these have validity for their intended interpretations and uses? And how can we tell? These are the types of questions I address in this chapter.

TRADITIONAL FORMS OF VALIDITY EVIDENCE: A HISTORICAL PERSPECTIVE

Early conceptualizations of test validity focused on the degree to which a test measured "what it is supposed to measure," as stated in an early definition attributed to Garrett (1939, as quoted by Sireci, 2009, p. 22). However, the definition of validity as a determination of whether a test measures what it purports to measure is problematic for at least two reasons. First, a test could "measure what it purports to measure" and still not be good for any useful purpose. To go back to the extroversion example, the developers of the fictitious online test could say that, according to their definition, the term *extroversion* simply means someone who speaks in a loud voice. Items on their instrument could be things like "People say I talk too loudly." Thus, the extroversion test would be measuring what it purports to measure. If that were the end of it, there would be no cause for concern. However, tests are usually developed to be used in some way, such as to predict future behavior or status, or to assess a person's qualifications or suitability for a job or training course. This is what is emphasized by the part of the definition from the *Standards* that says "interpretations of test scores for proposed uses of tests" (2014). If I were to use the "loud voice" extroversion test to, for example, predict who would make a good telemarketer, I would likely meet with little success. This is because the test would be unlikely to result in accurate predictions.

Another reason the definition of test validity as "measuring what it's supposed to measure" was found to be inadequate was that, given the slippery nature of social science constructs, it is often not possible to pin down a particular construct to determine whether a test measures it. As Urbina (2014) points out, although a limited number of constructs, such as narrowly defined content knowledge (e.g., single-digit addition) or skills such as speed or accuracy in typing, are fairly straightforward to define, most social science constructs do not fall into this category. But defining validity as the extent to which tests measure what they are supposed to measure implies that the items included on a test completely define the construct being measured. This can result in definitions that are not particularly useful. Urbina cites as an illustration Boring's (1923) definition of intelligence as "whatever it is that intelligence tests measure." Cliff (1989) called this the *nominalistic fallacy*, or the fallacy of assuming that a test measures something simply because its name implies that it does. The extroversion test alluded to previously is a case in point.

In an attempt to get around the issues inherent in this early definition of validity, a new definition emerged in which validity came to be operationalized as the degree to which scores on the test correlated with scores on another measure of the same attribute. As Guilford (1946), famously stated, "a test is valid for anything with which it correlates" (p. 429). The idea here was that, if there were a "gold standard" of the construct (often, an earlier test), and if scores on the test correlated with that gold standard, the test could be inferred to be a measure of the construct. Note that this is a variety of the "if it walks like a duck and quacks like a duck, it must be a duck" argument. The problem was, of course, that there was no gold standard for most constructs, probably because if there were it may not have been necessary to create the test in the first place. Despite these problems, the correlational view of validity held sway through the 1950s. As noted by many validity researchers (e.g., Angoff, 1988; Borsboom, Cramer, Kievit, Scholten, & Franić, 2009; McDonald, 1999; Newton & Shaw, 2013; Sireci, 2009), this was largely due to the influence of the scientific paradigm of that time. At the time, these early definitions of validity were introduced, the correlation coefficient was a fairly new development (by Karl Pearson in 1896) and was no doubt seen as quite state of the art. At the same time, the prevailing philosophical paradigm was based on logical positivism, which emphasized the use of empiricism, verification, and logical analysis in scientific inquiry, and influenced, among other things, the behaviorist movement of the time. The correlation coefficient provided a means of obtaining empirical and verifiable evidence of validity, so it is not surprising that early validity theorists should have seized upon it as the ideal method for obtaining validity evidence.

Original Validity Types

The first edition of the *Standards* (at that time titled the *Technical Recommendations for Psychological Tests and Diagnostic Techniques* [APA, 1954]) defined four types of validity: *content, predictive, concurrent,* and *construct*. Of these, predictive and concurrent validity reflected the emphasis on correlations, as both types of validity were evidenced by the correlations or regression coefficients of the test of interest with some criterion. These two forms of validity evidence were later subsumed under the heading of *criterion-related* validity. Predictive validity was defined as the correlation between scores on the test of interest (the predictor) with values of a criterion obtained at some point in the future. A widely known example is that of the SAT test, which is purported to predict college grade point average (GPA). In this case, the criterion is students' college GPA, and the predictive validity evidence is the correlation or regression coefficient that measures the relationship between SAT scores and GPA. *Concurrent validity* was defined in a similar fashion, except that, as the name implies, scores on the predictor and on the criterion were obtained at the same time. This type of validity was typically of interest for situations in which one test was proposed as a measure of the same attribute as the other. For example, a newly developed ability test that could be administered in a group format might be proposed as a substitute for a more time-consuming individually administered ability test. For this substitution to be viable, the test developers would have to demonstrate that the group-administered test measured the same attribute(s) as the individually administered test. A high correlation between scores from the two tests would serve as evidence.

Consistent with the empirical orientation of the time, the criteria in predictive validity situations were typically observable or behaviorally defined variables, such as job performance or success in school. Although satisfying the logical positivists, the use of such criteria was problematic on several levels. Job performance measures such as supervisor ratings, for example, often lacked validity evidence themselves. In addition, rating criteria are often inconsistently applied, resulting in a lack of reliability. Other measures of performance, such as the number of widgets produced, were found to measure only a part (and perhaps not the most important part) of job performance. Jenkins, who served during World War II in the Navy Department where he helped to develop tests to select combat pilots, wrote in the aftermath of that war that "there is always the danger that the investigator may accept some convenient measure (especially if it be objective and quantifiable) only to find ultimately that the performance which produces this measure is merely a part, and perhaps an unimportant part, of the total field performance desired" (1946, p. 97). As an example, Jenkins used a situation in which piloting skills might serve as a criterion because they could be objectively scored, but tests of judgment and emotional adjustment, though arguably at least as important, may not be included as criteria because of the greater difficulty in measuring these qualities.

Such observations led researchers such as Jenkins (1946) and Rulon (1946) to suggest that in some situations it is the content of the test and the cognitive processes required to produce a correct answer that is important, not the degree to which the test can predict a criterion. These considerations led to a new type of validity, which came to be called *content validity*. In achievement testing, for example, test scores are taken as indications of the amount of content knowledge examinees have learned, along with the level of cognitive processing skill they have attained. In such situations, interest is naturally focused on the match between the content of the test and the cognitive processes it elicits with the content and processes that have been taught. If the content and processing skills examinees have been taught are not the same as, or at least similar to, those on the test, it is difficult to make inferences about what students have learned. If, however, a content match can be demonstrated, this information would be much more useful than the degree to which the test predicts scores on, say, another test. Thus, content validity emerged in the 1954 *Standards* as a new type of validity, useful for any situation in which the content and/or processes measured by the test were of primary importance. Achievement and aptitude tests were most commonly included in this category.

Finally, as noted by Angoff (1988, p. 25), "it was no coincidence that the 1954 *Standards* listed construct validity . . . as one of the four types." The inclusion of construct validity was more or less assured by the presence on the Joint Committee of Lee Cronbach and Paul Meehl, whose seminal article "Construct Validity in Psychological Tests" (1955) introduced this concept to the world. As the authors stated, "construct validation is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not 'operationally defined'" (p. 282). They go on to state that "when an investigator believes that no criterion available to him is fully valid, he perforce becomes interested in construct validity. . . . Construct validity must be investigated whenever no criterion or universe of content is accepted as entirely adequate to define the construct being measured" (p. 282). These statements suggest that construct validity was seen as the option of last resort, to be used with those troublesome tests that were simply not amenable to content or criterion-related validation. The fact that evidence based on inspection of test content or correlations with criteria was not appropriate for this new type of validity made it clear that new types of evidence were needed.

To meet this need, Cronbach and Meehl (1955) proposed several forms of evidence that might fill the bill. Evidence based on group differences would be appropriate for situations in which two or more groups might be expected to score differently. For example, in measuring attitudes toward gun control, we might expect that those who own guns would have more positive attitudes than those who do not. Correlations of test scores with scores from other tests of the same attribute could also provide evidence of construct validity. In the eyes of Cronbach and Meehl, however, the most sophisticated evidence for construct validity was the elaboration and testing of a *nomological network* in which the construct was embedded. As they stated, "Scientifically speaking, to 'make clear what something *is*' means to set forth the laws in which it occurs. We shall refer to the interlocking system of laws which constitute a theory as a *nomological network*" (p. 290, italics in the original). They go on to specify that such a network must contain at least some observable (i.e., not latent) variables. As Borsboom and colleagues (2009) noted, the idea of the nomological network reflected the logical positivism of the time and mimicked the belief expressed in physics and other physical sciences that the meaning of a theoretical term was provided by the laws that governed its relationship with other things (preferably observable things). It is not surprising that psychologists, who were at that time struggling to gain acceptance for their field as a credible science, would adopt the epistemological stance of the more established physical sciences. Whatever its origins, however, the nomological network remains a valuable heuristic for organizing the theory and empirical findings through which the nature of a construct is understood. Shepard (1997) notes that this type of organizing framework is "quintessentially the model of scientific theory testing" (p. 7), which emphasizes the similarities between validity research and plain old scientific research.

Although Cronbach and Meehl (1955) seem to have felt, based on the statements quoted previously, that construct validity was only applicable when content or criterion-related evidence was inadequate or unattainable, in their very next sentence they state that "determining what psychological constructs account for test performance is desirable for almost any test" (p. 282). Sireci (2009) ascribes this seeming ambivalence about the utility of construct validity to the fact that Cronbach and Meehl, having just introduced the concept, were hesitant about overstating its usefulness. As it turned out, however, they need not have worried, as others were more than willing to so for them. Loevinger (1957), never one to mince words, stated flatly that "since predictive, concurrent, and content validities are all essentially *ad hoc*, construct validity is the whole of validity from a scientific point of view" (p. 636). Loevinger's argument was essentially that none of the other forms of validity required the development of theories that would advance scientific knowledge about the attribute of interest. She likened the difference between criterion-related and construct validities to the difference between learning how to boil an egg through trial and error and learning how to boil an egg by developing an understanding of the chemistry of protein molecules. In this context, she stated:

> The argument against classical criterion-related psychometrics is thus two-fold: it contributes no more to the science of psychology than rules for boiling an egg contribute to the science of chemistry. And the number of genuine egg-boiling decisions which clinicians and psychotechnologists face is small compared with the number of situations where a deeper knowledge of psychological theory would be helpful. (p. 641)

Arguments against the "Tripartite" View of Validity

Several years later, Loevinger's argument was taken even further by Messick (1989), who stated that "construct validity embraces almost all forms of validity evidence" (p. 17). By "almost all" Messick excludes only the appraisal of social consequences, although later in his nearly 90-page chapter, he brings these, too, into the fold of the "unified" view of test validity. I discuss Messick's views in more detail later in this chapter. For now, suffice it to say that in a series of papers (Messick, 1965, 1975, 1980, 1981, 1988), he argued against the traditional "tripartite" view in which content, criterion-related (which subsumed predictive and concurrent), and construct validity were treated as separate but equal frameworks for test validation. Instead, Messick has argued that, because information obtained from both content and criterion-related validity studies contributes to our understanding of the meaning of test scores, and because construct validity is concerned with the meaning of test scores, both content and criterion-related evidence contribute to construct validity and should not be considered as separate "types" of validity.

Instead, all available evidence should be integrated into an overall judgment about the meaning of test scores.

Others have argued against the tripartite view of validity on more pragmatic grounds. For example, Anastasi (1986) argued that the separation of validity into three "types" leads researchers to feel that they must "tick them off in checklist fashion" (p. 2), regardless of the purpose of the test. She goes on to rail against the view that "once this tripartite coverage was accomplished, there was the relaxed feeling that validation requirements had been met." (p. 2). Anastasis argument reflects a widely held view that the segmentation of validity into different "types" has led some researchers to practice what has been termed a "weak program" of validation (e.g., Benson, 1998). In such a program, validity evidence is accumulated in piecemeal fashion without giving sufficient (if any) thought to the types of validity evidence that would contribute most to our understanding of the attribute being measured and how it could be used. Instead, the researcher's efforts simply focus on obtaining evidence from each of the "three C's," regardless of whether this evidence helps to illuminate the meaning of test scores. In contrast, the "strong program" of validation research bears a striking resemblance to the conduct of research in general. The strong program is focused on developing and testing a theory of the attribute being measured, creating a test that reflects this theory, and accumulating evidence specific to the proposed uses of the test (i.e., can it really indicate who would make a good salesperson?) and the proposed interpretations to be made from scores (i.e., does a high score really indicate higher levels of extroversion?).

