Chapter 1

WHAT IS ARGUMENT-BASED VALIDITY?

The title of this book includes both the terms *testing* and *assessment* because the basic concepts of argument-based validity apply to the full range of activities encompassed by testing and assessment. Tests and assessments are used to make inferences about people's capacities on the basis of a sample of their performance. In testing and assessment, the inferences are used for such purposes as placing students into classes, drawing conclusions about learning, diagnosing specific challenges, judging candidates' performance adequacy for a job, making decisions about university admissions, and certifying qualifications. Tests and assessments are widely used by educators, human resources personnel, and researchers in education, government, health professions, and business, for example. In these varying contexts, some users of tests and assessments favor one term over another, but in this book no conceptual distinction is intended as both terms are used to refer to the same process of using systematically gathered samples of performance, summarized as scores, to make inferences about human capacities from which conclusions are drawn. The professionals responsible for all facets of testing and assessment are referred to as "testers."

The central concern for testers is validity, and therefore the meaning of validity and how to conduct validation research are ongoing topics of discussion and debate in the field. As academic discussion continues, testers need to meet the many demands of society for tests that can help in making a range of decisions. Since 1954, this need has been addressed, in part, by the consensus about validity and validation research expressed in the *Standards for Educational and Psychological Testing*, referred to throughout the book as the *Standards* (American Educational Research Association [AERA], American Psychological Association [APA], and the National Council on Measurement in Education [NCME], 2014). The *Standards* was developed and is periodically revised by the American professional associations directly concerned with testing: the AERA, the APA, and the NCME (Plake & Wise, 2014). Members of these organizations include the theorists, researchers, and practitioners who formulate conceptual and methodological approaches for testing and assessment that are applied across subject areas in and beyond North America. Because of the wide use of the *Standards*, comments considered in the most recent revision came from others concerned with testing and assessment, including other

professional associations (e.g., American Counseling Association and National Association of School Psychologists), testing companies (e.g., ACT and Pearson), academic and research institutions (e.g., Human Resources Research Organization), credentialing organizations (e.g., National Board of Medical Examiners), and other institutions (e.g., Fair Assess Coalition on Testing). This chapter introduces argument-based validity as a means for implementing the validity guidelines in the *Standards* and sketches the evolution of concepts about validity that have informed both the general guidance in the *Standards* and the specific conventions of argument-based validity.

Introducing Argument-Based Validity

Argument-based validity, as formulated primarily by Kane (1992, 2006, 2013), provides the conceptual tools needed to carry out the guidance in the *Standards*. The *Standards* defines validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of test scores" (AERA et al., 2014, p. 1). This definition expressed by professionals is different from the common sense notion that validity refers to tests themselves and that tests can, therefore, be either valid or invalid. According to the *Standards*, "statements about validity should refer to particular interpretations for specified uses," and "it is incorrect to use the unqualified phrase 'the validity of the test'" (p. 1). Argument-based validity provides a means for defining the interpretations and uses of test results so that the intended interpretations and uses can be validated.

Another commonly held perception is that validation research is carried out by calculating a correlation between sets of scores on two tests. In contrast, the *Standards* alludes to a more complex process for doing validation that begins with "an explicit statement of the proposed interpretation of the test scores, along with a rationale for the relevance of the interpretation to the proposed use" (AERA et al., 2014, p. 1). In addition, propositions supporting the proposed interpretations need to be identified, and then "one can proceed with validation by obtaining empirical evidence, examining relevant literature, and/or conducting logical analyses to evaluate each of the propositions" (p. 1). Specifically, the *Standards* names five types of evidence that can be used to investigate validity: evidence based on rationales and expert judgment of test content, evidence based on the study of test takers' response processes, evidence based on statistical testing of the internal structure of response data, evidence based on relationships to other variables (including convergent and discriminate

evidence), and evidence about the consequences of testing (*Standards*, pp. 13–21). These five types of evidence are intended to be integrated to make a professional judgment about validity. Overall, the *Standards* treats validation as a process of scientific hypothesis testing consisting of formulating propositions and evaluating their plausibility in view of empirical data. It also includes expert judgment and theoretical rationales in the validation process.

