2

Validity in Testing and Psychometrics

The concept of validity originated and evolved within the fields of educational and psychological testing. In this chapter we will examine this evolution, with respect to how the concept has been applied to the interpretation and use of various kinds of tests.[1] In the first half of the 20th century validity was discussed and debated predominantly within a psychometric framework, relating to how the development and use of tests could be improved. In time it began to be applied in other contexts of social science inquiry, especially the assessment of research quality (covered in Chapter 3), but it continued to evolve within psychometrics as well.

In the psychometric uses of the concept of validity, one can see the priorities that have been applied to the term in all of its contexts. The specification and determination of validity are concerned with an appeal to truth and accuracy, and that orientation certainly describes its origins in psychometrics. As Angoff (1988, p. 19) wrote: "Validity has always been regarded as the most fundamental and important [*concept*] in psychometrics."

Many excellent texts on psychometrics discuss validity in detail (e.g., Bandalos, 2018; Cronbach, 1990; Urbina, 2004). Our purpose in this chapter is more selective: we discuss the psychometric view in a way that prepares for our discussion later in the book. Given that the point of validity is to assess the truth and accuracy of the use of a test, how has the concept needed to evolve in order to accommodate the steadily growing sophistication of test theory? Further, how can conceptions of validity take the leap

[1] In this chapter I use the term "test" as a shorthand to refer to a variety of measurement strategies, to which the assessment of validity can be applied. Thus a measure could consist of behavioral observations, observer judgments, etc., as well as a conventional test.

into other elements of social science methodology? We analyze how the "truth function" of validity has evolved to incorporate modern perspectives on testing and psychometrics.

Validity in the Psychometric Tradition

Beginning in the 1920s, validity was frequently defined as "the extent to which a test measures what it purports to measure" (Urbina, 2004, p. 154). That definition, although widely adopted and repeated in many textbooks, has long been abandoned by psychometricians as too simplistic. The scope of what *validity* should encompass—and, by extension, its definition—have been and continue to be debated, especially in light of current conceptions that view validity as complex and multidimensional. Related questions involve how validity should be assessed and what that assessment should include. The conceptualization has been provided some coherence by a consensus of social science organizations. A consortium of three leading societies—The American Educational Research Association, The American Psychological Association, and The National Council on Measurement in Education—first issued a joint statement of standards for educational and psychological testing in 1954 (AERA et al., 1954). Those standards have been revised and updated approximately every decade since.

The 2014 *Standards* define validity as follows: "Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA et al., 2014, p. 11). Several aspects of this definition deserve attention. First, it squarely consigns validity as a property of test score *interpretations* rather than as a static property of the test itself. Thus, it would be incorrect to make a single, initial determination of a test's validity and then accept it as a permanent feature of the test. Second, the definition also introduces the consideration of how the test will be used. This can include, for example, selection into a restricted program (e.g., entry into a program for students with special skills or talents, or conversely, into a remedial program for additional academic assistance), individual diagnosis of a student's current strengths and weaknesses (as in a classroom test of academic content), assessment of an individual's personality profile (e.g., for online matching of potential romantic partners), and innumerable others. By this definition, test validity will depend in part on the appropriateness of the decisions that will follow, which in turn are based on the interpretation of scores.

Evolution of the Conceptualization of Validity

Given the multitude of interpretations, uses, and functions of tests, it will not be surprising that the concept of validity—which targets the truth and accuracy of test score interpretations—would take on a wide variety of forms. Early conceptions of test validity, beginning in the 1920s, were based on the examination of correlational evidence, i.e., whether the test correlated with some designated criterion (Sireci, 2009). The correlation coefficient had been introduced by Karl Pearson at the turn of the century, and test validity was considered to be a practical application of that new statistical procedure. Sireci cites a 1946 quote from the psychologist J. P. Guilford that "a test is valid for anything with which it correlates."

The correlational approach was soon supplemented by a second technique, factor analysis, for other types of questions about tests. Factor analysis had also recently been introduced, by Charles Spearman, and it provided a completely different approach to determining the qualities of a test. Whereas correlational techniques focused on whether a test provided similar information to that from some other measure or criterion, factor analysis examined the internal structure of the test items.