CURRENT CONCEPTUALIZATIONS OF VALIDITY

In the previous section, I reviewed the history of the traditional tripartite view of validity, which divided validity evidence into the three C's of content, criterion-related, and construct. However, the concept of validity has evolved considerably over the past few decades, and modern validity theory is no longer congruent with the tripartite view. In this section, I therefore discuss the most important aspects of current validity theory, while at the same time introducing some of the current theorists. Because these views are increasingly represented in the measurement literature and will likely dominate that literature in the near future, it is important for you to have an understanding of their major themes. These include the general (though not universal) preference for a unified view of validity; the focus on test score inferences and uses rather than on the test itself; the focus on explanation and cognitive theory in validity studies; and the inclusion of test consequences and values in the validity framework.

The Unified View of Validity

The unified view of validity is now widely held, and it has become more common to refer simply to "validity" rather than to any "type" of validity, as evidenced by this statement from the most recent version of the *Standards*:

> 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*). (2014, p. 14)

This quotation from the *Standards* illustrates several features common to current conceptualizations of test validity. First, following the work of Loevinger (1957), Messick (1975, 1980, 1981, 1988), and others, the term *validity* is now used in its broad sense to refer to all forms of evidence that support the intended interpretations and uses of a test. "Types" of validity such as content-related and criterion-related are subsumed under this broad definition because these contribute to validity in the broader sense of the term. As Messick (1989) states, "Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment (p. 13, original emphasis). Thus, most modern conceptualizations of validity emphasize an integration of all forms of evidence that are useful in elucidating the meaning(s) that can be attached to test scores. It is up to the test user to evaluate the available evidence to judge the degree to which his or her intended interpretation or use is appropriate.

Another unifying theme in Messick's work is the argument that all threats to construct validity are subsumed under the headings of *construct-irrelevant variance* and *construct underrepresentation*. Construct-irrelevant variance is irrelevant in the sense that, although it contributes variability to test scores, this added variability is not due to differences in the construct of interest. Instead, the additional variance is due to sources other than the construct being measured. For instance, some examinees might obtain higher scores on essay questions because of a greater ability to bluff their way through an answer, not because they have greater knowledge or skill. Construct underrepresentation refers to situations in which a test does not completely capture all salient aspects of the construct of interest. Underrepresentation can result from a narrowing of the content of a test. For example, a test of managerial skills might include many items measuring knowledge of budgeting processes but only two items on personnel management. Underrepresentation can also result from a mismatch of the skills elicited by the type of item used with the skills that are of real interest. In the previous example, suppose that both personnel management items were multiple choice. It could be argued that skills in this area are best demonstrated by having examinees respond to scenarios describing personnel issues, or by actual performance in handling a sticky personnel problem. In these examples, the construct of managerial skill would be doubly underrepresented owing to underrepresentation of both content and response processes.

Focus on Interpretation and Use of Test Scores

Another important aspect of current validity theory, illustrated by the previous quotation from Messick (1989), is that it is not the test itself that is validated but the inferences made from test scores or the uses for which the test is intended. As Newton and Shaw (2013) point out, every edition of the *Standards* since the first (1954) has specified that statements about validity are made about the interpretations for particular types of decisions. For example, making inferences about the spelling knowledge of third-grade students on the basis of a third-grade spelling test would likely be supported by available validity evidence. In contrast, making inferences about students' intelligence or likelihood of success in later life on the basis of the same test would not be supported. With regard to test usage, validity evidence might support the use of the test in making classroom decisions about spelling instruction, such as whether remediation is required for some students. However, use of the test to determine which students should be required to repeat the third grade would not be supported. The view that the appropriate object of validation is the inferences made from test scores and not the test itself is widely, though not universally, accepted (Moss, 1992; Shepard, 1993; Zumbo, 2009). Some validity theorists, most notably Borsboom and his colleagues (Borsboom, Mellenbergh, & van Heerden, 2004; Borsboom et al., 2009), disagree with the view that it is the interpretation of a test that is validated, preferring the original definition of validity as the extent to which a test measures what it purports to measure. Borsboom and colleagues state that "the notion of a test score interpretation is too general," applying to "every possible inference concerning test scores" (p. 139). That is, a score can be interpreted in an infinite number of ways, only some of which make sense. Borsboom and his colleagues argue that this makes it difficult to pin down exactly what is meant by test validity.

My own view on the issue of whether validity applies to the test itself or to the interpretations of test scores aligns with that of Markus (2014), who refers to this issue as a nonproductive "pseudo-argument." Markus points out that validity does not refer to a test or an interpretation in isolation, but rather to a relationship between the test, the test scores, the test interpretation, and the test use. That is, a test can be considered valid in the context of one scoring/interpretation/use but not another. Going back to the earlier example of a spelling test, the test would likely be considered valid in a scoring/interpretation/use context in which it is scored correctly (scoring), and the scores are interpreted as indications of spelling knowledge (interpretation) and used to determine which students need more practice in spelling (use). In contrast, the same test would likely be considered invalid in a context in which correct answers to easy spelling words received twice as many score points than answers to difficult words (scoring), or test scores were interpreted as indications of intelligence (interpretation), or if scores were used to determine placement in gifted programs (use). Markus's view is in alignment with Gorin's (2007) definition that "validity is the extent to which test scores provide answers to targeted questions" (p. 456), where by "targeted" Gorin means the test/test score/interpretation/use relationship noted by Markus.

Finally, through its reference to "the accumulated evidence," the definition in the *Standards* emphasizes that obtaining validity evidence is a *process* rather than a single study from which a dichotomous "valid/not valid" decision is made. The attributes researchers attempt to measure in the social sciences are typically latent constructs that, by their very nature, are somewhat elusive. There is thus no definitive study that can pin down, once and for all, the meaning of a construct such as “intelligence” or “creativity.” As some validity theorists have noted, the process of conducting validity studies is very similar to that of conducting research studies in general. Students are taught in their introductory statistics courses that it is not possible to “prove” the null hypothesis but only to disprove it. However, if enough studies accumulate in which the evidence is supportive of a given research hypothesis, we begin to give some credence to that hypothesis. In the same way, it is not possible to *prove* that a test is valid for some purpose, although it is possible to provide evidence that it is not valid for such a purpose. And, as is the case with research studies, the more evidence that accumulates in support of a proposed use of a test, the more credence we are likely to give that use. This is why test validation is best thought of as a program of research in which one attempts to obtain a body of evidence that, taken as a whole, would support the intended uses of and inferences from the test scores.

Focus on Explanation and Cognitive Models

Early tests of achievement were based on theories of learning in which knowledge was thought to be accumulated incrementally. According to such theories, students must first learn factual and procedural information and, when such knowledge is sufficient, can then progress to higher-level skills such as reasoning with, synthesizing, and expanding upon the knowledge. However, more recent learning theories have moved beyond such so-called behaviorist theories of learning and focus more on aspects of learning such as how learners organize information in long-term memory in structures known as *schemas*. Research into learners’ schemas in areas such as physics (Chi, Glaser, & Rees, 1982) and chess (Chase & Simon, 1973) has shown that those who are experts in an area have much more sophisticated schematic structures than novices. In the book *Knowing What Students Know* (National Research Council, 2001), it is argued that more detailed theories that take into account learners’ organizational schemas, common misconceptions, and response processes are necessary to make accurate inferences about student learning. Contributors to this book point out that to make valid inferences about learners’ knowledge, researchers must have an explicit cognitive model of learning based on what is known about how people learn in the domain of interest. Such a model might be more or less detailed, depending on the state of cognitive research in the particular domain and on the complexity of the knowledge or skills being tested, with more complex tasks requiring more detailed models. A cognitive model should include specification of how learners at different levels organize, apply, and transfer knowledge. Although a full explication of cognitive models is outside the scope of this book, it is important to understand this basic framework because such models are central to the thinking of several current validity theorists (i.e., Embretson, 1998, Embretson & Gorin, 2001; Gorin, 2005, 2006; Mislevy, Steinberg, & Almond, 2003).

Cognitive models have been the focus of so much attention in achievement testing because they provide instructors with specific information on learners’ strengths, weaknesses, and misconceptions. A common criticism of many current tests of knowledge is that they do not, in general, provide such information. For example, if a student answers a question incorrectly, it is often difficult to determine the exact source of the problem. It could be that the student did not understand the question, did not possess needed factual information, knew that information but was unable to integrate it to arrive at the correct answer, or any of a host of other possible reasons. Such difficulties occur because many achievement tests have not been based on cognitive models that explicate common misconceptions in the domain, what the knowledge structures of novices and experts look like, and how students progress from basic to more advanced knowledge states. Tests that are based on such cognitive models are much more suited to the job of providing detailed information on learners' levels of knowledge and skill.

Inclusion of Values and Test Consequences in the Validity Framework

One of the more contentious aspects of modern validity theory is the focus on value implications of test scores and consequences of testing. As noted by Kane (2013), among others, the arguments regarding consequences are not about whether the consequences of testing are important to consider when making test-based decisions. Instead, the issues center on whether consequences should be included as part of validity, and if so, whether test developers or test users are responsible for evaluating these. As Kane and other theorists have noted, consequences have always been an important consideration in evaluating tests, for the simple reason that tests are usually given with the expectation that they will yield certain positive consequences (e.g., selecting the most capable employees, preventing unqualified people from entering professions, determining the best type of therapy for a client). Because the main purpose of validation is to determine the likelihood that these benefits will be realized, Kane argues that consequences should be included as part of a validity research program. As Shepard (1997) puts it, once test use is brought into the validity arena, we are "obliged to think about effects or consequences" (p. 6).

Values in Testing

Messick (1995) famously stated, "Validity judgments *are* value judgments" (p. 748). By this he meant that value implications are an inherent part of the meaning of scores, and, because validity has to do with understanding what scores mean, values are inextricably linked to validity judgments. To take a concrete example, we as a society attach certain meanings to construct names such as "assertiveness" and "intelligence; if we did not, we would not be interested in measuring them in the first place. However, the implications of low or high scores often go considerably beyond simple statements such as "Sancho has a high level of assertiveness" or "Jon has lower than average achievement." If Sancho had been female, we might attach a different meaning to his high assertiveness score.

And the fact that Jon received a low score on an achievement test will likely result in his being labeled a low performer or as being in need of remediation—a label that may follow him for the rest of his life. As is well known, such labels can have important implications for one's education. In response to arguments that the inclusion of values under the umbrella of validity would unduly complicate the concept, Messick (1989, 1995) has explained that the inclusion of value implications is not meant to *add anything to* the conceptualization of validity, but rather to make explicit an aspect of validity that is often hidden. The importance of making values explicit is that the implications attached to test scores are then exposed and can be openly debated.

Another aspect of value implications is the broader societal value that was the impetus for obtaining the scores in the first place. The fact that a school, organization, or society at large is interested in a particular type of test score implies that some value, either positive or negative, is associated with the construct being measured. For example, the fact that colleges use aptitude tests as part of their college admissions criteria implies that high aptitude is valued. In fact, this value is so firmly entrenched in our society that you may wonder why I even bother to point it out. But as Shepard (1993) notes, colleges could put more weight on other criteria, such as obtaining a diverse student body. Diversity is also a value espoused by many in American society, but the fact that aptitude is typically weighted more heavily than diversity considerations in college admissions may indicate the value placed on each in educational institutions. As another example, admission to medical school is very competitive, typically requiring high scores on the Medical College Admission Test (MCAT). This reflects the appropriately high value placed on medical knowledge for doctors. However, the emphasis on high levels of medical knowledge may result in less emphasis being placed on attributes such as compassion and communication skills, which many patients feel are also valuable. These examples illustrate Messick's (1995) point that, if such values are not made explicit, those using test scores may not consider whether the values inherent in their testing process are, in fact, the most important ones.

OBTAINING EVIDENCE OF VALIDITY

In this section, I descend from the philosophical heights of validity conceptualizations to the more practical matters of what constitutes validity evidence and what types of evidence are most relevant to different testing situations. I begin by introducing the argument-based approach to validity (Cronbach, 1988; Kane, 1992, 2013). Although this approach was originally suggested by Cronbach (1988), Kane's work has done much to popularize the argument-based approach to validity. Space considerations preclude a full account of the intricacies of this approach, but I hope that the abbreviated version presented here is still useful in demonstrating how to make a basic validity argument. I encourage you to consult the original articles for more details on this useful approach. After introducing the argument-based approach, I discuss the five sources of validity evidence outlined in the *Standards* (2014): evidence based on test content, evidence based on response processes, evidence based on internal structure, evidence based on relations to other variables, and evidence for consequences of testing. Each of these is illustrated with a fictitious example.