However, the *Standards* is not a methodology book with details about how to design a program of validation research and references to academic sources supporting its guidance. The argument-based approach to validation provides testers with a framework for conceptualizing the complex validation process suggested in the *Standards*, concepts and procedures for designing validation programs to yield the evidence called for by the *Standards*, and a common language for communicating within and across testing programs about the meaning of research results for the validity of test interpretation and use. Argument-based validation is not a single method yielding one type of validity evidence. Instead, argument-based validation encompasses a research program consisting of activities whose findings need to be integrated into a logical conclusion about the validity of test score interpretations for particular uses. Like the *Standards*, argument-based validation has its roots in an academic tradition of more than 100 years.

The Academic Tradition of Validity

The *Standards* does not make reference to the academic literature on validity, but the consensus views expressed in each successive version reflect the contemporary concepts, practices, and values of researchers in educational and psychological testing. These technical foundations have been conceived and refined over the past century (Kane, 2013; Messick, 1989; Shepard, 1993; Sireci, 2009). Playing a key role in this historical evolution have been the multiple editions of *Educational Measurement*, an authoritative edited volume published first in 1951 and then updated three times, in 1971, 1989, and 2006. Each volume contains a chapter on validity and validation research from one author's viewpoint and serves as a catalyst for discussion, research, and practice, which in turn influence the following edition of the *Standards*.

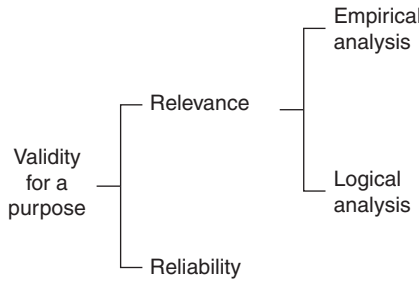
These chapters have proved to be influential because they provide useful snapshots of a dynamic evolution of concepts that remain important for testing today, and therefore build the background for

argument-based validity. The four chapters in successive editions of *Educational Measurement* have been analyzed from a philosophical perspective to show how they reflect the evolution that has taken place across the social sciences (Markus & Borsboom, 2013). Such an analysis emphasizes change and disjuncture. What is needed to work with validity arguments in practice today is an understanding of the basic concepts in testing that the chapters introduce and how the roles of these concepts have shifted in validation research. Most central is the evolution in the conceptualization of what gets validated, which has shifted from the idea that the test itself is validated to the statement in the *Standards* today that interpretations and uses of a test need to be validated. This conceptual shift, which occurred in the 1950s, began to reveal the complexity of the validation process as portrayed in the chapters of the successive editions. Argument-based validity was developed as a way of managing the complexity of the process of validation, and it does so by accommodating the important concepts introduced by previous generations of testers.

1951: Validity of a Test for Its Purpose

In the first edition of *Educational Measurement*, Cureton (1951) defined validity as a characteristic of a test but acknowledged that it means "how well the test serves the purpose for which it is used" (p. 621). He defined purpose as "the function to be appraised" and "the group in which the appraisal is to be made" (p. 621). For him, validity included both the relevance of a test for its purpose and its reliability, as illustrated in Figure 1.1. Reliability was defined in terms of the size of sample of performance, whereas relevance was further defined as consisting of empirical relevance and logical relevance. The latter requires a demonstration that the test content results from an appropriate definition of the criterion domain and sampling of content from that domain. The former is demonstrated by the correlation of the test with the appropriate criterion. "A direct quantitative estimate of the test's validity is provided by the actual test-criterion correlation corrected for attenuation in the criterion scores but not for attenuation in the test scores" (p. 623). The correction for attenuation allows for treatment of the criterion measure to be interpreted as a "true" score, which refers to the proportion of the score variance that is not error. A correlation based on the true score of the criterion can be estimated from the observed score, making the procedure for calculating a validity coefficient clear. With such a straightforward procedure in place for estimating validity, Cureton had no need for construct theory, which would not have fit in the operationalist perspectives of that period.

Figure 1.1 Schematic Diagram of the Components of Validity as Defined by Cureton (1951)



In view of the central role of the criterion measure in estimating an empirical validity coefficient, the chapter is almost singularly focused on how relevant behaviors (i.e., performance) can be identified from a "universe of behavior" (Cureton, 1951, p. 631) and assessed in a manner that will allow them to serve as criterion measures in validation studies. Ironically, criterion measures are beset by the same challenges as any test, and readers are left to conclude that a credible validity coefficient is purely hypothetical because acceptable criterion measures can be described only in hypothetical terms as a sample of performance from a defined series of criterion behaviors. In view of this irony, Cureton acknowledged the limited utility of validation research for test use:

> Often we are called upon to make action judgments on the basis of the best available tests in situations wherein we do not know what tests are the best available, nor the validities for any tests for the purposes at hand. Such situations are the rule, rather than the exception, in educational and vocational guidance, and most of the tests which are used in guidance are intelligence tests, aptitude tests, interest tests, personality tests, and the like, rather than educational achievement tests. (p. 664)

Cureton's (1951) rigorous definition of criterion scores explains some concepts that are still in use today, such as sampling of content, universe scores, and true scores. But it also created a value-laden dilemma by assembling an impossible set of requirements for researchers working within the constraints and demands of the real world. His advice was that "ideally, we should not use the term *criterion scores* for any measures that fail to meet the requirements of random or representative selection of acts from the criterion

series, and unbiased observation and evaluation" (p. 632). In the real world where criterion scores never satisfy theoretical ideals and tests have consequences, Cureton cautioned, "A set of non-representative or biased criterion scores may well be less relevant to the ultimate criterion than are a set of scores on a carefully worked-out test" (p. 634). In the end, then, the acknowledged reality of educational and psychological testing falls largely outside the requirements for a validity coefficient as defined by Cureton.

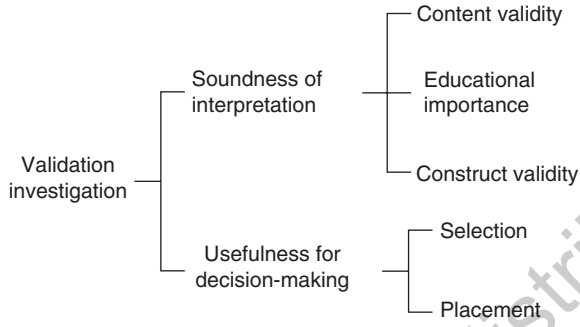
1971: Validity of Test Interpretations

In the second edition of *Educational Measurement*, Cronbach's (1971) definition shifted the object of validation to the test interpretations. Underscoring the shift, he wrote, "The phrase *validation of a test* is the source of much misunderstanding. One validates, not a test, but *an interpretation of data arising from a specified procedure*" (p. 447). Cronbach saw Cureton's validity coefficient expressing prediction of a criterion measure as being too narrow, pointing out the "paradox" that it "rests on acceptance of the criterion measure as being perfectly valid (save for random error), yet common sense tells one that it is not" (p. 487).

Cronbach (1971) presented a broader conception of validation, which "examines the soundness of all the interpretations of the test—descriptive and explanatory interpretations as well as situation-bound predictions" (p. 443). He acknowledged that the 1966 edition of the *Standards* had described three types of validity—criterion-related validity, content validity, and construct validity—but for him, there were not three validities. Instead, he referred to gathering types of evidence for "*what in the end must be a comprehensive, integrated evaluation of the test*" (p. 445). As illustrated in Figure 1.2, he defined two primary types of evidence: evidence supporting the soundness of interpretations—which can be done through inquiries into content validity, educational importance, and construct validity—and evidence about the usefulness of scores for decision making about selection and placement, for example.

Cronbach (1971) framed the process of validation as scientific hypothesis testing. The hypotheses state that test takers' performance on a particular testing procedure can be interpreted as an indicator of the construct that the test is intended to measure. Drawing upon the introduction of construct validity presented by Cronbach and Meehl (1955), he saw constructs as the logical basis for interpretations and essential when score interpretations cannot be made on the basis of a criterion or a domain of content. "Whenever one classifies situations, persons, or responses, he uses constructs. The term *concepts* might be used rather than *constructs,* but the latter term emphasizes that categories are deliberate creations chosen to organize

Figure 1.2 Schematic Diagram of the Types of Investigations of Validity as Defined by Cronbach (1971)



experience into general law-like statements" (Cronbach, 1971, p. 462). The lawlike statements form the basis for construct theories that are intended to explain test performance.

Test performance serves as the empirical data in the process of hypothesis testing and, therefore, must be gathered with great care through the testing procedure, which Cronbach (1971) referred to as the operational definition of the construct. An operational definition is a full description of the procedures, including test content, and the allowable variations that are repeatedly administered to gather consistent samples of test takers' performance. The consistency afforded by a good operational definition plays a critical role in gathering relevant performance data from test takers from which the tester makes inferences about the construct. Consistency, or reliability, needs to be achieved not only by gathering a sufficient number of samples of performance but also by gathering the appropriate samples of performance. Consistent performance samples play an important role in Cronbach's view of construct validation as scientific hypothesis testing because constructs ascribe meaning to systematic observations.