For a variety of reasons, these early conceptions of validity and the corresponding methods of test validation were eventually judged to be inadequate and limited, and they came under increasing criticism by psychometricians. For example, with regard to the correlational approach, there were no standard expectations for how substantial a correlation needed to be before it could be considered as sufficient evidence to use a test in a particular way. This growing frustration led to the development of the Standards.

Introduction of the Standards

In the early 1950s, the American Psychological Association developed a proposal for common standards for the development and interpretation of psychological tests and measures. This led to the formation of the joint committee, which published its Standards in 1954. This document proposed four different types of test validity: content, concurrent, predictive, and construct. In later revisions through 1985, these were shortened to three categories—criterion-related, content, and construct validity—which form the basis of what is called the tripartite view of validity.

Following that initial appearance, revisions of the Standards were published in 1966, 1974, 1985, 1999, and 2014. Beginning with the 1985 edition, the Standards take a sharply different approach to validity. Rather than describing distinct types, it proposes that validity is a unitary phenomenon, and it describes five different kinds of validity evidence, as we will detail later.

Tests and Their Constructs

For a term that is used so frequently and with such import for the theory and practice of validity, the term *construct* is often used with little attempt to specify exactly what it means. Table 2.1 provides a summary of some of the definitions and explanations that have been offered for the term.

TABLE 2.1 ● What Is a *Construct*? Some Variations

Source	Definition
Cronbach and Meehl (1955)	"A construct is some postulated attribute of people, assumed to be reflected in test performance. In test validation the attribute about which we make statements in interpreting a test is a construct." (p. 283)
Shadish, Cook, and Campbell (2002)	"A concept, model, or schematic idea." (p. 506)
Anastasi (1986)	"Let us consider the nature of the constructs employed in test development. Essentially they are theoretical concepts of varying degrees of abstraction and generalizability which facilitate the understanding of empirical data." (p. 5)
Embretson (1983)	"Here, *construct* refers to a theoretical variable that may or may not be a source of individual differences." (p. 180)
AERA et al. (2014)	"The term *construct* is used in the *Standards* to refer to the concept or characteristic that a test is designed to measure." (p. 11)
Messick (1981)	"Constructs thus provide organized interpretations of observed behaviors as well as a means of predicting previously unobserved behavioral consistencies from the theoretical implications of the nomological network." (p. 580)
Borsboom et al. (2009)	"[W]e do not know what constructs are, that is, we have rarely come across a clear description of what something should be like in order to deserve the label 'construct.' Constructs, as far as we are concerned, are truly shrouded in mystery,...in the sense that we don't really know what we are talking about in the first place." (p. 150)

In many cases, especially in years past, *construct* has been used to refer to latent traits and entities that don't have immediate and objective representations in the real world. Examples would be abstract concepts such as intelligence, sociability, aggression, attractiveness, or need for achievement. This was the approach originally taken by Cronbach and Meehl (1955): "A construct is some postulated attribute of people, assumed to be reflected in test performance" (p. 283). Their focus on constructs as latent traits was understandable since their paper, and their original formulation, was specifically focused on the domain of psychological testing. Thus the examples they cited involved trait-like characteristics: "The constructs in which tests are to be interpreted are certainly not likely to be physiological. Most often they will be traits such as 'latent hostility' or 'variable in mood,' or descriptions in terms of an educational objective, as 'ability to plan experiments.'" (p. 284) Cronbach and Meehl theorized that the focal construct of a psychological test should be supported by a body of theory and evidence, which they called a nomological network. In their formulation, the process of test validation consisted of identifying the nomological network and confirming its support of the construct.