Introduction to the Argument-Based Approach to Validity

Recall that validity is defined as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA, APA, & NCME, 2014). This definition implies that it is not possible to obtain evidence to support all possible interpretations or uses of a test. Tests are interpreted and used in many ways, some of which are justifiable and some of which are not. The first step in test validation, therefore, is to specify the intended interpretation(s) and use(s) of the test scores. This should be accompanied by an explanation of how the proposed interpretation is relevant to the proposed use of the test. Suppose that I wanted to develop a test of the ability to generate sound research hypotheses, which I will call the GOSH (Generation of Sound Hypotheses) test. I propose to use this test in selecting students for graduate school. My proposed interpretation of the test scores is that those with higher scores on the test have a greater ability to generate sound research hypotheses. The rationale behind the use of these test scores is that students in graduate school will be engaged in research of some kind, and the ability to generate research hypotheses is crucial to conducting research. Thus, students who do well on the GOSH test should be more successful in graduate school than students who do not.

You may already have identified several problems with my description of the GOSH and its proposed interpretation and usage. For one thing, what do I mean by "*sound* research hypotheses"? Before proceeding with my development of the test, I would have to fully define what is meant by "sound" research hypotheses and identify the skills and knowledge that make up the ability to generate these. Another issue is that not all graduate students are engaged in research that requires the generation of research hypotheses, sound or otherwise. Students in the performing arts, for example, may more likely be engaged in practicing for performances than in hypothesis generation. Thus, my test may not be useful in selecting students for such programs.

This example, and my accompanying critique, represent a streamlined version of Kane's (1992, 2013) argument-based approach to test validation. According to Kane (2013), "The core idea [of the argument-based approach] is to state the proposed interpretation and use explicitly and in some detail, and then to evaluate the plausibility of these proposals" (p. 1). The statements about proposed interpretations and use constitute the validity argument. Claims about validity then depend on the degree to which the accumulated validity evidence supports the claims made in this argument. As I pointed out in the context of the GOSH test, arguments about test interpretations and use are based on a series of assumptions and inferences. For example, I may have initially assumed that the GOSH could be used for selecting students for any type of graduate program. A related underlying assumption of the GOSH is that students in all graduate programs use the same skills and knowledge to generate research hypotheses. This may not be the case, however. Although knowledge of the research process is likely necessary in hypothesis generation, content knowledge is also clearly important. Content knowledge will, by definition, vary across content areas. Thus, some of the assumptions underlying my argument are likely to be violated. If so, my claims that GOSH scores can be interpreted as measures of general hypothesis-generating ability and used to select students for graduate school will not be supported.

Types of Validity Evidence

In the following sections, I describe the five types of validity evidence outlined in the *Standards* (2014): (1) *evidence based on test content*, (2) *evidence based on response processes*, (3) *evidence based on internal structure*, (4) *evidence based on relations to other variables*, and (5) *evidence for consequences of testing*, and I illustrate them using the fictitious GOSH test. Table 11.1 shows the five types of evidence, the general type of validity argument addressed by each, and illustrative methods for obtaining each type of evidence. In the last column I indicate how these relate to the traditional "three C's" (content, criterion-related, and construct validity). Although the tabular format necessitates presenting them separately, the five types of validity evidence are relevant to different aspects of the validity argument and are not intended to be viewed as different "types of validity." Instead, all five types contribute in some way to our understanding of the meaning of test scores, and this meaning is at the heart of validity. However, different types of evidence are needed because test scores may be interpreted and used in different ways, necessitating different validity arguments with different underlying assumptions. The type of evidence that best supports an argument for one use or interpretation may not support others, as is illustrated in the following section. Similarly, support for a particular interpretive/use argument may not require all types of evidence. As Kane (2013) points out, simpler claims regarding interpretation and use of scores require less support than more ambitious claims. Thus, although some types of evidence are often associated with certain types of tests, no type of evidence is exclusive to a particular test type. Researchers should determine the types of evidence that are appropriate based on the type of interpretation and use to be made.

Evidence Based on Test Content

Evidence based on content has to do with the degree to which the content included on a test provides an adequate representation of the domain to be measured. In most, if not all, testing situations, it is not possible to include every item that is part of a construct's domain. For example, the inclusion of every possible two-digit addition item on a test of addition knowledge would clearly result in a prohibitively lengthy test. As noted in Chapter 3, researchers are interested in the responses to specific test items because these responses are thought to be representative of the broader domain. That is,

TABLE 11.1. Types of Validity Evidence with Associated Validity Arguments and Methods

Evidence based on:	Validity argument	Examples of methods for obtaining evidence	Mapping to traditional forms of validity
Test content	Test contains a set of items that are appropriate for measuring the construct	• Table of specifications • Expert review of content/ cognitive processes • Identification of possible construct-irrelevant variance • Identification of possible construct underrepresentation	Content
Response processes	Test items tap into the intended cognitive processes	• Specification of chain of reasoning from item responses to desired inferences • Think-aloud protocols • Eye tracking • Response time • Expert–novice studies • Concept maps • Manipulation of item features and other experimental studies	Construct
Internal structure	Relations among test items mirror those expected from theory	• Item and subscale intercorrelations • Internal consistency • Exploratory and confirmatory factor analysis • Item response theory • Generalizability theory • Studies of differential item functioning	Construct
Relations to other variables	Relations of test scores to other variables mirror those expected from theory	• Test–criterion relations ○ Correlations with other scales or variables ○ Predictive ○ Concurrent ○ Sensitivity and specificity • Group differences • Convergent and discriminant relations • Identification of method variance • Multitrait–multimethod matrices	Criterion-related
Consequences of testing	Intended consequences are realized; unintended consequences are not due to test invalidity	• Determination of whether intended benefits accrue • Identification of unintended outcomes ○ Determination of whether unintended outcomes are due to test irrelevance or construct underrepresentation	Not included

researchers are not so much interested in a student's specific ability to add 17 + 11, but in the broader ability to add two-digit numbers. Thus, researchers interested in evidence based on test content are interested in the degree to which the test items constitute a representative sample of the domain of interest, from which inferences about that domain can reasonably be drawn. Evidence based on test content is therefore relevant to any validity argument that scores can be interpreted and used as representative measures of the knowledge, skills, abilities, or other attributes that make up an identifiable domain. Achievement and employment tests fall into this category, as do certification and licensure tests.

Anastasi and Urbina (1997) state that evidence of content validity is not appropriate for tests of aptitude or personality because "these tests bear less intrinsic resemblance to the behavior domain they are trying to sample than do achievement tests" (p. 117). Certainly, it is easier to delineate the relevant content for tests of knowledge or skill than for tests of aptitude or personality, because tests of knowledge are typically based on a common curriculum or set of experiences. In contrast, content on personality and aptitude tests is often based on a particular theory. For example, a personality test based on the Big Five theory of personality would be quite different from one written by a theorist who did not subscribe to that theory. Thus, in my view, content evidence is relevant to personality and aptitude tests, but definition of the content domain is likely to be more theory-based than curriculum-based. And there will likely be less agreement among researchers about what constitutes an appropriate content domain for such tests than for achievement tests.

In all cases, however, evidence based on content begins with a detailed definition of the domain of interest. For achievement tests, this often takes the form of a table of specifications. Items are then written to reflect these specifications, as discussed in Chapter 3. Such tables delineate both the content to be covered and the cognitive level at which items should be written. Similar tables could be prepared for personality tests, although this is seldom done in practice. More commonly, content for personality tests is based on the researcher's understanding of the theory underlying the construct or on diagnostic criteria, if relevant. For tests assessing job-related knowledge or skills, a job or task analysis should be conducted. In these analyses, the knowledge and skills underlying commonly performed tasks in a particular job are evaluated with the goal of determining which are most important for proficiency on the job.

Once the content domain has been delineated, through a table of specifications, job analysis, personality theory, or diagnostic criteria, it should be reviewed by a panel of experts. The makeup of such a panel will necessarily depend on the content area, but the level of expertise of panel members should be as high as possible. This is because many of the analyses conducted to obtain content-based evidence depend on this expertise. One common form of content-based evidence is that in which panel members independently match the items to the original table of specifications. Crocker and Algina (1986) suggest that the matching process be structured by supplying panel members with copies of the table of specifications and the test items and asking them to write the number of each item in the cell to which they think it belongs. Experts could also be asked to rate the items' relevance and importance for the domain. For employment tests, experts could be asked to rate how important each of the test's items or tasks is to successful performance in the job. Information from such ratings or matching processes can be summarized by computing the average rating, the percentage of matches overall and in each content area or cognitive level, and/or the percentages of items the experts felt were unimportant to performance or did not match the table of specifications. Agreement among the experts' assignments or ratings of items could also be calculated, using the methods for interrater agreement described in Chapter 9. Additional information can be obtained by asking experts to suggest other content areas or job tasks that should be included. For licensure or certification tests, some knowledge or skill areas may be considered more important than others. In such cases, experts should consider whether the areas considered most important are appropriately represented by more items. Finally, in addition to matching the items to the table of specification, experts are often asked to rate items in terms of their clarity and freedom from bias or stereotyping. Because these require different types of expertise, different panels are often convened for these tasks, as noted in the discussion of these aspects of the item review process in Chapter 3.

Construct Underrepresentation and Construct-Irrelevant Variance

Although Messick (1989) stated that the presence of construct underrepresentation and construct-irrelevant variance threaten all validity claims, these two threats are particularly associated with evidence based on content. As discussed previously, construct underrepresentation refers to a situation in which test content is defined too narrowly, leaving out important aspects of the construct. For example, the GOSH test is a measure of the ability to generate sound hypotheses. Suppose that students were provided with a question, such as "Why do young people join gangs?" and asked to generate as many hypotheses as possible. Their score on the GOSH test could then be calculated as the number of hypotheses generated. However, this approach does not include any measurement of the *soundness* of the research hypotheses and would thus not represent an important aspect of the ability to generate sound research hypotheses. Construct underrepresentation can also occur if a test covers the intended content adequately but does not do so at the intended cognitive level. For example, a licensure test for physicians that requires memorization of facts but has no items requiring application of these facts to diagnose diseases would result in underrepresentation of the appropriate knowledge domain.

Construct-irrelevant variance occurs when test scores are influenced by factors that are not part of the intended construct. This is problematic because when irrelevant features of an item affect people's scores, we cannot be sure of the meaning of these scores. The scores are now influenced by both the construct of interest and the irrelevant source, but we do not know how much influence each has on a given score. In the context of achievement testing, Messick (1989) defines two types of construct-irrelevant variance: construct-irrelevant easiness and construct-irrelevant difficulty. The first occurs when features of the test items that are irrelevant to the construct being measured make them easier for some examinees. For example, the correct answer for a multiple-choice item may contain a grammatical clue or be longer than the other options, as discussed in Chapter 4. Students who are test-wise are then likely to choose the correct answer even if they do not have the requisite content knowledge. Or there may be examples in the test items with which some examinees are very familiar, making these items easier for such examinees. If the GOSH test included the question on gang membership mentioned previously, examinees who were familiar with gangs would likely have an advantage. Construct-irrelevant difficulty, in contrast, occurs when features of the item that are irrelevant to the construct being measured make the item more difficult for some examinees. A classic example is the inclusion of overly complex language in achievement test items. Some students, notably those whose first language is not English, may fail to answer correctly not because they lack the content knowledge but because they do not understand the question. Unless the purpose of the test is to measure language ability, language complexity should be kept to a minimum to avoid contamination by this source of irrelevancy.

For personality or attitude measures, construct-irrelevant variance is introduced if responses are influenced by social desirability, a tendency to respond in an extreme fashion, or other response styles. Some respondents have trouble answering negatively oriented questions, as discussed in Chapter 5. Others may not read the questions carefully or may deliberately misrepresent themselves. All such tendencies will result in scores that are contaminated with some degree of construct-irrelevant variance. Messick (1989; see also Shadish, Cook, & Campbell, 2002, p. 452) points out that construct-irrelevant variance due to social desirability or other response artifacts occurs because of *mono-operation bias*, or using only one method to measure an attribute. These authors suggest that researchers use several methods of measurement, such as self-reports, ratings by others, and behavioral observations. One advantage of using different measurement methods is that each method provides a somewhat different view, thus triangulating on the construct. In addition, when such measures are combined, the construct-irrelevant variance should wash out because only the construct-relevant variance will be correlated across methods.