Cronbach's (1971) portrait of researchers engaged in scientific hypothesis testing implied scientific values of rigor and an unrelenting quest for developing theories useful for explaining test performance. The discovery-oriented values of a scientist are evident in his view of validation:

> A test score has an endless list of implications, and one cannot validate the entire list. Construct validation is therefore never complete. Construct validation is better seen as an ever-extending inquiry into the processes that produce a high or low test score and into the other effects of those processes. (p. 452)

Such programs of inquiry are undertaken by a community of scientists with a common set of values motivating them to discover useful ways to define and measure constructs, which "requires the concurrence of persons who have thought deeply about the problem and have given due weight to research from laboratories with other orientations" (Cronbach, 1971, p. 480).

Cronbach's (1971) presentation of a never-ending process of construct validation in which no coefficient of construct validity exists and a series of studies does not "permit a simple summary" (p. 464) was overwhelming to many textbook writers and practitioners, and remains so today. Many textbooks today still teach students that there are three types of validity, even though by 1985 the *Standards* presented construct validity as central to a single, integrated judgment, rather than as one of three validities. Nevertheless, Cronbach's prescient vision of validation as a research program serves as the foundation for argument-based validity.

1989: Validity of Interpretations and Actions

In the third edition of *Educational Measurement*, Messick (1989) elaborated on a unitary conception of validity by defining validity as "an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores" (p. 13). Messick saw the three validities—content validity, criterion-related validity, and construct validity—as a historic idea, which promoted the poor practice of choosing one of the validities rather than engaging in the type of scientific hypothesis testing depicted by Cronbach (1971). Construct validity, for Messick, was central to all validation research, and "because content- and criterion-related evidence contribute to score meaning, they have come to be recognized as aspects of construct validity. In a sense, then, this leaves only one category, namely, construct-related evidence" (Messick, 1989, p. 20).

The distinction between types of validity and types of evidence is lost on many researchers and practitioners who continue to use the pre-1985 terminology of multiple validities and even add new types of validity (e.g., see the analysis by Newton & Shaw, 2014). However, for Messick and others viewing validation as a scientific process of investigating score meaning, the distinction between "types of validities" and "evidence for validity" is important, so much so that Messick chronicled the shift in definitions of validity from types of validity in the 1950s to the 1980s view of validity as unitary. "Types of evidence" fits well with the perspective of validation as hypothesis testing, whereas "types of validity" does not.

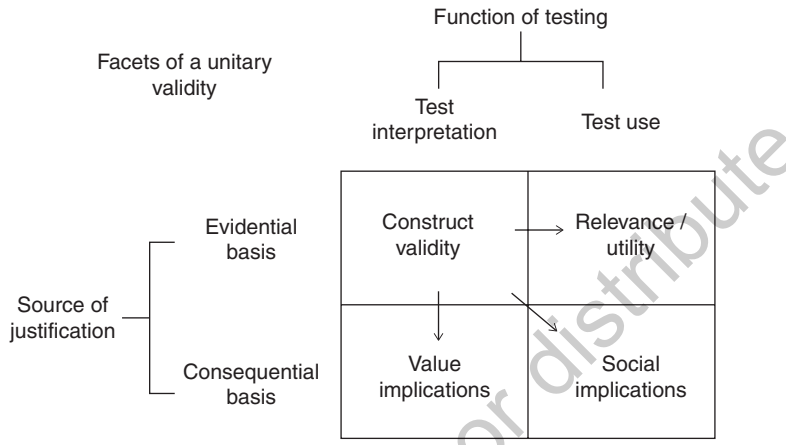
Constructs are central to Messick's (1989) presentation of validity. He defined a construct as a meaningful interpretation of performance consistency. Consistency, or reliability, is important because any test score, as a summary of performance, can have meaning only if the individual samples of performance mean something in combination.