This approach to terminology was more prominent in prior years when construct validity was seen as just one of several validity types, to be applied in some test settings but not others. In those cases it was invoked when the focus of testing was on an abstract and unobservable characteristic. With validity now being viewed as a unitary concept rather than a collection of types, and construct validity no longer seen as that specialized kind of validity that is invoked when there is a lot of debate about what exactly is being measured, the term *construct*—at least with regard to validity theory—is usually given a more universal role, reflecting the particular focus of any test that is under consideration. The 2014 Standards provide the following description: "The proposed interpretation [*of test scores*] includes specifying the construct the test is intended to measure. The term *construct* is used in the *Standards* to refer to the concept or characteristic that a test is designed to measure....Examples of constructs currently used in assessment include mathematics achievement, general cognitive ability, racial identity attitudes, depression, and self-esteem." (p. 11) Within this list, the inclusion of mathematics achievement—however hard it is to define—is noteworthy in its complete departure from the realm of latent psychological traits. Every test has its construct, whether that construct is steeped in psychological theory or behaviorally defined. There is no attempt to restrict the term to unobservable variables.

The psychometric theorist Anne Anastasi provided this elaboration:

> [*Constructs*] are ultimately derived from empirically observed behavioral consistencies, and they are identified and defined through a network of observed interrelationships. In the description of individual behavior, such a construct corresponds closely to what is generally termed a trait. A simple example, with narrowly limited generalizability, is speed of walking. If we take repeated measurements of an individual's walking speed, we still obtain a whole distribution of speeds...Nevertheless, it is likely that an analysis of such varied measures would reveal a substantial common factor that reliably differentiates one person from another in overall walking speed. This common factor would be a construct; it does not necessarily correspond to any single empirical measure. (Anastasi, 1986, pp. 4–5)

Some constructs are highly general while others are more specific. The critical test is functionality. Cronbach (1990) noted:

> A formal construct is invoked when inference reaches out to diverse situations. “Musical talent” is a more convenient dimension than “talent for stringed instruments.” That, in turn, is handier but less definitive than “dexterity in rapid finger movements” and “pitch discrimination.” The broad construct is neither true nor false; it is adequate for some purposes and inadequate for others. (p. 52)

Some constructs invite definitional inconsistency not because of a theoretical debate about underlying components or a nomological network, but simply varied options for definition. However, it is important to conceive of the construct as existing apart from the test itself, such that the test is an attempt to capture it. In years past it was occasionally suggested that the construct and its test were identical, a concept known as *operationism*—famously illustrated in an early quote about intelligence by the American psychologist Edwin Boring: “...intelligence as a measurable capacity must at the start be defined as the capacity to do well in an intelligence test. Intelligence is what the tests test” (Boring, 1923). But operationism in measurement is now widely rejected.

The Traditional, Tripartite View of Validity

Until approximately the 1980s, the dominant view of validity was that it was a multifaceted concept comprising a collection of subtypes. The Standards adopted this view as well, beginning with their initial formulation in 1954 and continuing until the third revision in 1985, although the taxonomy varied to some degree over the revisions. For the most part, the major categories of validity were labeled as *criterion-related*, *content*, and *construct*.

Criterion-Related Validity

Criterion-related validity involves correlational approaches, and has been often conceived to encompass two subtypes, concurrent and predictive. As the names imply, *concurrent* validity refers to the degree to which the target test aligns with criterion variables measured more or less simultaneously, while *predictive* validity refers to the test’s prediction of future events, such as success at a job, or its alignment with test scores obtained at a later point in time. In later versions of the Standards these were combined under the single heading of *criterion-related* validity. What the two types share in common is a reliance on empirical evidence from specific identified sources to arrive at a judgment about the test’s validity status. This was historically the dominant view of what *validity* means, through approximately the 1950s.