Evidence Based on Response Processes

Interpretations of item responses typically presuppose that respondents have used certain cognitive processes to produce their answers. For example, cognitive theories of responses to noncognitive items assume that respondents have read and understood the question, searched their memory for relevant information, integrated this information into an answer, and correctly mapped this answer onto the response options provided. However, what if the respondent has simply chosen the middle response option without even reading the item, or otherwise satisficed (Krosnick, 1991; see Chapter 5). Such responses should engender doubt about whether the question has measured the intended construct. In the context of achievement testing in mathematics, it is common to use word problems that require a student to determine which type of solution is required. But what if a student has memorized key features of such problems that allow for determining the appropriate solution strategy without using the intended reasoning processes? Such responses would not allow for inferences about the student's level of the reasoning ability of interest.

These scenarios illustrate the importance of understanding the response processes used to produce answers to test items. In many situations, inferences about the meaning of test scores center on the response processes used. If our proposed score interpretations are based on the assumption that certain cognitive processes have been used, these inferences are on shaky ground if this assumption has not been met. Embretson (1983) refers to this aspect of validity as *construct representation,* which she defines as being concerned with "identifying the theoretical mechanisms that underlie item responses, such as information processes, strategies, and knowledge stores" (1983, p. 179). Although Embretson's work is in the area of ability testing, it is important to point out that the same principles can be applied to personality or other noncognitive testing situations.

As Mislevy and his colleagues (Mislevy, 1994; Mislevy, Steinberg, & Almond, 2003; Mislevy, Steinberg, Breyer, Almond, & Johnson, 2002) note, the relevance of item responses and their value in supporting our inferences depends on the chain of reasoning that leads from item responses to the claims or inferences that we wish to make. In the context of the GOSH test, the chain of reasoning might be something like this: "I observe that a student has produced a number of well-reasoned hypotheses based on the scenario given. Based on past experience, I have confidence that the student would likely be able to do this in future situations. I therefore infer that the student has the ability to generate sound research hypotheses in general." Such inferences from responses to some type of stimulus are made every day by physicians, psychologists, social workers, and employers. The stimuli may take the form of responses to interview questions, results of medical tests, observations of behavior or nonverbal communications, or performance of a task. The inference may be based on scientific theory, empirical data, subject-matter expertise, or personal experience. My example from the GOSH test represents a very basic reasoning chain, whereas the examples described by Mislevy and colleagues (2003) are much more complex. For example, Mislevy and colleagues would likely identify the specific cognitive processes needed to produce reasonable hypotheses, such as the abilities to keep the scenario in memory, search memory stores for relevant information, combine the information into a hypothesis, and judge the reasonableness of the generated hypotheses. Nevertheless, my example illustrates the basic process of reasoning from evidence.

One part of this process is to rule out alternative explanations for the observations. If the student in the GOSH example has not been able to generate a single hypothesis, I might reason that the student lacked the ability to do so. However, another explanation is that the student did have the ability but lacked the motivation to use it, or had hypothesis-generating ability but was completely unfamiliar with the scenario provided. You might recognize that these explanations represent construct-irrelevant variance in that they result in variations in performance that are not due to the attribute of interest. Construct-irrelevant variance is problematic because it does not contribute to measurement of the attribute of interest. Or, in the words of Mislevy and colleagues (2003), construct-irrelevant variance does not accumulate. These authors note that a total test score is more informative than a score from a single item because each item score provides a nugget of information that contributes to the attribute of interest. These nuggets of information are cumulated in the total test score, yielding a total score that is the sum of what the items have in common (which, we hope, is the attribute of interest). Because construct-irrelevant variance is not relevant to this common attribute, it does not cumulate across items and therefore does not contribute to measurement of that attribute. As Mislevy and colleagues point out, "the more examinees differ from one another on the knowledge that does not accumulate, the less accurate are the scores over the same number of items. A cardinal principle of our view of assessment design is that what accumulates over tasks should be intentional rather than accidental" (p. 50). Although this statement was made in the context of achievement testing, the same principle applies to any type of test in which item responses are combined in some way. How does one make sure that what is captured by a total score is "intentional rather than accidental"? One way is to devise test items, tasks, or performances in such a way that the aspect of the attribute intended to be measured is isolated as much as possible. That is, make sure that influences of extraneous features such as reading ability, motivation, and contextual variables such as the physical setting or test instructions are minimized to the extent possible. Such test standardization practices are discussed in Chapter 3, but in the framework described by Mislevy and his colleagues, standardization is taken to new heights.

Another way in which researchers can design assessments that capture intentional rather than accidental variance is to determine the influences on item difficulty and generate items in which these influences are manipulated systematically. This is the approach taken by Embretson (1998), who has pioneered this type of item generation in her work on spatial ability. Embretson's approach is to build validity into the test during the item development process by manipulating item features known or hypothesized to affect response processes. For example, in her work on measurement of abstract reasoning, Embretson developed items consisting of matrices of shapes or symbols, similar to those used on Raven's Advanced Progressive Matrices (APM) test (Raven, Court, & Raven, 1992). She manipulated item features such as the number of characteristics that varied across the sequence of symbols and the manner in which these characteristics changed across rows or columns of the matrix. She then used a mathematical model called a *cognitive item response theory model* to test the degree to which the manipulated item features affected item difficulty.

Although the work of Mislevy (Mislevy, 1994; Mislevy et al., 2002, 2003) and Embretson (1998) has been in the area of achievement and ability testing, there is no reason that similar approaches could not be used in the noncognitive arena. Note that the approaches described by these researchers build validity into the test during the item development phase. That is, items are specifically engineered to test hypotheses about the constructs of interest, such as the nature of the cognitive processes underlying item responses. Items could also be designed to reflect features thought to affect the task's difficulty level or, for noncognitive items, the intensity level. Doing this requires a fairly well-developed theory that would point to the characteristics to be manipulated. As Embretson points out, most currently used tests are not based on such theories. Instead, achievement and aptitude tests have been based on manipulation of relatively simple content and cognitive features (e.g., Bloom's taxonomy; see Chapter 3). And although most tests assume that items measure only a single piece of knowledge or a single skill, it is certainly possible to design items that tap into multiple sources of knowledge, skill, or ability (for more on this subject, see Chapter 15). This allows the items to "multitask," and provides more information than more traditional items that tap into only one skill.

In the view of Borsboom and colleagues (2009), evidence about response processes is crucial for validity. They argue that validity is concerned with whether "a measurement instrument for an attribute has the property that it is *sensitive* to differences in the attribute; that is, when the attribute differs over objects then the measurement procedure gives a different outcome" (p. 148, original emphasis). They go on to point out that this requires the researcher to understand the underlying response processes and how these processes are influenced by variations in the attribute being measured, which they refer to as "how the test works" (p. 149). Borsboom and his colleagues are not alone in their focus on response processes. As Gorin (2006) states, "In comparison to earlier theories of assessment design . . . recent test development frameworks rely more heavily on cognition than ever before" (p. 21). The work of researchers such as Embretson (1983, 1998), Gorin (2005, 2006, 2007), Mislevy (1994, 1996; Mislevy et al., 2003), and Wilson (Wilson & Sloane, 2000) exemplify this trend of focusing on how test items "work."

Of course, in some situations the underlying response processes are not crucial to our inferences. The GOSH test may be a case in point. If my only interest is in *whether* students are able to generate sound hypotheses for a given scenario, *how* they do so may be immaterial. I might argue that for a student to succeed in graduate school, the important thing is that students are able to come up with research hypotheses in some way, but it does not really matter how they do so (assuming, of course, that it is not by plagiarizing the ideas of others). The important question to ask in determining whether cognitive process evidence is relevant to a particular testing situation is, "For the inference I wish to make, does the response process matter?" If the answer is "yes," evidence based on response processes is needed.

The current focus on process models for item responses is an exciting development in the testing world and has already resulted in better understandings of the mechanisms by which respondents arrive at answers to test items. However, understanding this process is, as Borsboom and colleagues (1999) point out, "no small matter," as it requires the researcher to "explicate what the property's structure or underlying process is and how this structure or process influences the measurement instrument to result in variations in the measurement outcomes. This seems to be a very daunting task indeed for many psychological properties that researchers claim to measure" (p. 148). I agree with Borsboom et al. that this is a daunting task because it requires the researcher to have a theory of how variations in the attribute of interest produce variations in responses. As these researchers imply, such theories are rare in the area of psychology and likely in most social science areas. Perhaps the most progress in developing such theories has been made in certain areas of achievement testing. Much of this research has focused on explicating what it is that makes some types of test items more difficult than others. Such models have been developed in such areas as physics (Chi, Feltovich, & Glaser, 1981; Chi & Van Lehn, 1991), reading comprehension (Gorin, 2005; Gorin & Embretson, 2006), children's math learning (Brown & Burton, 1978; Griffin & Case, 2007), and abstract reasoning (Embretson, 1998). But, as noted in the book *Knowing What Students Know* (National Research Council, 2001), much more work is needed in this area (p. 179).

Developing Process Models

How, then, does one go about developing such process models? The best place to start is with the theory underlying the construct and/or empirical studies of the phenomenon of interest. Theory can help in understanding the types of processes respondents might use in answering. For example, the cognitive process model of responding to attitude items described by Sudman and colleagues (1996; see Chapter 5) is a good starting place for developing a model of noncognitive response processes. Much of Embretson's (1998) work on abstract reasoning relied on the theories of Carpenter, Just, and Shell (1990), which suggested that working memory capacity was the primary cognitive resource needed for solving matrix problems. Empirical studies in which respondents are asked to say their thought processes out loud as they solve a problem or respond to a noncognitive item can be rich sources of information (see Ericsson & Simon, 1984). In the context of the GOSH test, for example, such *think-alouds* might provide insight into the ways students judge the "soundness" of a hypothesis.

A related method is the *analysis of reasons* in which respondents are asked to provide rationales for their responses. For achievement items, such rationales would provide insight into the algorithms, reasoning processes, and/or cognitive schema underlying students' responses. For noncognitive items, rationales might refer to the types of considerations respondents deliberated in forming their answers, how these were weighted to yield a single response, and/or how the response was mapped onto the rating scale provided.

An *analysis of errors* could also be used for cognitive items. In this method, students' incorrect responses are examined in an attempt to make inferences about possible misconceptions, improper application of algorithms, or faulty reasoning.

Expert–novice studies are designed to elucidate features of the knowledge structures that differentiate beginners from experts in a particular area. Studies of this type have revealed that experts and novices do not simply differ in their *amounts* of knowledge, but in how this knowledge is organized (e.g., Chi et al., 1981, 1982). This suggests that research might profitably be focused on the *schemas* or knowledge representation systems, through which experts and novices encode and organize their knowledge. One way of doing this is to ask respondents to create a *concept map* of a set of terms, problems, or other information that depicts the respondents' understanding of how these are related.

In the 1982 study by Chi and colleagues, experts and novices were asked to create concept maps showing the organizational structure underlying a series of problems in mechanics.

Experimental studies, such as those used by Embretson (1998), can be used to manipulate characteristics of items to determine whether these characteristics affect items responses in expected ways. For example, items features thought to affect the ability to generate a response to a noncognitive item, such as the complexity or familiarity of the attribute, could be manipulated systematically to determine their effects on responses. In some cases, it is possible to manipulate the attribute itself to determine whether responses show a corresponding change. This has been done in studies of test anxiety, in which the construct has been measured during a regular class session and again after a difficult examination has been given. Experimental studies might also involve tracking eye movements or the time taken to respond to an item. Eye-tracking studies can be used to study such things as reading comprehension, by showing specific parts of the text on which readers focus (or fail to focus) and how long this focus lasts. Longer periods of focus may be indicative of greater complexity of the material or of greater interest in the material. Response time measures can be used in similar ways. Items that are more difficult or complex should engender correspondingly longer response times. If this is not the case, it may be that respondents are responding randomly or otherwise lack engagement in the material. In studies of the GOSH, I might vary the complexity of the scenarios to determine whether response times increase with complexity, as expected. I might also track respondents' eye movements in an effort to determine the specific part of the scenario on which respondents focused, or whether respondents reread parts of the scenario in the process of producing their answer.