> The key point is that in educational and psychological measurement inferences are drawn from scores, a term used here in the most general sense of any coding or summarization of observed consistencies on a test, questionnaire, observation procedure, or other assessment device. (p. 14)

Building on Cronbach's (1971) entrée into validation as scientific hypothesis testing, Messick (1989) explored aspects of philosophy of science to lay a principled basis for defining constructs, conceptualizing validation as inquiry, and developing the facets of a unitary conception of validity. Ontologically speaking, construct meaning can be conceptualized from realist, constructivist, and realist-constructivist positions. Realists view constructs as true. A constructivist position, such as the one taken by Cronbach (1971), does not presume a search for the truth. Cronbach explicitly used the term "*usefulness*, not *truth*" (p. 477) to refer to explanatory theoretical networks of constructs. Messick (1989) presented a constructivist-realist position as a middle ground. Epistemologically speaking, Messick examined five modes of inquiry from which researchers can investigate meaning. The goal was to go beyond declaring validation to be scientific inquiry to laying out potential modes of scientific inquiry (or epistemologies) for discovery of construct meanings reflecting different ontologies. In doing so, he explicitly recognized the need to take into account the contextual social and cultural aspects of test interpretation and use.

Mesick's (1989) unitary validity is made up of four facets resulting from each of the two functions of testing (interpretation and use) being justified two ways (evidence and consequences), as shown in Figure 1.3. The evidential basis for test interpretation is construct validity. The evidential basis for test use is construct validity and relevance/utility because "*general* evidence supportive of construct validity usually needs to be buttressed by *specific* evidence of the relevance of the test to the applied purpose and the utility of the test in the applied setting" (p. 20). This means that validity inquiry encompasses the investigation of the local, cultural meanings of test scores and their use.

Figure 1.3 Schematic Diagram of the Facets of Validity as Defined by Messick (adapted from Messick, 1989, p. 20)



The consequential basis of the framework, which includes appraisal of value implications and social implications, is seen by some as a controversial departure from previous conceptions of validity. But Messick (1989) saw the examination of values inherent in construct names and theories as being within the scope of validity inquiry. For Messick, "the consequential basis of test interpretation is the appraisal of the value implications of the construct label, of the theory underlying the test interpretation, and of the ideologies in which the theory is embedded" (p. 20). He described "the consequential basis of test use [as] the appraisal of both potential and actual social consequences of the applied testing" (p. 20), adding a complex sociocultural layer to the validation process. Decades later, debate continues about the breadth of Messick's definition of validity (e.g., Cizek, 2012; Lissitz, 2009). There is no dispute, however, that value implications and social consequences are matters of importance in all testing and assessment and that Messick's treatment of these topics is seminal. Nevertheless, even as the 1999 *Standards* included consequences in the chapter on validity, some disagree that such issues should be encompassed in the meaning of validity, as noted in Chapter 3.

Messick (1989) expanded the scope of validation by incorporating constructs, test use, values, and consequences all into a validity framework. He explained the framework in terms of his analysis of the evolving philosophy of science that recognized the culturally and historically situated values guiding the process of validation. In view of the breadth and depth of this

chapter, it has been recognized as a profound desideratum by some professionals and as a bewildering, erudite addition by others. Messick's presentation of validity has fueled rich reflection and debate in the field, but it has also added the complexity of social and cultural dimensions to the process of validation.

2006: Validity of Interpretation and Use

Kane's 2006 chapter in *Educational Measurement* defined validity by describing the action of validation: "to validate an interpretation or use of measurements is to evaluate the rationale, or argument, for the claims being made, and this in turn requires a clear statement of the proposed interpretations and uses and critical evaluation of these interpretations and uses" (p. 17). Whereas Messick had referred to interpretations and actions, Kane used the expression "interpretations and uses," which might be seen as narrower in scope. However, the more telling difference in Kane's presentation from that of Messick is the shift from the scientific language of empirical evidence and theoretical rationales to the language of rhetoric. Kane wrote about evaluating "the rationale, or argument for the claims" made about test interpretation and use. From a theoretical perspective, Messick had spanned the divide between the science of validation and the sociopolitical context of testing, but Kane went further by actually framing the validation process in its social context with multiple potential participants: "Ultimately, the need for validation derives from the scientific and social requirement that public claims and decisions be justified" (p. 17).

Kane's goal was to provide a pragmatic approach for doing validation as a means of putting into practice analytic frameworks such as Messick's for defining validity (Kane, 2001). Kane later reflected on the argument-based approach as a means of extending "the construct-validity model by substituting an IUA [interpretation/use argument] that specifies the inferences and assumptions inherent in the proposed interpretation and use of the test scores for the kind of scientific theory envisioned by Cronbach and Meehl (1955)" (Kane, 2016, p. 208). Kane's approach was to shift from positioning constructs as the basis for test score interpretation to requiring the tester to specify claims expressing the meaning of the score and the inferences required to make such claims. The useful insight accomplished by the shift from constructs to claims is that test interpretations and uses entail multiple different types of meanings, only some of which can be expressed by construct definitions. Moreover, it is up to the tester to formulate the relevant claims, depending on the intended meanings, and then to investigate the defensibility of the intended meanings for the intended users. This "contingent approach to validation," as Kane put it, has the advantage of

customizing the validation research program to meet the needs of the test users in their particular contexts of test use, rather than requiring all testing programs to engage in a prescribed process of validation (p. 208). In other words, a contingent approach is intended to recognize context-specific validation needs.