For example, a newly developed measure of some construct of interest may be introduced into a crowded field in which other measures already exist. The supposed advantage of the new measure might involve a gain in practicality rather than improved truth or accuracy. Thus it may be shorter or less psychologically sensitive than existing measures. For example, the Rosenberg Self-Esteem Scale (Hyland et al., 2014) consists of only 10 items, considerably shorter than other available self-esteem measures, and is the most widely used scale for measuring that construct. As a different kind of advantage, it may be that the new test is designed for group-based administration whereas prior measures require individual administration by a trained tester. Or the new test might be less expensive to implement due to paper-and-pencil self-administration or savings on proprietary copyrights. For these or other reasons, the new test of a construct might be welcome even if other measures already exist. If the test is shown to correlate with one or more measures to some anticipated degree, that finding provides evidence for the test's validity, that is, the appropriateness of interpreting its scores as a measure of the construct in question. However, the level of that association with the criterion measures will be carefully considered and may be a point of contention among critics.

Concurrent and predictive validity typically have somewhat distinct purposes, despite their underlying similarities in perspective and approach. Predictive validity has an intuitively appealing functional clarity—specifically, the prediction of future status—that makes it well-suited for purposes of selection, such as hiring from among job applicants or colleges' selection of incoming students. By contrast, the classic purpose of concurrent validity is to make comparisons *between* tests. As noted, if an existing, well-established test is expensive or time-consuming to administer, a newly developed test can be used in its place, provided that a sufficient degree of equivalence or compatibility can be demonstrated. Thus the theoretical underpinnings of concurrent and predictive validity—that is, the assumptions about what "validity" means and the type of evidence it depends on—are similar and congruent, but the two types have distinct niches with regard to their purposes and uses.

Content Validity

Content validity makes no claims about correlational relationships, but rather addresses whether the test adequately captures, represents, or samples the universe of content that the test has been developed to measure. Content validity is most relevant for tests that are designed to assess the attainment of skill or mastery in some domain, and thus it is particularly applicable to educational achievement tests, which may be focused on knowledge acquisition (e.g., vocabulary), cognitive skills (e.g., long division), or motor skills (e.g., keyboard skills). Outside of the educational domain, content validity is also relevant for uses of tests that involve selection.

The assessment of content validity does not involve measuring congruence with an outside criterion, but rather focuses on identifying the hypothesized content domain and determining how well it is represented by the specific content of the test. The test represents a sampling from that domain, and therefore the content universe must be carefully defined and mapped. Once this is done, one can determine whether that universe is represented appropriately, with an appropriate balance of elements.

Another aspect of content validation is the determination that the cognitive processes required to answer the test questions, whatever they may be, are relevant to the abilities about which judgments are being made. For example, if a test item is intended to assess mathematical reasoning in answering a complex problem, producing the correct answer should require the replication of those reasoning processes. That process is subverted for a respondent who happens to remember the answer without needing to go through the reasoning.

These two sets of criteria for content validity—adequate sampling of the test items from a content domain and the requirement of engaging in specified response processes to arrive at the correct answer—are often assessed through expert appraisal, such as in the form of an invited panel of professionals with strong expertise in the relevant content domain. In sum, content validity focuses on the substance, selection, and composition of the test questions rather than agreement with an external comparator.

Construct Validity

This final form of validity in the conventional tripartite model was a new concept when it was proposed in the first edition of the Standards in 1954. Lee Cronbach and Paul Meehl, who were both members of the joint committee that produced the Standards (with Cronbach as chair), followed up the next year with a paper that has become one of the seminal articles in the history of validity theory (Cronbach & Meehl, 1955).

Construct validity refers to how well the test represents and measures an individual's status on the construct that comprises the focus of the test. Cronbach and Meehl did not identify a limited set of processes that could be used to establish construct validity for the uses of a test. Instead, they postulated that the validation process for construct validity requires the researcher to demonstrate that the test scores are consistent with the nomological network that already exists for the construct. For example, if a test is developed to measure a psychological trait such as sensation seeking (Zuckerman, 2007), construct validity will be confirmed if the test scores correspond to other measures in a manner consistent with the theoretical formulation. Another reason for the relative elusiveness of construct validity has been the sheer scope of the concept. Angoff (1988) wrote: "...we can see that construct validity as conceived by Cronbach and Meehl cannot be expressed in a single coefficient. Construct validation is a process, not a procedure; and it requires many lines of evidence, not all of them quantitative." (p. 26). Messick stated: "In its simplest terms, construct validity is the evidential basis for score interpretation. As an integration of evidence for score meaning, it applies to any score interpretation—not just those involving so-called 'theoretical constructs.'" (Messick, 1995, p. 743).