Evidence Based on Internal Structure

Some tests are designed to measure a single *dimension*, or narrow aspect of a test, such as the propensity to buy a particular type of product or ability to add two-digit numbers. Other tests are designed to measure broader, multidimensional constructs such as general intelligence. Many personality constructs are thought to be multidimensional, such as Sarason's (1984) conceptualization of test anxiety, which posits four test anxiety dimensions: worry, bodily arousal, tension, and test-irrelevant thinking. Tests are typically based on theory, either implicit or explicit, about the dimensionality of the construct being measured, and our interpretations of test scores are based on this assumed dimensionality. We might, for example, refer to the "worry" component of test anxiety or to specific abilities such as spatial ability, and we might make inferences on the basis of these narrower dimensions. Because these inferences assume a specific dimensional structure for the test, it is important to determine whether the test items actually form the separate, identifiable dimensions that are hypothesized. Determining the degree to which test items live up to our dimensional expectations is therefore an important type of validity evidence. Such evidence is crucial in determining the degree to which we are justified in interpreting test scores as representing the posited dimensions. In the 2014

Standards, this type of evidence is referred to as "evidence based on internal structure" but is closely related to what Messick (1995) and Loevinger (1957) referred to as the "structural aspect of validity."

Types of Internal Structure Evidence

At the most basic level, evidence about a test's internal structure could be obtained by examining the pattern of intercorrelations among the items. At the very least, items developed to measure the same dimension should have some level of positive correlation (after re-coding negatively oriented items, if needed; see Chapter 5). Correlations among items on a test considered to be unidimensional should be fairly uniform in magnitude. This is because items on a unidimensional scale should all tap into the same construct, in more or less the same way. Of course, each item will be somewhat idiosyncratic simply because it is worded differently from the other items. However, this should not result in patterns in which some items are much more highly correlated than others. A heterogeneous pattern of correlations would suggest that the more highly correlated items share a source of variance that is not shared by the less highly correlated items. The presence of this additional variance may be evidence that the construct is not unidimensional after all and that the highly correlated items represent one or more narrower subdimensions. Another possibility is that the additional variance is due to construct-irrelevant sources, such as similarity in wording or some type of method effect.

Items on a multidimensional scale should exhibit fairly uniform patterns of interitem correlations *within a dimension*, and these within-dimension correlations should be higher than the across-dimension correlations of the items. Such a pattern is shown in Table 11.2, which depicts the intercorrelations of nine items. As can be seen from the table, items 1–3 are much more highly correlated with each other than with the other six items. The same is true for items 4–6 and 7–9. This pattern suggests that the items are tapping into three different dimensions.

TABLE 11.2. Hypothetical Intercorrelations of Nine Items Measuring Three Dimensions

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9
Item 1	1.0								
Item 2	.60	1.0							
Item 3	.50	.70	1.0						
Item 4	.20	.30	.15	1.0					
Item 5	.15	.20	.20	.60	1.0				
Item 6	.10	.20	.25	.65	.70	1.0			
Item 7	.20	.25	.10	.25	.20	.25	1.0		
Item 8	.30	.10	.15	.20	.20	.15	.55	1.0	
Item 9	.20	.15	.20	.15	.10	.20	.60	.65	1.0

Note that I have not said anything about how high the correlations should be. Intuitively, it may seem that correlations among items measuring the same dimension should be as high as possible. However, as Cronbach and Meehl (1955) have pointed out, high interitem correlations support validity inferences only if they are theoretically warranted. In other words, the theory underlying the construct should determine the appropriate level of interitem correlations. It should be kept in mind that scales based on highly correlated items will be quite narrowly focused. In the case of the GOSH test, asking students to generate hypotheses for two very similar scenarios would yield a much more limited measure of hypothesis-generating ability than use of two very different scenarios. But the scores based on the similar scenarios would be more highly correlated than those based on the different scenarios. Of course, narrowly focused tests based on highly correlated items might be appropriate in some situations. For example, items on scales designed to measure respondents' attitudes about a specific topic, or items developed to measure employees' ability to perform a specific skill within a limited context would be expected to correlate very highly. Messick (1989) suggests that indices of internal consistency, such as coefficient alpha, can be used to provide evidence about internal structure. As he states, "This is relevant validity information because the degree of homogeneity in the test . . . should be commensurate with the degree of homogeneity theoretically expected for the construct in question" (p. 51). Thus, the main point with regard to internal structure analysis is that the structure, whatever it is, should align with theoretical expectations.

Factor Analytic Evidence. Another common form of evidence based on internal structure is that obtained from *factor analytic* procedures. Because I discuss these procedures in detail in Chapters 12 and 13, I will provide only a brief description here. In essence, factor analytic methods answer the question, "What is causing the item scores to correlate in the particular way we observe?" In factor analysis, the answer to this question is that the scores are correlated due to a common cause: the factor(s) or attributes being measured. In other words, respondents vary in their levels of the factors being measured, and these differences in factor levels cause them to provide different responses. For instance, a respondent with a high level of anxiety will provide different answers to items on an anxiety measure than a respondent with low anxiety. Thus, factors cause respondents to answer in particular ways, and this will result in particular patterns of correlations among the items. As another example, note that the pattern of correlations in Table 11.2 suggests three factors.

The input to most factor analytic methods is the matrix of interitem correlations. Factor analytic methods parse these correlations into blocks that represent highly correlated sets of items. This is done by transforming the interitem correlations into a set of *factor loadings* that measure the relation between the factor and the variable. In *exploratory factor analysis* (EFA), researchers can either allow the number of factors to be determined by the computer software being used or can specify that a specific number of factors be obtained. In either case, the loadings of each variable on each factor are estimated. Researchers examine these loadings to determine whether the items written to measure a specific factor all have high loadings on that factor and lower loadings on other factors. In *confirmatory factor analysis* (CFA) the researcher must prespecify the number of factors and the variables associated with each. Both EFA and CFA can help researchers understand both the number and composition of the latent constructs underlying the variables of interest.

Validity arguments based on factor analytic evidence could take several forms. Scales in the social sciences are often designed to measure multiple dimensions, and if so, evidence of this multidimensionality should be provided. A common validity argument is that the hypothesized number of factors will be found *and* that the hypothesized items will load on each. Support for such an argument provides evidence for the hypothesized dimensionality of the test. For example, evidence that the items on Sarason's (1984) measure of test anxiety, the Reactions to Test (RTT) scale loaded as expected onto the four posited test anxiety factors would support interpretations and use of the scores on these factors as measures of worry, test-irrelevant thinking, and so on.

IRT Evidence. IRT methods constitute another set of procedures for examining the degree to which test items conform to a hypothesized structure. Because I discuss IRT methods in Chapter 14, I provide only a brief description here. These methods are similar to those of CFA, but whereas the use of CFA methods assumes that items are measured on a continuous scale, IRT methods are generally applied to dichotomously scored items, such as those commonly found on achievement tests (but see Chapter 14 for extensions to items with multiple scoring categories). IRT models can be used to estimate the probability that an examinee with a given level of ability will answer an item correctly. As discussed in Chapter 14, there are different IRT models, distinguished by the inclusion of different item parameters. The most basic IRT model includes one parameter, known as the difficulty parameter, which is analogous to the difficulty index in classical test theory (see Chapter 6). Thus, IRT models can be used to provide evidence for validity arguments about the relative difficulty of items on a test. For example, I might hypothesize that it will be much more difficult to generate hypotheses for one GOSH scenario than for another scenario. I may therefore make a validity argument that examinees who obtain a high score on the difficult scenario have more ability than those with low scores. This argument would be supported if the IRT difficulty parameters for the two scenarios conformed to my hypotheses.

IRT methods are also commonly used to determine whether items are biased, or, in IRT parlance, whether items exhibit *differential item functioning* (DIF). Although readers may think of item bias as occurring when examinees from different ethnic- or gender-based groups obtain different scores, this is not how bias is defined in the measurement literature. Instead, as discussed in Chapter 16, DIF is described as a situation in which examinees from different groups *who have the same level of ability or achievement* obtain different scores. The distinction in italics is necessary because groups may have valid differences in construct levels. For example, students in the fifth grade would likely do better on a general math test than students in the second grade because fifth graders have more math knowledge. However, this would not be considered bias because we would expect fifth graders to know more. In the same way, groups may differ in knowledge because of such things as differential curricula or opportunities to learn the material. Such differences would not be considered as reflective of test bias because they stem from construct-relevant differences in the construct of interest. In contrast, bias or DIF reflects the influence of construct-irrelevant variation on item responses. A common example is the presence of unnecessarily complex language in a mathematics item. Some test takers may miss the item not because they do not understand the mathematics, but because they do not understand the language. Thus, the presence of DIF suggests that some irrelevant construct (such as language ability) is being measured along with the intended construct.

DIF studies are relevant to validity because they can be used to ferret out sources of construct-irrelevant variance that may otherwise remain hidden. In addition, validity arguments typically make the assumption that the items on a test measure the same construct for all groups of interest—an assumption that is called into question if DIF is present.

Generalizability Theory Evidence. Finally, *generalizability theory* was discussed in Chapter 10 as a method for assessing the extent to which scores are affected by different sources of measurement error. For example, students' writing skills might be assessed by requiring them to write essays in different genres, and these essays might be rated by different raters. The extent to which students' scores are similar across the different genres provides evidence about the breadth of their skills within the writing domain. It may be that some students perform well in narrative writing but not in persuasive writing. It may also be the case that scores obtained from different raters are dissimilar, suggesting that we cannot legitimately generalize scores from one rater to those of other raters. Thus, generalizability theory analyses provide information on the limits that should be placed on our interpretations of scores. Because this information informs the meaning we can make from test scores, it is relevant to test validity.

Evidence Based on Relations to Other Variables

In many cases, interpretations and uses of test scores rely on their relation with other variables. For example, a test may be used to predict which job applicants will make the best employees. Or the theory underlying a test might suggest that those with a specific psychological diagnosis, such as depression, should obtain higher scores than those without such a diagnosis. Such relations are measured by correlation coefficients or regression coefficients from either linear or logistic regression, or by tests of mean differences in test scores across groups, with the choice among the methods depending on the types of variables involved. Categories of evidence discussed in the *Standards* include *test–criterion relationships*, *group differences*, and *convergent and discriminant evidence*. Test–criterion relationships refer to situations in which test scores are used to predict future performance or current status on some criterion. For example, recall that I proposed using the fictitious GOSH test as a means of selecting students for graduate school, arguing that students with high scores on the GOSH test should be more successful than those with low scores. My validity argument is therefore that GOSH scores are predictive of graduate school success. To back up this claim, I would have to demonstrate empirically that GOSH scores are predictive of the criterion of graduate school success—an example of the prediction of future performance.

As an example of the prediction of current status, suppose that a researcher develops a multiple-choice test designed to assess the same skills as the GOSH test. If the GOSH test were an established instrument, the researcher's validity argument might be that scores on her test are highly correlated with, or predictive of, scores on the GOSH, and are therefore measuring the same construct. Of course, to back up her argument, the researcher would have to show that scores on her test are, in fact, highly correlated with scores on the GOSH test. Another type of test–criterion relationship involves the use of test scores for assigning individuals to different treatments, jobs, or educational programs. Within the educational system, for example, tests are sometimes used to determine whether students should take an advanced placement or remedial class. The logic behind such placements is that students with higher scores are more likely to benefit from an advanced placement course, whereas students with lower scores will benefit from remediation before moving on in the curriculum. The validity of this type of test use depends on the degree to which the hypothesized benefits actually accrue. If students assigned to the remedial class would actually have done well in the advanced placement class and/or if students assigned to the advanced placement class are unable to keep up in that class, doubt would be cast on these uses and interpretations of the test scores.

Some validity arguments are predicated on hypotheses that respondents in different groups should score differently. This is common for tests designed to identify those with a particular mental disorder. For such tests, the validity argument is that the scores of those diagnosed with the disorder should differ from the scores of those without such a diagnosis. For example, an appropriate form of validity evidence for a test designed to measure social anxiety would involve comparing the scores of those who had been independently diagnosed with social anxiety and those who had not been. Such studies are commonly referred to as *known groups* studies. Clearly, if scores of the two groups do not differ, the test is of little use in diagnosis, so evidence of group differences is essential for such tests.

Finally, validity arguments involving *convergent and discriminant evidence* state that test scores should be related to scores from other tests of the same or similar constructs (convergent evidence) and should be less strongly related to scores from tests measuring dissimilar constructs (discriminant evidence). For instance, if a researcher wanted to show that scores on an attitude measure were not influenced by social desirability, attitude scores could be correlated with scores on a measure of social desirability. A correlation close to zero would provide the desired discriminant evidence. In the context of the GOSH test, it might be argued that GOSH scores should be related to scores on a test of inductive reasoning (convergent) but should not be related to writing ability (discriminant).

In the following sections, I briefly discuss the types of evidence relevant to arguments based on test–criterion relationships, group differences, and convergent and discriminant relationships. I also highlight practical considerations in designing studies to obtain such evidence.