Kane's representation of validation as a process, as shown in Figure 1.4, consists of three types of actions. In the first, the test developer creates an interpretation/use argument that will serve "as the framework for collecting and presenting evidence" for the test score interpretation and use (Kane, 1992, p. 527). The rest of this book explains how interpretation/use arguments are constructed by assembling the intended claims about the test score interpretation and use into a logical structure, or chain, through the use of inferences. The second action is to design and carry out the research required to provide support for making the inferences that lead to each claim. The research can consist of documenting test development practices, expert analysis of content, statistical item analysis, theoretically motivated correlational analyses, and standard-setting research, for example. The research needed depends on the inferences and claims in the interpretation/use argument. The third action is to summarize the research results in a validity argument. The validity argument states the claims and the rationales for their support, insofar as support is warranted on the basis of research results.

The central contribution of the argument-based approach is that it helps testers use the concepts introduced in previous definitions of validity to conceptualize a concrete process for conducting relevant research and interpreting its results. Despite the pragmatic goals of a validity argument framework, at first glance, concepts such as "inferences" and "claims" seem at least as abstract as those in previous frameworks. Especially for testers still thinking in terms of three validities, the conceptual leap into

Figure 1.4 Schematic Representation of the Process of Validation Based on Kane (2006)

validity argument may seem daunting. The sketch of the academic background of the validity argument in this chapter provides some scaffolding because it introduces the central concepts in testing required for developing validity arguments.

Evolving Concepts in Testing

Concepts such as test performance, constructs, and values have been introduced and used in the successive chapters of *Educational Measurement* to conceptualize validity and validation, and these all remain important in argument-based validity. Seven such expressions are included in Table 1.1, each with a note about its respective role and degree of significance in each of the four presentations of validity from 1951 through 2006.

Cureton (1951) saw content as central to validity because both the relevance of a test and the selection of criterion performance depend on content. Cronbach (1971) did not dispute the importance of test content, in particular for its role in the operational definition of the construct and for descriptive interpretations of test scores, but he acknowledged the limitation of test content for explaining test scores. Even though Messick (1989) did not maintain the expression "content validity" in his unitary definition of validity, he recognized content-related evidence as important for investigating construct validity, which for him was central. Kane (2006) saw test content as relevant for inclusion in a validity argument because test tasks define the nature of the samples of performance that can provide one basis for score interpretations. Chapter 6 explains how the definition and selection of test content during the development process can be included in a validity argument to assert the role of test content in score interpretation. Chapter 4 offers a second avenue: Argument-based validity provides for inclusion of test content, with a claim that the score interpretation is relevant to the content of certain tasks in the classroom, curriculum, or real-world contexts.

Reliability has been seen as central to validity at least since the 1950s, when Cureton (1951) defined reliability as one of the two aspects of validity (see Figure 1.1) and demonstrated that the "validity coefficient" was limited by the reliability of both the test and the criterion measure. Cronbach (1971) also saw reliability as central to validity, but whereas Cureton emphasized the number of samples of performance, Cronbach focused on the consistency of the sample of performance obtained from test tasks developed from the operational definition. Messick (1989) built upon Cronbach's emphasis on the substantive interpretation of consistency by defining a construct as an interpretation of performance consistency. Accordingly, Kane (2006) created a means of making claims about reliability for inclusion in validity arguments, as described in Chapter 5.