Limitations of the Tripartite View

Over time, the idea that there are different varieties of validity came to be seen as problematic. Despite the ubiquity of the tripartite view through the 1980s, the field of psychometrics was moving to abandon it. Lee Cronbach wrote, in 1988: "The 30-year-old idea of three types of validity, separate but maybe equal, is an idea whose time has gone" (Cronbach, 1988, p. 4).

A large part of that problem was in the pragmatic applications of the concept. Test developers tended to choose one or another of the subtypes for their validation studies, and then consider the case closed. Critics also noted that test developers, in their validation studies, tended to collect data that were most available and accessible—that is, easiest to collect—rather than data that conformed with a conception of what the test is designed to do, which would lead to a determination of which type of validity would be most appropriate to pursue. Another issue was that the validity types often overlapped, which made labels of construct, content, and criterion validity somewhat arbitrary.

The unit within the tripartite view that has the most in common with scientific inquiry, in general, is construct validity. Due to this perspective, many theorists had been moving toward a view that construct validity encompasses the other forms. The prominence of construct validity continued to rise in the second half of the 20th century, and because of the perceived similarity of construct validation to the overall research process, many psychometricians came to see it as central and essential to all forms of validity. For example, Zumbo (2009) wrote:

> Although it has been controversial, one of the current themes in validity theory is that construct validity is the totality of validity theory and that its demonstration is comprehensive, integrative, and evidence-based. What becomes evident is that the meaning of "construct validity" itself has changed over the years and is being used in a variety of ways in the current literature. Arguably in its most common current use, construct validity refers to the degree to which inferences can be made legitimately from the observed scores to the

> theoretical constructs about which these observations are supposed to contain information. (p. 68)

Cronbach himself maintained this perspective as well, writing in the final edition of his textbook on psychological testing: "The end goal of validation being explanation and understanding, construct validation is of greatest long-run importance" (Cronbach, 1990, p. 152).

The view that "all validity is based on construct validity" presaged the unitary theory of validity, and made it a short conceptual jump to the view that validity itself cannot be divided into subtypes.

The Current View: Validity as a Unitary Phenomenon

Samuel Messick (1989, 1995) provided the most comprehensive rejection of the traditional view, advancing the idea that validity is a unitary concept. He wrote: "One or another of these forms of evidence, or combinations thereof, have in the past been accorded special status as a so-called 'type of validity.' But because all of these forms of evidence fundamentally bear on the valid interpretation and use of scores, it is not a type of validity but the relation between the evidence and the inferences drawn that should determine the validation focus. The varieties of evidence are not alternatives but rather supplements to one another. This is the main reason that validity is now recognized as a unitary concept." (Messick, 1989, p. 16).

Consistent with this view, the current conception of validity presented in the 2014 Standards does not enumerate separate kinds of validity. What used to be the different forms—content, predictive, concurrent, and construct—have been reconfigured to represent different *forms of validity evidence* (AERA et al., 2014).

Approaches to the Process of Test Validation

Traditionally, there have been several well-established approaches for assessing the validity of a test or measure. The approach that is most appropriate in a particular instance depends on the purpose of the test and the theoretical underpinnings of its target construct(s). Thus, for example, predictive and concurrent validity (both tied theoretically to external criteria) would typically be validated through correlational approaches: investigations would determine whether the test could predict the relevant event, or, alternatively, whether it would correlate with other tests that purport to measure the same construct. Tests that were intended to reflect a psychological construct would be subjected to factor analytic studies to determine the loadings of the items. Tests that purported to sample a universe of academic content would be subjected to content validity studies.

With the ascendance among psychometricians of the unified view of validity, many of these same approaches are still used, but they are considered to be simply different kinds of evidence rather than different phenomena. The 2014 Standards identifies five specific types of validity evidence, which, to some degree, are traceable back as restatements of the tripartite forms of validity.