Test–Criterion Relationships

Recall my earlier proposal to use the fictitious GOSH test as a means of selecting students for graduate school, on the basis that students with high scores on the GOSH test should be more successful than those with low scores. However, it would be foolish to accept this claim without any evidence to back it up. To obtain such evidence, I would have to show empirically that those with high GOSH scores are more successful in graduate school than those with low scores. Arguments that test scores can be used to predict such things as success in education or training programs or performance in employment settings are common in the testing arena. Such claims underlie the use of test scores to select those who will be admitted to college or given a job. In these situations, evidence of the test's predictive ability is crucial to the validity argument. If the test does not predict success, using it as a selection tool is questionable, at best. Another form of test–criterion relationship is that in which scores on two tests or on a test and a criterion are obtained concurrently. Evidence of concurrent test relationships is necessary whenever it is argued that one test can be used as an alternative to another, as in my example of the multiple-choice GOSH test. Other examples are the development of a shorter form of a longer test or a less expensive measure designed as an alternative to an existing, more expensive test. An example from the latter category might involve a paper-and-pencil test of anxiety proposed as an alternative to an existing physiological test. If the shorter or less expensive test is intended to replace the longer or more expensive test, there must be empirical evidence showing that scores from the two tests are highly correlated. Another situation in which concurrent evidence is relevant is that of psychodiagnostic tests. Such tests are often validated against the clinician's diagnosis. If the test yields the same diagnosis as the clinician, use of the test for diagnostic decisions is supported.

Issues with Test–Criterion Relationships

Selection of an Appropriate Criterion. A common issue with evidence based on test–criterion relationships is the selection of an appropriate criterion. When a new test is suggested as a substitute for an existing test, as in concurrent validity situations, selection of the criterion is straightforward as the criterion is simply the existing test. For tests purported to predict a particular outcome, such as success in school or on the job, however, obtaining scores on an appropriate criterion can be problematic. In my previous example of the GOSH test, I stated that GOSH scores should be related to success in graduate school. In presenting that example, I skirted the issue of what I meant by "success in graduate school," but if I were to actually conduct such a study success would have to be defined. Does success in graduate school mean a high GPA, the number of conference presentations or research publications produced, the number of citations of a student's research, all of these, or something else? GPA and number of citations have the advantages of being easily obtained and quantifiable but may not correspond to what many people think of as "success." Number of presentations or publications is arguably more closely related to the ability to generate sound hypotheses because such products presumably are based on such hypotheses. But should all presentations be weighted the same, or should international presentations count more than local presentations?

Similar questions arise in other areas in which tests are hypothesized to predict performance, such as employment testing. What should the criterion be for a test designed to identify the best employees? Actual performance on the job is probably the most relevant criterion, but how should it be measured? In most cases, this is not as straightforward as counting the number of widgets produced. One possible criterion is supervisor ratings, but as readers who have held jobs can imagine, such ratings may not always be accurate reflections of performance. One commonly discussed influence on supervisor ratings is *criterion contamination*, which occurs when supervisors know employees' scores on the employment test and allow this knowledge to affect the ratings they give. This is similar to the halo effect discussed in Chapter 4.

Restriction of Range. As noted previously, evidence for test-criterion relations typically takes the form of a correlation or regression coefficient and is therefore subject to the factors influencing these coefficients. Restriction of range, as discussed in Chapter 8, occurs whenever the full range of scores on either the predictor or criterion variable (or both) is not obtained. Because the whole point of using tests for selection purposes is to choose the highest (or in some cases, the lowest) scorers, restriction of range is almost always an issue. Recall that restriction of range in either the predictor test or the criterion can result in lower values of the correlation coefficient than would be obtained if the full range of scores was available. Thus, the true correlation of the test with a criterion may be higher than that obtained from a restricted sample.

Other Factors Attenuating Predictor–Criterion Relations. Another potential issue is that commonly used correlation coefficients such as Pearson's correlation are based on the assumption that the relation between the predictor and criterion is linear. This may not be the case in many situations. For example, suppose that a performance requires a certain level of ability but after that point, having more ability does not increase performance. In this case, performance will increase with ability up to the requisite level, but the relationship will then remain constant, resulting in a nonlinear pattern. Another attenuating factor in predictor–criterion relations is a lack of reliability in either. In general, the relation between a predictor test and a criterion will be attenuated to the extent that either the predictor or criterion is not measured reliably. This effect is ubiquitous in measurement theory, as evidenced by Equation 11.1, which provides the relation between the predictor–criterion correlation and the reliabilities of the two.

$$\rho_{XY} \circ \sqrt{\rho_{XX'}\rho_{YY'}} \tag{11.1}$$

ρ_{XY} is the correlation between predictor and criterion, and $\rho_{XX'}$ and $\rho_{YY'}$ are the reliabilities of the two. Equation 11.1 shows that the correlation between two scores is restricted by their reliabilities. When measuring such relationships, what we would really like to know is the correlation between the two true scores, or the correlation between the error-free measures of *X* and *Y* ($\rho_{t_X t_Y}$) The so-called correction for attenuation formula in Equation 11.2 provides an estimate of the correlation between the true scores of *X* and *Y*; that is, it is the correlation we would have obtained if the measures were perfectly reliable. Note that the observed score correlation (ρ_{XY}) is adjusted upward by dividing it by a function of the reliabilities of the two scores. This makes sense conceptually because the correlation between the true scores should be higher than that between the observed scores to the extent that the observed scores are unreliable.

$$\rho_{t_X t_Y} = \frac{\rho_{XY}}{\sqrt{\rho_{XX'}\rho_{YY'}}} \tag{11.2}$$

However, as noted by Crocker and Algina (1986), researchers rarely, if ever, have perfectly reliable measures in practice, so Equation 11.2 yields an overestimate of the correlation that would be obtained from actual tests. McDonald (1999), however, points out that "in many applications, we wish to know how well the given test predicts, in spite of its measurement properties" (p. 227), and Equation 11.2 answers this question.

Logistic Regression

The discussion to this point has been based on the assumption that both the test score and the criterion can be treated as continuous. However, in some situations, the outcome to be predicted takes the form of categories, such as whether a person passes or fails, responds to treatment or not, or drops out or does not drop out of school. With categorical outcomes, the overarching validity argument is that the test can accurately classify respondents into the correct category. To obtain evidence supportive of this argument, researchers would administer the test to respondents and obtain information on their subsequent status on the categorical variable. A regression method such as logistic regression, which is suitable for categorical outcomes, would then be used to determine the degree to which the test accurately predicts the outcome. Although space concerns preclude a full treatment of logistic regression here, readers unfamiliar with logistic and other regression techniques suitable for categorical data are referred to Cohen, Cohen, West, and Aiken (2003) or other regression texts. However, because the topic of classification accuracy, which is relevant to measurement validity, is not covered in depth in most regression texts, I include a brief discussion of this topic in the following paragraphs.

Classification Accuracy. One outcome of logistic regression models is the predicted probability, given a particular score on the test, that a respondent is either "positive" (passes, drops out, responds to treatment) or "negative" (fails, does not drop out, does not respond to treatment). Note that for dichotomous outcomes, only one probability is obtained because the probability of a negative classification is simply one minus the probability of a positive classification (because the probabilities must sum to one). Cut scores are chosen by first determining a probability above which a person would be classified as a "positive" and below which the person would be classified as a "negative." For example, probabilities of 0.6 or above might be considered as a positive, and probabilities less than 0.6 as negatives. After deciding on this probability, the logistic regression equation can be used to determine the test score that corresponds to that probability. This score is then set as the cut score for classifying people as positives or negatives. Classification accuracy can then be determined by comparing the results of these classifications to the actual status of the person.

With a dichotomous (two-category) outcome, there are four possible outcomes: classification of a person as being in the positive category either correctly (*true positive*) or incorrectly (*false positive*), and classification of a person as being in the negative category either correctly (*true negative*) or incorrectly (*false negative*). These four terms give rise to two further specifications: *sensitivity*, or *true positive rate*, and *specificity*, or *true negative rate*. Sensitivity is calculated as the number of true positive classifications divided by the total number of actual positive outcomes (i.e., number of true positive + false negative classifications). Sensitivity measures answer the question, "Of those with observed positive outcomes, what proportion were correctly classified as positive on the basis of the test?" Specificity is calculated as the number of true negatives divided by the total number of actual negative outcomes (i.e., true negatives + false positives) and addresses the question, "Of those with observed negative outcomes, what proportion was correctly classified as negative on the basis of the test?" The calculations for sensitivity and specificity are illustrated in Table 11.3.

In this illustration, the sensitivity of the test, or the proportion correctly classified as having a positive outcome, is .7, or 70%. The specificity, or proportion correctly classified as having a negative outcome, is .9, or 90%. Is this good? The answer to this question depends on the relative consequences of false positives and false negatives in any given situation. The terms *sensitivity* and *specificity* were originally developed in the context of medical studies. In those studies, a "positive" outcome is usually the presence of a disease (although this does not seem particularly positive), so a false positive would

TABLE 11.3. Numbers of Respondents Correctly and Incorrectly Classified by a Test

Predicted outcome (based on test)	Actual outcome		Total predicted positive and negative
	Positive	Negative	
Positive	True positives (TP) = 35	False positives (FP) = 5	40
Negative	False negatives (FN) = 15	True negatives (TN) = 45	60
Total actual positive or negative	50	50	Total N = 100

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} = \frac{35}{50} = .7 \qquad \text{Specificity} = \frac{\text{TN}}{\text{TN}+\text{FP}} = \frac{45}{50} = .9$$

mean a patient is classified as having a disease when it is not really present, and a false negative would mean that the patient is not classified as having the disease when it really is present. In this case, a false negative is probably more consequential than a false positive because a person classified as a false negative would not receive needed treatment, whereas a false positive would likely be identified as such upon further testing. In educational settings, sensitivity and specificity are often calculated for the purpose of diagnosing learning disabilities. Because students diagnosed with such disabilities usually receive additional resources, it could be argued that a false negative classification is more serious than a false positive. Whereas a false positive diagnosis would result in a student receiving unneeded services, a false negative diagnosis would result in needed services being withheld.

Although ideally we would prefer to minimize both false positives and false negatives, this is possible only if the test is 100% accurate, which is an unlikely scenario. The interplay between false positives and false negatives can be seen from Figure 11.1, which shows the proportions of true and false positives and negatives for three different test cut points. Recall that decisions regarding classification of a respondent as positive or negative are made on the basis of a particular cut score on the predictor test. These cut scores can be chosen by the test user to minimize false positives or false negatives. In Figure 11.1, three cut scores are shown as vertical lines in the center of the figure. The shaded ellipse represents the obtained test scores and outcomes. The cut score represented by the heavy black line in the center of the figure would balance false positives and negatives fairly evenly. Choice of the lower cut point to the left would minimize false negatives but increase false positives, whereas the higher cut point to the right would minimize false positives at the expense of false negatives. This is somewhat similar to the situation in statistical hypothesis testing, in which minimization of Type I errors comes at the expense of Type II errors. To determine the "optimal" cut score, therefore, researchers must first decide whether it is more important to minimize false positives or false negatives. Researchers can then try out different cut scores, using them to make positive and negative classification decisions, and calculating the sensitivity and specificity values for each. To illustrate, I constructed the hypothetical data in Table 11.4 in which dropout status is predicted by a test score. Three different cut scores on the test are used to determine dropout status: low, medium, and high. Numbers in the body of the table represent the number of people with each dropout status.