Table 1.1 The Roles of Testing Concepts in Four Editions of *Educational Measurement* (1951–2006)

	Edition of Educational Measurement			
Aspects of testing	*1951: Validity of a test*	*1971: Validity of interpretations*	*1989: Validity of interpretations and actions*	*2006: Validity of interpretation and use*
Test content	Critical to relevance (one aspect of validity); must be demonstrably relevant to criterion performance	Important for interpretations required for validation (content validity); part of the operational definition, which is critical to validation	Key to content evidence for validity as one source of evidence to be used when investigating score meaning	Relevant for interpretations about the quality of the test tasks and about the appropriateness of scores for particular uses
Reliability	One of the two a spects of validity and required to obtain a high "validity coefficient"	Important because it is connected to the operational definition (which must be consistent across test tasks) and the construct (based on consistent categories)	Central to validation because performance consistency is summarized and expressed by test scores	One interpretation that can be made based on test scores; if it is made, it must be included in the argument
Performance	Central to both aspects of validity: Relevance requires a sample of performance to represent a defined universe of behavior; reliability requires a sample of performance of sufficient size	The critical result of a good operational definition	The consistency in performance is critical for score interpretation	A legitimate source of meaning for score interpretation; if it is used to express the score meaning, it must be specified in the argument

Construct	Not given any role	Introduced as critical to validation because it identifies the categories created to interpret test scores	The foundation for the validity framework and basis for a unitary definition	One interpretation that can be made based on test scores; if made, it must be included in the argument
Test use	Not central to validity; recognized as creating a dilemma for testers because of the limits of validation research	One of the two types of validity investigations focusing on usefulness for decision making	Central to validity definition: interpretations and actions; "test use" is an action performed with test scores	A critical part of an argument for interpretation and use
Values	Not central to validity, but limitations of criterion measures and users' pressing needs for tests create a value-laden dilemma for responsible professionals	The basis for professional responsibility of testers to pursue a rigorous program of scientific hypothesis testing and for decision making	Critical to the validity framework, which includes "values" as they connect with both construct meaning and consequences of testing	Inherent in the validation process; can be specified as assumptions underlying their respective inferences in a validity argument
Consequences	Not central to validity, but prompts recognition of the limits of validation research for contributing to desired social consequences	The source of concerns about effects on students when tests are used in educational decision making	Critical to the validity framework, which includes "consequences" of testing	The cause for public engagement, which motivates the need for validity arguments

Performance is treated as central to both of Cureton's (1951) aspects of validity: Reliability requires a sample of performance of sufficient size, and relevance requires the sample of performance to represent a defined universe of behavior. The latter requirement makes performance integral to defining score meaning in the absence of constructs. Cronbach (1971) did not require performance to serve as a basis for score interpretation, but, for him, eliciting relevant performance was nevertheless integral to validation. Messick (1989) amplified the relative roles Cronbach placed on performance and constructs by relying on constructs as the basis for the meaning of test scores and defining the construct as a meaningful interpretation of performance consistency, rather than simply an interpretation of performance. If Cronbach and Messick appeared to de-emphasize the role of performance in score interpretation, Kane (2006) opened the door for re-emphasizing it. Performance is viewed by Kane as a legitimate source of meaning for score interpretation in a validity argument. He therefore places on the table for consideration by testers both the performance-oriented view of score interpretation presented by Cureton and the construct-oriented view of Cronbach and Messick. Chapter 4 explains how one or both of these approaches can be expressed in argument-based validity.

Constructs have been the aspect of validation whose role has had the most dramatic metamorphosis. Cureton's account of validity in 1951 did not even mention constructs. Cronbach (1971) included constructs as critical to validation because they identify the categories created to interpret test scores, and Messick (1989) considered constructs central to validity inquiry. Kane (2006) presented constructs as one way of expressing score interpretation but not the only way the substantive meaning of scores can be expressed. Chapter 4 explains how testers can formulate a validity argument with or without a theoretical construct.

Test use in terms of placement, selection, certification, or grading was not central to Cureton's (1951) definition of validity. He defined relevance, one aspect of validity, with reference to the behaviors or criterion measures in the domain of interest, rather than as relevance for decision making in educational, clinical, and work settings. For him, validity had to be for a particular purpose, but ironically, he defined purpose as what is tested and who is tested, but not what for. In fact, Cureton saw test use as creating a challenge for testers because of the limits of validation research. Cronbach (1971), in contrast, included test use for decision making in his validity framework (see Figure 1.2.). He depicted studies of specific decision-making uses of tests as one way of investigating validity, but he stopped short of defining validity as pertaining to both interpretations and uses. Messick (1989) took the additional step of including test use by defining validity as a judgment about the interpretations and actions based on test

scores. Test use is an action performed with test scores in a particular social and cultural context, and therefore evidence of the relevance and utility are required to make a validity judgment. Kane (2006) defined validity as an appraisal of the rationale for test interpretation and use. In doing so, he strengthened the imperative for a validity argument to take into account the sociocultural context of test use. Accordingly, Chapter 3 demonstrates how to include test use in a validity argument.