1. **Evidence based on test content.** Consistent with earlier conceptions of content validity, this form of validity evidence is particularly suited to tests that are hypothesized to sample from a universe or domain of content. This includes educational achievement tests, which often sample from a body of knowledge or desired academic skills, as well as personnel selection tests, which sample from a domain of high-priority job performance skills. Rather than demonstrating the test's agreement with an external criterion or confirming its theoretical structure, content analysis seeks to demonstrate that the items comprising the test are a suitable representation of the larger domain of content from which it is drawn. This may be true for an area of subject matter knowledge or a type of skill set, such as various kinds of mathematical operations, the knowledge needed to be a lawyer in a particular state, and so on.

   The processes of content analysis often take the form of review by an expert panel, which can pass judgment on considerations such as the representativeness of the test content with regard to the content domain and the priority and relative weight of components of the test. A second approach is to develop a table of specifications (Bandalos, 2018), which details the scope and focus of the knowledge domain being measured, and, thus, what materials the test must sample. This table is then used to generate the specific content of the test, and the test content can be mapped onto the specifications.

   Two important concepts here are *construct-irrelevant variance* and *construct underrepresentation*, both of which are content-related threats to validity. Construct-irrelevant variance refers to differences between test scores that are due to something other than differences on the construct in question. As illustration, consider a carelessly constructed multiple choice test item. An alert student might recognize that one response option is longer and more detailed, or more carefully conditional, than the other options, and correctly infer that this option is correct. In that case, differences in students' scores might be due in part to their familiarity with the specific type of test, rather than whatever knowledge is presumably being tested.

Construct underrepresentation refers to a situation in which the construct is not fully represented. For example, a standardized test of math skills might include items related to arithmetic operations and fractions, but omit items related to decimals. In such cases score differences could indeed be due to differences in mastery of the construct, but if components of the construct are fully included in appropriate proportions, those differences will be more accurate, and possibly either larger or smaller.

2. **Evidence based on response processes.** This category of evidence refers to investigations of how the test takers answer the questions on the test, and is affirmed if it can be demonstrated that the processes necessary for arriving at correct answers on the test are those that are hypothesized for the target construct. For this category, reference to the underlying construct is key to providing evidence. For example, the determination of an answer to a question about long division should require going through the division process. The validation process involves task decomposition, that is, "an examination of test responses from the point of view of the processes, strategies, and knowledge stores involved in their performance" (Urbina, 2004, p. 159).

3. **Evidence based on internal structure.** This form of evidence relates to the relationships among the individual items in the test and the test's focal construct. The response patterns for the test items must demonstrate an internal organization or structure that conforms with the prediction of the construct. This evidence is based most commonly on some form of factor analysis, especially confirmatory factor analysis. The factor loadings, i.e., the correlations between the set of test items and the underlying dimension(s) of the test, can be examined to demonstrate evidence in support of the theory of the construct. Thus, if a construct is hypothesized to be unidimensional, a factor analysis should indicate that a single factor contributes to test scores. If the construct is presumed to be multi-dimensional, the factor structure would align with the theory underlying the construct, in terms of the number of factors (or subdimensions of the construct) as well as the items that load on each of those. Other kinds of analysis that demonstrate evidence based on internal structure include patterns of item difficulty and item response theory.

4. **Evidence based on relations to other variables.** Conforming with the older concept of criterion-related validity, this category of evidence for validation reflects the perspective that a test designed to measure a particular construct needs to demonstrate agreement with other indicators of that construct, which can include both other tests and different kinds of criteria. This form of evidence is correlational, examining whether the test takers' scores on the test conform with their status on other criteria relevant to the test's construct. A key distinction is whether we are talking about evidence that is predictive or concurrent, conforming with the traditional division within criterion-related validity that was presented earlier.