As can be seen in Table 11.4, use of the low cut score maximizes sensitivity (.98), or the proportion of true positives, and minimizes the proportion of false negatives (.02). In this scenario, a true positive is a student who is predicted to drop out and does drop out, whereas a false negative is a student who is not predicted to drop out but actually does. Use of the high cut score results in the greatest specificity (.97), or the largest proportion of students correctly predicted not to drop out. The proportion of false positives, or students who are predicted to drop out but do not, is lowest using the high cut score at 0.03. Use of the medium cut score results in the largest combined total for sensitivity and specificity. This cut score would therefore result in the greatest number of students being correctly classified overall. If test users were most concerned with minimizing false positives, the

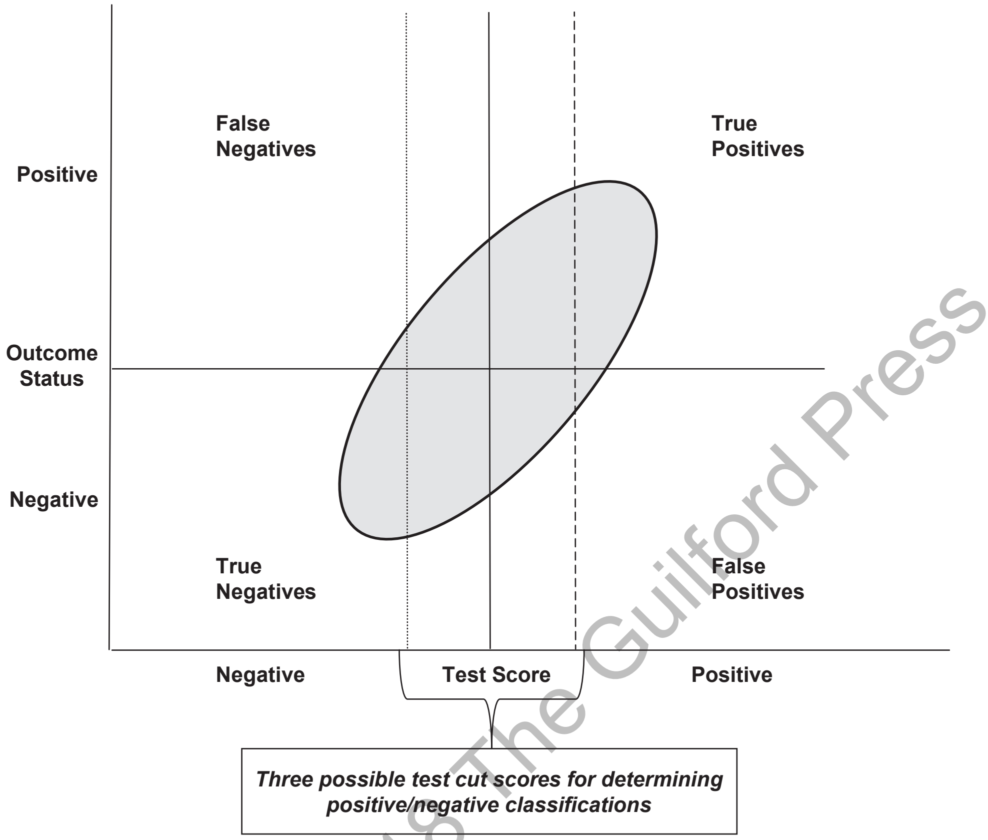

FIGURE 11.1. False positives and false negatives for three different cut scores.

high cut score should be used, whereas if false negatives were of most concern, the low cut score should be used. If test users want to obtain the largest number of correct decisions overall, the medium cut score should be used. Note that these results are consistent with Figure 11.1, which shows that the number of false positives decreases as the cut score increases, and the number of false negatives decreases as the cut score decreases.

As is evident from the previous discussion, selection of a cut point involves a tradeoff between sensitivity and specificity. Table 11.4 shows these values for three possible cut points, but a more fine-grained examination of the relationship between the two can be obtained from a *receiver operator characteristic* (ROC) *curve*, which plots the percentage of true positives (sensitivity) against the percentages of false positives (1 – specificity) for each possible cut score. The ROC curve thus provides a graphical way of showing the percentages of true positives and false positives for each possible cut score.

Group Differences

Evidence about group differences is relevant in any situation for which logic or theory (or, one hopes, both) dictates that one group should obtain higher scores than

TABLE 11.4. Determination of Dropout Status Using Three Cut Points

	Low cut score		Medium cut score		High cut score	
	Actual outcome					
Predicted outcome	Dropout	No dropout	Dropout	No dropout	Dropout	No dropout
Dropout	98	78	82	28	37	6
No dropout	2	122	18	172	63	194
Total	100	200	100	200	100	200
Sensitivity	98/100 = .98		82/100 = .82		37/100 = .37	
Specificity	122/200 = .61		172/200 = .86		194/200 = .97	
Sensitivity + specificity	1.59		1.68		1.34	

another on the test of interest. For example, those clinically diagnosed with depression would be expected to obtain higher (more depressed) scores on a depression scale than those who were not so diagnosed. Or, I might expect that scientists would obtain higher scores on the GOSH test than would professional wrestlers because scientists have presumably had more practice than wrestlers in hypothesis generation. Studies of group differences, typically referred to as *known groups* studies, can shed light on score interpretation and use. If diagnosed and nondiagnosed groups were found to have equivalent mean scores on a depression instrument, this would call into question the meaning of scores as measures of depression. If on one hand, students who had taken a measurement theory course and those who had not taken such a course obtained the same scores on a test of that content, I would be hesitant to use that test to assign grades in my measurement theory course. On the other hand, if the expected group differences were found, the meaning of the scores as measures of the intended construct would be enhanced.

Another type of differences are *within-person differences*, or differences in scores obtained from people at two or more different times. Many skills and abilities, such as motor coordination and reading ability, increase steadily throughout childhood. Other abilities increase up to a certain point and then decrease. For instance, vocabulary generally increases throughout early adulthood, reaches its peak in middle age, and declines slowly after that point. Reasoning ability shows a similar pattern but declines more quickly after its peak in late adolescence. In the case of the GOSH test, I would expect students' hypotheses to become increasingly sophisticated throughout their school careers, possibly reaching a peak in adulthood and then leveling off. For constructs such as these that are expected to change in a particular way across time, longitudinal studies can provide evidence for the validity of scores. A finding that scores change over time as expected would support interpretations of the scores as measures of the intended construct.

Convergent and Discriminant Evidence

An important aspect of the theory surrounding a construct is an explication of its expected relations with other constructs as well as with observed variables. These relations are important to understanding a construct because they help to situate the construct within existing theory. In addition, explication of a construct's network of relations provides a basis for distinguishing it from similar constructs. Finally, knowledge of such networks helps in understanding a construct's possible utility for practical purposes such as prediction, and for theoretical purposes such as elaboration of related theories. Explication of this network of relations is sometimes referred to as embedding the construct in a *nomological network*. As discussed earlier in this chapter, the concept of a nomological network was introduced in the classic article by Cronbach and Meehl (1955), who defined it as "the interlocking system of laws which constitute a theory" (p. 290). In this section, I focus on two important features of a construct's relations: its *convergence* with other constructs and its *discriminability* from other constructs.

Constructs are often hypothesized to share certain characteristics with other constructs. If such similarities make up part of the theory underlying a construct, evidence of such convergence is relevant. For example, depression and anxiety are both considered to be aspects of negative affect, and they often covary. An empirical finding of independence between scores from scales purported to measure depression and anxiety would likely cause researchers to question whether the scales used to measure one or both of these were valid. *Discriminant evidence* refers to the fact that constructs should not be redundant; that is, a construct should not be a reformulation of another construct. Thus, researchers proposing a construct such as test anxiety must be prepared to differentiate it from an existing construct such as performance anxiety. In the case of the GOSH test, I would have to explain how generating sound hypotheses differs from the broader construct of inductive reasoning or from creativity.

Messick (1989) defines *trait validity* as the notion that "constructs should not be uniquely tied to any particular method of measurement nor should they be unduly redundant with other constructs" (p. 46). The second part of this definition refers to discriminant evidence. In the first part of the definition, Messick refers to the fact that different methods of measurement can yield scores that are quite dissimilar. For example, it is well known in educational measurement that students can score quite differently on essay and multiple-choice tests of the same content because these testing formats require different types of knowledge and response processes. Similarly, measures of personality based on self-reports and on reports by others may show little convergence. The extent to which scores diverge across different methods of measurement is referred to as *method variance*. In general, the presence of method variance is undesirable because it suggests that scores are tied to a specific type of measurement, and this narrows the interpretation of the construct. Ideally, we would like to see scores converge across different methods of measurement.

In a classic article, Campbell and Fiske (1959) introduced the *multitrait–multimethod* (MTMM) *matrix* as a way of simultaneously assessing the degree to which measures of the same construct using different methods converge (lack of method variance) and the degree to which measures of different, but related, constructs converge (convergent evidence). To construct an MTMM matrix, a researcher obtains scores on traits, or constructs, that are similar but differentiable. All of the traits are measured by at least two methods, such as self and peer ratings, paper-and-pencil measures and observations, or whatever methods are appropriate. Correlations among all the traits measured by all methods are then entered into a matrix. A hypothetical MTMM matrix for the GOSH is shown in Table 11.5. In the matrix, there are three traits or constructs: ability to generate sound hypotheses (GOSH), ability to reason inductively (ARI), and creativity (CRE). Although the three constructs are likely related, my hypothesis is that they are differentiable. Thus, although I expect that they will be positively correlated, their correlations should not be so high as to suggest that they are measuring the same construct. Each of the three traits is measured by two methods: a written, open-ended test and a multiple-choice test. Ideally, three or more methods would be used. I illustrate the concept with only two methods here simply for ease in presentation.

The entries in the table are as follows: The diagonal entries in parentheses are the reliability values for each trait/method combination (i.e., the value of .79 is the reliability coefficient for the GOSH measured by an open-ended test). The bolded values on the bottom left-hand side (.60, etc.) are the so-called *validity coefficients*, or the *monotrait–heteromethod coefficients*. These are correlations of the same trait measured by different methods and should be high if method variance is not excessive. Low values indicate that scores lack generalizability across methods. That is, an examinee could obtain very different scores from the open-ended and multiple-choice tests. The coefficients in Table 11.5 are reasonably high, given that open-ended tests likely measure a different aspect of the three traits than can be measured using a multiple-choice test. These validity coefficients should be larger than the *heterotrait–monomethod coefficients* shown in the triangles enclosed in solid lines. The heterotrait–monomethod coefficients are correlations of different traits measured by the same method and should be lower than correlations of the same trait measured by different methods (e.g., the validity coefficients). If this is not the case, it would indicate that substantial method variance exists,

TABLE 11.5. Hypothetical MTMM Matrix

		Method 1: Open-ended			Method 2: Multiple-choice		
	Traits	A1	B1	C1	A2	B2	C2
Method 1: Open-ended	A1: GOSH	(.79)					
	B1: ARI	.40	(.75)				
	C1: CRE	.30	.20	(.70)			
Method 2: Multiple-choice	A2: GOSH	**.60**	.35	.25	(.85)		
	B2: ARI	.25	**.65**	.20	.45	(.83)	
	C2: CRE	.20	.20	**.50**	.35	.25	(.80)

resulting in high correlations among same-method measures, even though these are not measuring the same thing. In Table 11.5, the heterotrait–monomethod coefficients are lower than the validity coefficients, indicating that method variance is not a substantial problem. Finally, the *heterotrait–heteromethod coefficients* shown in triangles enclosed in dashed lines should be the lowest in the table, because these are correlations among measures that share neither trait nor method. This is the case in Table 11.5, as the heterotrait–heteromethod coefficients are lower than either their heterotrait–monomethod or monotrait–heteromethod (validity) counterparts.

At this point, you may be wondering how to determine whether a given correlation should be considered "high" or "low." The answer is that these are relative terms and depend on the particular application. One advantage of the MMTM is that it provides some context for determining what is high or low. With a matrix, the monotrait–heteromethod (validity) correlations should be highest, heterotrait–heteromethod correlations should be lowest, and heterotrait–monomethod correlations should be somewhere in between. It is this pattern of correlations, rather than their absolute sizes that is important. In the current example, I relied on the interocular method, more commonly known as eyeballing, to discern this pattern. However, methods of CFA described in Chapter 13 can provide more rigorous tests of MTMM structures.

Evidence Based on Consequences of Testing

Although not all theorists agree that test consequences fall under the purview of validity, evidence for consequences of testing has been included in the latest version of the *Standards*, so here I include examples of the possible forms such evidence might take. Before doing so, however, I provide a brief overview of recent discussions of consequences in testing. In a previous section I alluded to the fact that tests are typically given with the expectation that they will yield certain positive consequences (e.g., selecting the most capable employees, preventing unqualified people from entering professions, determining the best type of therapy for a client). In these examples, the consequences share two important features. First, they are *intended* consequences; that is, the purpose of testing, at least in part, is to accrue these benefits. Second, they are *positive* consequences, in the sense that the benefits of testing (obtaining qualified employees) outweigh the negative consequences (e.g., potential employees who did not do well on the test but have important job skills not measured by the test may not be hired).

In contrast, much of the recent theorizing in the area of test consequences has focused on those consequences that are unintended and negative. For example, if medical school admissions decisions were based solely on high scores on the MCAT, this may have the unintended consequence of yielding doctors who lack communication skills (although I am not, of course, suggesting this is the case). A commonly cited unintended consequence of state-level achievement testing is that such testing programs result in a narrowing of the curriculum to focus only on the material included on the test. Turning to negative testing consequences, situations in which the test scores of certain gender- or ethnic-based groups result in their being selected at lower rates for jobs or scholarships, or at higher rates for remedial classes (known as *adverse impact*), are common examples.