Values underlying validation research are evident throughout the four chapters, although their prominence and roles shift. Values are expressed in Cureton's (1951) and Cronbach's (1971) depictions of a community of testers guided by their earnest professionalism as they attempt to provide well-justified advice to test users. Cureton acknowledged that the definition of validity requiring a perfect criterion measure created a moral dilemma in view of both the need to advise prospective test users about validity and the impossibility of doing so because of insufficient criterion measures. Despite Cronbach's framing of validation in scientific terms, he recognized that values and judgments were integral to a never-ending validation process propelled by the ethical pursuit of defensible test interpretations. He also saw values as the basis of decision making, regardless of the statistical data analysis serving in the process.

Messick (1989) included values explicitly in his validity framework (see Figure 1.3.) with an extensive discussion of value implications. In particular, Messick pointed out the implicit cultural and political values concealed in constructs such as "intelligence" and "aptitude" in the first half of the 1900s. Gould's (1996) analysis shows how the racist values of this time period formed the basis for intelligence research and how the tests were used to perpetuate these values. Zwick (2006) summarizes the foundation of modern college admissions testing in the northeastern United States at a time when 2% of the population attended college and three-quarters of them were white men. Values instituted in these tests were arguably exclusionary, even though today similar practices are intended to implement the values of an inclusive, merit-based system. The historical shift may partially explain today's irony that, as Zwick points out, "to some, tests like the SAT are harsh and capricious gatekeepers that bar the road to advancement; to others, they are the gateways to opportunity" (p. 649).

Kane (2006) further shifted the role and meaning of values for validation by explaining how they extend beyond the community of testers working on a particular testing issue to responsibility for communicating the logic of the validity argument. In a sense, the raison d'être of argument-based validity is to have sufficient technical language to develop the logic behind test use so that a rationale can be judged by others for its soundness and evaluated for its relevance to other test takers and other contexts. In short,

today values are recognized to permeate all aspects of the validation process—as illustrated in Cizek's (2012) revision of Messick's (1989) characterization of the validation process—and are therefore relevant to every chapter.

The social consequences of testing were recognized by Cureton (1951), who wrote about the limits of validation research for serving society, with recommendations about the validity of tests for such purposes as career guidance, qualifications certification, and educational advancement. Cronbach's (1971) discussion of decision making as one of the two types of investigations for validity inquiry is concerned with school-based consequences, which included concerns about effects on students. Cronbach's highly social and political treatment of validity appeared much later (Cronbach, 1988), when he introduced the need for validity arguments that speak to a variety of audiences. Messick (1989) placed social consequences within his validity framework, and influenced by Messick, the *Standards* included "evidence based on consequences of tests" (p. 30) as one type of validity evidence. Building on Cronbach's 1988 positioning of validity argument in the social and political context, validity argument provides a mechanism for inclusion of consequences, as explained in Chapter 3. The relationship between testers and society from the early 1950s into the 2000s reflects a change in perspective from educational and psychological testing as a neutral science to its conception today as a culturally situated social responsibility. As a result, today the fairness of test interpretation and use for individuals and subgroups within the population are areas of continuing analysis and research, which take into account the consequences of testing (Camilli, 2006).

Conclusion

The basic testing concepts introduced over the past decades remain useful for understanding and developing validity arguments today. Moreover, the chronology of definitions of validity and validation in the field should enable readers to see both the complexity of validation issues and the variation in how they can be conceived. Like the history outlined in this chapter, the academic discussion of validity argument continues (e.g., see the papers by Brennan, Haertel, Moss, and Sireci in the *Journal of Educational Measurement*, 2013). Despite the value of the continuing academic discussion about the epistemological frameworks (e.g., Lissitz, 2009; Moss, 1994) and methods underlying professional conceptions of validity, in the real world of testing practice, professionals need to be able to justifying test

score interpretations and uses. As Shepard (1993) pointed out, understanding validation is cultivated, in part, through the study of examples of actual validation practices. Examples of argument-based validity in practice, however, are rare, meaning that the academic discussion of validity argument is undertaken largely in the abstract. This book is intended to expand the circle of professionals able to use argument-based validity for designing and conducting research on tests used for a range of purposes. Most professionals and students of testing are familiar with the basic concepts, if not their history, as introduced in this chapter. The following chapters will build on them to explain how testers can construct their own validity arguments.