    Predictive evidence refers to the test's ability to predict test takers' status on some criterion that is measured at a later point in the future. This present-to-future relationship is the essence of a selection test. For example, the SAT test is designed to be administered to high school students with the purpose of predicting their future success in college, and has been used by colleges and universities as a part of the admissions process for many decades. Accordingly, for this test, the most direct form of validity evidence would be any of several indicators of later college success, such as grade point average, college graduation, or first year retention versus dropout. However, the SAT has been criticized as reflecting cultural and economic bias in its test scores and its susceptibility to test preparation courses that are more accessible to students from affluent families (e.g., Soares, 2020). The SAT has both detractors and supporters, but these concerns would constitute a threat to the test's validity based on the fifth form of evidence—the uses of test scores—discussed below.

    Concurrent evidence, by contrast, refers to the test's agreement with other measures that are assessed at approximately the same time as the test in question. For example, a test that is measuring a personality-related construct would be expected to correlate sufficiently well with other existing tests that purport to assess the same construct.

    Despite the emphasis on correlation in the description of criterion-related validity, some diversity exists in the potential methodological approaches that they employ, and the connection to "correlation" should not be interpreted as referring solely to a reliance on the correlation or regression coefficient. Besides straightforward correlational evidence based on criterion variables, the validation process can involve investigations of whether the test can distinguish between existing groups that would be expected to vary on the construct being measured. For example, a psychological test, e.g., of obsessive compulsive disorder, could be administered to a sample of known OCD individuals in comparison to a general population sample. The statistical analyses in these studies could involve t-tests or other tests of group differences. However, in this last case the conception of criterion-related validity begins to overlap with that of construct validity. Such ambiguity is one of the factors that led psychometricians to develop the view of validity as a unitary concept.

Donald Campbell, although best known for his contributions to the theory of validity in relation to experimental design (see Chapter 3), made a significant contribution to psychometric validity in a collaboration with Donald Fiske (Campbell & Fiske, 1959). Their procedure, called the multitrait-multimethod matrix, involves developing a pattern of correlational results based on other measures, some of which are hypothesized to measure the same construct as the target test and some of which are hypothesized to measure something else. In addition, the measures can differ from each other in their use of different methodologies, e.g., survey self-report, self-ratings, ratings by others, or essay examinations. A case for validity of the target test is demonstrated if the highest correlations exist for the measures of the same construct using similar methods, while measures of the same construct using different measures would be expected to display a somewhat lower level of correlation. Measures of different constructs using different measures should display the lowest correlations.

5. **Evidence for consequences of testing.** This form of evidence, introduced by Messick (1995), is the most recent of the five types. It refers to whether the uses of the test's scores—that is, the test's consequences—are appropriate. If it can be demonstrated that test-based decisions and other consequences of using the test have value compared to not using it (e.g., in producing more accurate hiring decisions), that will constitute evidence that these particular uses of the test are valid. It must also be determined that the uses and consequences of the test are fair, equitable, and free from bias to the extent possible.

   This criterion is particularly relevant for tests that are used for purposes of selection, into either academic programs or employment settings. Thus this form of validity evidence is directly relevant to considerations of social justice in test use. As an example, consider again the SAT. Given the criticisms and extended debate about socioeconomic and cultural bias in the use of this test for college admissions decisions, an increasing number of colleges have dropped it as a requirement of the application process, either making it optional or eliminating it entirely.

The Argument Perspective on Validity in Measurement

Cronbach (1988) introduced the perspective that test validation should be viewed explicitly as a process of argument, rather than one that should adhere strictly to the procedures and protocols of scientific research. He noted, "Validation speaks to a diverse and potentially critical audience; therefore, *the argument must link concepts, evidence, social and personal consequences, and values*." (p. 4, italics in original). He was making several points. First, the validation process must anticipate, and be responsive to, the perspectives of the audiences that receive the information. Further, it should incorporate elements of argument and persuasion that go beyond the presentation of empirical evidence. The argument perspective was offered in contrast to the perspective of scientific inquiry. Finally, the process must be able to accommodate and incorporate uncertainty with regard to the judgment. "'What work is required to validate a test interpretation?' That question, with its hint that we are after a 'thumbs up/ thumbs down' verdict, I now regard as shortsighted and unanswerable" (Cronbach, 1988, p. 4). His chapter went on to identify five specific perspectives that should be accommodated in the validation argument, which he labeled the *functional*, the *political*, the *operationist*, the *economic*, and the *explanatory* perspectives.