Although consequences originally entered into the validity framework in their positive form, more recently the focus has shifted to studies designed to detect negative consequences of testing. In addition, the topics of whether such studies should be included in the validity framework and of whether a finding of negative consequences invalidates a test are the subjects of much discussion. There is no consensus on these issues, but most theorists feel that a validity study should, at the very least, include information on whether negative consequences are outweighed by positive consequences. If there are negative consequences, the test and testing procedures should be investigated to determine whether these consequences are due to a source of invalidity in the test. For example, if students whose first language is not English obtain low scores on a math test because the test contains many word problems, the test probably will not yield valid inferences about such students' math achievement. Thus, the negative consequence of these students' obtaining low scores would be due to a source of invalidity in the test. Note, however, that this would not necessarily affect the inferences that could be made about the math achievement of students who are native English speakers; valid inferences may still be made on the basis of these students' scores.

As noted in an earlier section, many theorists have argued for the consideration of test consequences as part of test validity. Others, however, argue that the concept of validity is already sufficiently complex and that the inclusion of yet another aspect of validity overburdens the concept. For example, Mehrens (1997) stated that proponents of the inclusion of consequences under the validity heading apparently feel that "one should confound the results of using data in a decision-making process (which is what I think is what such individuals mean by consequential validity) with the accuracy of the inference about the amount of the characteristic an individual has" (p. 17). Mehrens notes that such a confounding of concepts may not be a good idea. He goes on to point out that consequences are, by definition, tied to a specific use of a test but that valid inferences can be made whether or not the test is used for some purpose. Thus, he argues that (some) test inferences may be valid even if negative consequences accrue from (some) test uses.

Others have argued that consequences are important but should not be included under the validity heading. These researchers have argued instead that terms such as *utility* (Lissitz & Samuelsen, 2007), *evaluation* (Shadish et al., 2002), *overall quality* (Borsboom et al., 2004), or *justification* (Cizek, 2012a) be used instead. Theorists such as Markus (2014) state that consequences have a bearing on test use whether or not they are included in one's definition of validity. If they are not included in one's definition of validity, consequences are simply discussed separately. Markus points out that inclusion of consequences in the validity definition helps to ensure that consequences are not overlooked or marginalized and makes it easier to link consequences to sources of test invalidity. Markus feels that exclusion of consequences from the validity definition would remove the investigation of consequences from the purview of those gathering validity evidence and make these the responsibility of others (such as schools or other organizations), which may or may not be a good thing.

This leads us to the final point of contention in the consequences-as-part-of-validity debate: if consequences are to be included as part of validity, who should be responsible for investigating these? Clearly, unintended consequences are more difficult to investigate than intended consequences for the simple reason that we do not necessarily know what the unintended consequences are. Kane (2013) argues that, because consequences are tied to test use, test users are in the best position to evaluate them. Test developers, however, often have greater technical expertise and more experience in testing issues, and so may be better at spotting potential consequences. Haertel (2013) suggests that those in academic disciplines such as sociology, anthropology, economics, and law, may, depending on the nature of the test, be able to help identify likely consequences. He suggests that testing experts work together with colleagues in these areas, as well as with interested stakeholders, to carry out studies of test consequences. The inclusion of stakeholders in the process is important because arguments about consequences depend on values, and for the argument to persuade stakeholders (e.g., parents, employers, health care professionals, members of the general public), the stakeholders must share the values of the test developers.

Intended Testing Consequences

Suppose I argue that using the GOSH test in graduate school admissions decisions will result in a greater likelihood of selecting students who will graduate within a specified time-frame. Such claims are common in testing because tests are typically used in the expectation that they will yield some type of benefit. These claims may be either explicit, as in my GOSH test claim, or implicit. Explicit claims can and should be evaluated to determine whether the anticipated benefits actually accrue, and if so, whether these can reasonably be attributed to use of the test. Simply showing that students with high GOSH scores graduate within a certain amount of time does not completely verify my claim because students with low GOSH scores might graduate in the same amount of time. To rule out this possibility, I would have to show empirically that students with low GOSH scores take longer to graduate. If high GOSH scores are required for admission to an institution, I may not be able to obtain graduation information from low GOSH scorers. However, I may be able to find a school in which the GOSH test is not used for admission, give the test to all the students, keep track of their times to graduation, and determine whether the expected pattern occurs. Or I could make a strong logical argument in which I show that the skills measured by the GOSH are the same as those used by successful graduate students.

In some cases, so-called *extra-test* claims are made. These are claims that go beyond the interpretations and uses of test scores specified by the test developers. For example, one argument for the use of performance assessments in lieu of paper-and-pencil tests was that performance assessments would be fairer tests of the knowledge of minority students. Unfortunately, there appears to be no evidence that this is the case (Linn, Baker, & Dunbar, 1991). As noted in the *Standards*, those making such claims are responsible for providing evidence that they actually accrue. Because such claims are outside the specified interpretations and uses of test scores, additional studies would likely be required to obtain evidence to support them. For example, if I were to claim that use of the GOSH test in graduate school admissions would enhance the reputation of graduate programs, I would need some data to back up this claim. Because data on people's perceptions of academic programs are not commonly collected, I would have to conduct research studies to gather this evidence.

Unintended Testing Consequences

As discussed earlier in some detail, the use of tests can have both intended and unintended consequences. For example, given the extended-answer format of the hypothetical GOSH test, it would likely be expensive and time consuming to score. If so, this may add to the cost of an admissions application for programs using the test. Admissions decisions might also be delayed because of the time needed to score the test. As in any use of testing, these negative consequences would have to be weighed against the anticipated positive consequences in making a decision about whether to use the test. A commonly discussed unintended consequence in educational and employment testing is the presence of score differences for groups defined by gender, race, ethnicity, age, or disability status. As noted in the *Standards*, "In such cases . . . 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, selection tests for positions as firefighters require applicants to carry heavy weights, and there may be members of some groups, such as women and older adults, who are not able to do so. However, the ability to carry heavy weights is clearly necessary for carrying out the job of a firefighter, so the fact that members of some groups are disproportionately represented does not imply any lack of validity of the test. But, in addition to physically fighting fires, firefighters are required to interact with the public. Suppose that women were found to be better at these interactions than men (not that I am suggesting this is the case). If a measure of public-interaction ability were not included in the test battery, fewer women might be hired as firefighters. This unintended consequence can be traced to construct underrepresentation—a source of invalidity—and is thus relevant to validity concerns.

Of course, unintended consequences are not always negative. For example, in many states, teachers are hired to score assessments given as part of the state's testing program. In this way, teachers meet other teachers from around the state and gain knowledge of their teaching practices. In addition, teachers obtain a better understanding of the state's testing program. Such consequences of the test, though not necessarily intended, are clearly positive. Therefore, in evaluating evidence based on consequences, all of the consequences, both positive and negative, must be weighed to make an overall judgment. One piece of evidence that should be included in such a calculation is the consequence of *not testing*. In the firefighter example, officials may decide that the consequences of testing are sufficiently onerous that the test should not be used. In that case, however, how would firefighters be selected? Should all applicants be hired on a trial basis, and those who are successful given permanent jobs? This option has a certain appeal but would not be cost effective because many more applicants than needed would be trained, at considerable expense. Another test battery could be developed, but this would also be expensive. Thus, as in any judgment, decisions regarding the use of tests must involve careful weighing of the pros and cons.

SUMMARY

The *Standards* (2014) suggest five types of validity evidence: evidence based on test content; evidence based on response processes; evidence based on internal structure; evidence based on relations to other variables; and evidence for consequences of testing. The basic validity argument underlying evidence based on test content is that the items on the test are appropriate for measuring the construct in terms of both content and cognitive level. Evidence for this can take the form of the match of items to the test blueprint, or table of specifications; expert reviews of the test content and of the cognitive processes required; and identification of any sources of construct irrelevance or construct underrepresentation. Evidence based on response processes should address the degree to which test items require respondents to use the cognitive processes that were intended. For example, if a respondent is able to adequately answer a critical thinking item by simply reciting factual information from memory, the intended cognitive process has not been evoked. Evidence based on response processes can be obtained from think-aloud protocols in which respondents are asked to verbalize their thoughts as they respond to items; tracking respondents' eye movements and/or response times as they answer questions; studies comparing the performance and response strategies of experts and novices; respondents' concept maps; and experimental studies in which item features thought to affect responding are systematically manipulated. Researchers should also specify the logic model that underlies the processes through which item responses are thought to lead to the desired inferences.

The validity argument underlying evidence based on internal structure is, broadly speaking, that the relations among items, and among items and subtests, mirror those expected from theory. Such evidence may take the form of item and/or subscale intercorrelations, internal consistency coefficients for items thought to form an identifiable scale, results from exploratory and confirmatory factor analyses, item response theory and/or generalizability theory, and DIF studies. Establishing evidence based on relations of test scores to other variables should be based on a theory explicating how and why test scores should relate to other variables. This theory should include information about the expected direction and strength of the hypothesized relations. The evidence can take many forms, such as correlations of test scores with the hypothesized variables, prediction of outcomes, differences among groups thought to differ on the construct being measured, or studies designed to reveal the extent, if any, of contamination resulting from method effects.

Finally, evidence of the consequences of testing should be reported whenever possible. Both positive and negative consequences of testing should be discussed, so that potential test users can make an informed decision about whether the positive are likely to outweigh the negative in a particular testing situation. Although researchers cannot be expected to anticipate every possible consequence, careful consideration of likely positive or negative consequences may bring to light possibilities that had not previously been thought of. Such an exercise might therefore be a valuable addition to the validation process. Evidence based on consequences can also be focused on the degree to which intended benefits of testing actually accrue. If unintended consequences are found, researchers should determine, to the degree possible, whether these are due to sources of test invalidity such as test irrelevance or construct underrepresentation. Negative consequences that are due to such sources undermine the validity of the test and suggest that the test should be modified to avoid these consequences. Alternatively, such negative consequences might alter the interpretations that can be made or the ways in which the test can be used.

EXERCISES

1. What, if anything, is problematic with the definition of validity as the degree to which a test "measures what it's supposed to measure"?
2. What is the "tripartite" view of measurement validity? What is one objection that has been raised against this view?

Questions 3–10 refer to the following hypothetical situation: A researcher has developed a test to measure problem-solving imagination (PSI), defined as "the ability to imagine new solutions to existing problems." The PSI consists of short descriptions of existing problems to which test takers respond by writing as many possible solutions as possible within a given time limit. These responses are scored by trained raters in three areas: (a) number of ideas, (b) creativity of ideas, and (c) likelihood of success of the ideas. PSI subscale scores are provided in each of the three areas, and an overall score, which is an equally weighted average of subscale scores, is also provided. The proposed interpretation of PSI total score is that higher scores indicate greater levels of problem-solving imagination. The PSI has been designed for use in research on imagination and problem solving. The test developer does not recommend use of the PSI for selection or classification decisions.

3. Could construct-irrelevant variance could be a threat to the validity of this test? Why or why not?
4. Could construct underrepresentation be a threat to the validity of this test? Why or why not?
5. As one form of validity evidence, the test developer showed that scores on the PSI were related to tests of creativity, problem solving, and imagination, as shown in Table 11.6. Is this evidence persuasive? Why or why not? What other information would be useful in evaluating this evidence?

TABLE 11.6. Correlation Values for Question 5

	Creativity	Problem solving	Imagination
Correlation of PSI with:	.60	.33	.65

6. The developer of the PSI would like to design a study to obtain validity evidence based on response processes. Give an example of this type of evidence that would be relevant to the PSI.

7. The developer of the PSI obtained correlations among the three subscale scores as evidence of validity based on internal structure. The correlations are shown in Table 11.7, with values of interrater agreement (nominal agreement) shown in parentheses on the diagonal.

TABLE 11.7. Correlations among PSI Subscores

	Number of ideas (Number)	Creativity of ideas (Creativity)	Likelihood of success of ideas (Success)
Number	(.95)		
Creativity	.68	(.60)	
Success	.25	.34	(.53)

   a. Do these correlations support the validity of PSI scores as measures of problem solving imagination? Why or why not?

   b. The correlations between Number and Success and between Creativity and Success are quite low. Why might this be?

8. The test developer did not provide any evidence showing that PSI scores were predictive of a particular outcome, such as success in gifted education programs. Is this problematic? Why or why not?

9. What are some possible consequences (either positive or negative) of using the PSI?

10. Suppose that the test developer decided that the PSI could be used to select students for special educational programs for gifted children. How, if at all, would this change your answer to Question 9?

Guilford Publications
370 Seventh Avenue, Suite 1200
New York, NY 10001
212-431-9800
800-365-7006
www.guilford.com