This view goes beyond what many other theorists were saying about validity, but it is compatible with the unitary view, and the central role of construct validity, because it speaks to the need for a wide array of types of evidence and the conception of validation as an ongoing process. In presenting his perspective Cronbach cited contemporary thinking in the field of evaluation as a model, particularly Ernest House. Cronbach wrote: "Validation of a test or test use *is* evaluation.., so I propose here to extend to all testing the lessons from program evaluation. What House (1977) has called 'the logic of evaluation argument' applies, and I invite you to think of 'validity argument' rather than 'validation research'." (p. 4). The argument view was taken up by Michael Kane (2013), who has become its most prominent proponent.

The perspective among some psychometricians that test validation is a process of argument opens a bridge to applying concepts of validity to other topic areas within the field of evaluation, including, e.g., recent scholarship on evaluative thinking (Buckley et al., 2015).

In this chapter we have reviewed the origins of validity theory in the psychometric tradition, and the evolution of the concept over the past century. A quote here from Messick (1989) is apropos: "Validity always refers to the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores." (p. 13) In this book I try to represent that principle, with regard to other kinds of judgments. If we replace the term "test scores" with "elements of the evaluation plan," we have a concise summary of one of the themes of this book.

In the next chapter we will examine how the concept of validity was borrowed and extended beyond test scores to apply to an entirely different set of considerations: the adequacy of research, experimental designs, and evaluation studies.

Chapter Summary

- The concept of validity originated and evolved within the fields of educational and psychological testing. A consortium of professional societies—The American Educational Research Association, The American Psychological Association, and The National Council on Measurement in Education—first issued a joint statement of standards for educational and psychological testing in 1954 and have updated that statement several times since, most recently (as of this writing) in 2014.
- Validity refers to the truth and accuracy of test score interpretations. It is a property of those interpretations rather than a static property of the test itself. Validity also takes account of how test scores will be used.
- According to the 2014 Standards, a *construct* is "the concept or characteristic that a test is designed to measure."
- The traditional, tripartite view of validity recognizes three validity subtypes that are used to understand the accuracy of a test:
    - Criterion-related validity involves correlational approaches. It can refer to the test's correlation with criterion variables that are measured at the same time (concurrent validity) or in the future (predictive validity).
    - Content validity addresses whether the test adequately represents the universe of content that it has been designed to measure. Content validity is particularly appropriate for tests of skill or mastery.

  - Construct validity refers to how well the test represents the individual's status on the construct that comprises the focus of the test.

- The modern view of validity rejects the idea of distinct validity subtypes and views validity as a unitary phenomenon. What used to be considered different forms of validity are now seen as different forms of evidence that supplement each other rather than serving as alternative approaches. Five specific types of validity evidence have been identified:
  - Evidence based on test content
  - Evidence based on response processes
  - Evidence based on internal structure
  - Evidence based on relations to other variables
  - Evidence for consequences of testing
- Lee Cronbach advanced the idea that test validation should be viewed as a process of argument rather than purely a process of empirical research. This view is consistent with Ernest House's view of evaluation as a process of argument, which is expanded upon in later chapters.

Questions for Reflection—Chapter 2

1. In your judgment, what are some constructs for which it would be relatively straightforward to develop measures? What are some constructs for which the development of measures would be particularly complex?
2. From your own professional experience, identify a measure that you have used or that you are familiar with. What is the construct that is addressed by this measure?
   - If you needed to gather evidence to demonstrate the construct validity of this measure, how would you go about the task?
3. As this chapter describes, testing theorists now generally view validity as a unitary phenomenon, which can be examined using different kinds of evidence. This replaces the older tripartite view, which tended to associate different kinds of evidence with different kinds of validity. What were the shortcomings of the tripartite view that led it to be eventually abandoned?