2

THE EVOLUTION OF VALIDITY

CHAPTER OUTLINE

2.1 FIRST AND CURRENT DEFINITIONS OF VALIDITY

The *Merriam-Webster Dictionary* defines validity as "a) the state of being acceptable according to law," or "b) the quality of being well-grounded, sound, or correct" (https://www.merriam-webster.com/dictionary/validity). The second definition is more applicable in the context of the validity of test scores. An early definition of validity is the "degree to which a test measures what it purports to measure," a definition initially proposed by Garrett (1937, p. 324). In one of the first peer-reviewed articles on validity, Rulon (1946) criticized this definition in connection with educational achievement tests as not being useful "because under it the validity of a test may be altered completely by arbitrarily changing its purport" (p. 290). Rulon's argument against the notion of validity as the degree to which a test measures what it's purported to measure is that this definition does not account for the variety of potential applications. For example, a test measuring high school algebra might be used as a graduation requirement or as a formative assessment of a student's strengths and weaknesses. In both cases, it is important to determine that the algebra skill is indeed what is being measured. On the other hand, these two instruments are likely to be designed differently to meet their purposes. As a result, the evidence to support validity in these two applications of the same content would very likely be different. It is interesting to note that although standardized tests have been in existence for over 4,000 years, notions of validity have arisen only in the last hundred years. And even then, Garrett's writing on the topic was published forty years after the first modern psychological test, Binet's intelligence scale, was implemented. Until then, the user of any test was responsible for demonstrating that the test was actually useful for its intended purpose.

By contrast, the current version of the *Standards* (AERA, APA, & NCME, 2014) offers this definition of validity:

> *Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. (p. 11)*

In other words, validity is about the interpretation of test scores and the evidence that supports that interpretation, not the test itself. Clearly, views on validity have changed dramatically over the past 77 years. In the next sections, I describe the evolution of thinking about validity. Knowing about this evolution is important because validity and validation will continue to evolve as the methods for test development, delivery, and scaling likewise evolve.

2.2 CRITERION MODEL OF VALIDITY

In the first half of the 20th century, psychometricians tended to view validity from a practical standpoint. Garrett (1937) wrote that "A test is valid for a particular purpose or in a particular situation—it is not *generally* valid" (p. 324). For example, Garrett reports on the use of the Army Alpha exam to select applicants for clerical positions. The test turned out to be a poor predictor of workplace performance. As a result, a test could be valid for one purpose but invalid for another. Most of the first standardized psychological tests were selection tests, such as the Binet intelligence scale and the Army Alpha and Beta Tests. Their validity depended on test scores doing what was intended, selecting at-risk students or selecting recruits for officer training. Pearson's (1896) correlation coefficient offered a statistical approach to validation and was used widely. For example, in one of the first measurement textbooks, Guilford (1946) wrote that "in a very general sense, a test is valid for anything with which it correlates" (p. 429).

This pragmatic view of validity came to be known as *predictive* validity. Validity was expressed by Cureton (1951), in the validity chapter of the first edition of *Educational Measurement*, as "how well a test does the job it is employed to" (p. 621). If this job is selection, a criterion is usually available, such as job performance for a personnel selection test or college GPA for an admissions test. Cureton went to write that a way to validate a test score is:

> *to give the test to a representative sample of the group with whom it is to be used, observe and score performances of the actual task by the members of this sample, and see how well the test performances agree with the task performances. (p. 623)*

At the same time, the concept of *concurrent* validity was introduced as a separate type of validity (APA, AERA, & NCME, 1954). Here, the test and criterion scores are obtained at the same point in time. This type of validity usually involved correlating test scores with another widely accepted measure of the same construct, although, as Kelley (1927) pointed out, just because two tests measure the same construct by name, they do not necessarily measure the same construct. Still, concurrent validity was considered to be an important source of validity evidence. Eventually, these two types of validity were combined into a single type—criterion-related validity (APA, AERA, & NCME, 1966).

Kane (2006) points out two major advantages of the criterion model of validity. First, the criterion is directly related to the test score and therefore clearly relevant to test score interpretation and use. Second, a quantifiable indicator of validity appears, at least on the surface, to be objective. On the other hand, Cureton (1951) and Anastasi (1986) noted several difficulties measuring the criterion. For one thing, the criterion may not be measured on a quantitative scale. For example, Binet's scale was intended to identify at-risk children. One possible criterion would be teacher judgments on the appropriateness of an educational intervention made as a result of the Binet test score (which originally was a comparison of mental age to chronological age). Such a criterion would be subject to measurement error, such as bias or imprecision on the part of the teachers. Additionally, even for quantitative criteria, some degree of measurement error is likely.

2.3 CONTENT-BASED VALIDITY MODEL

At the same time that the criterion model gained wide acceptance, during the 1940s and 1950s, other psychometricians noted the weaknesses of the criterion model and argued for a validity model based on test content (Rulon, 1946). An initial content-based validity concept was "face validity," defined as the degree to which a test's content appeared to be measuring the intended construct. For example, the item "I am sad most of the time" would appear to be measuring depression, but an arithmetic problem would not. Face validity is still a term occasionally found in use today, but it is one not taken seriously by psychometricians. Angoff (1988) wrote that "the effort to make a test face valid was, and probably is today, regarded as a concession, albeit an important one, to gain acceptability rather than a serious psychometric effort" (p. 24).

Instead, a view arose that, for many tests, particularly those for measuring educational achievement or for credentialing purposes, the criterion model was inadequate. For one thing, there was likely to be no infallible criterion against which to compare test scores. Furthermore, the focus of test score interpretation rested on the measurement of the specific knowledge and skills represented by the test items. This requirement led to the idea of claiming that the test content and associated item format is a representative sample of the universe of all possible items so that the test score is an unbiased estimate of overall performance. If the sample is large enough, that is if the test is long enough, then sampling error can be minimized.

For a time, content validity was challenged by psychometricians who found statistical evidence more convincing. For example, Loevinger (1957) argued that, on most tests, item formats such as multiple-choice were written to measure only some of the processes deemed important, and they were selected to represent particular levels of difficulty and discriminating power. In other words, a representative sample of content is usually impossible to achieve. However, when the relevant content domain has been carefully specified, items have been written that are representative of that domain, and they have been scored appropriately, then content evidence came to be seen as important, particularly for tests that measure mastery of a specific set of knowledge and skills, as in credentialing exams.

2.4 CONSTRUCT VALIDITY MODEL

The criterion and content models worked well for selection and achievement tests and tests measuring cognitive constructs, but they were not as applicable for tests of noncognitive constructs, such as personality tests, whose purpose was to provide psychological interpretations often used by counselors and clinical psychologists. For these tests, no specific criterion is available except for other tests claiming to measure the same construct, and no definitive content is often found. Operational definitions of constructs can vary among tests of the same construct (see Chapter 6), leading to vastly different content. To address this shortcoming, psychometricians in the 1940s explored the idea of construct validity for the validation of tests measuring theoretical constructs. Construct validity appeared for the first time in the *Technical Recommendations* (1954) as a new type of validity to go along with predictive, concurrent, and content validities.

Cronbach and Meehl (1955) published a seminal paper on construct validity that transformed validity into a much different concept, one that has led to modern models of validation. For psychological constructs, Cronbach and Meehl argued that while criterion and content evidence were insufficient for many constructs, those constructs often came with a theory of what they were and a set of hypothesized relationships with other constructs. For example, while a test of emotional intelligence has no definitive criterion or content, one can predict certain relationships to hold true. For example, one theory of emotional intelligence (Goleman, 1995) holds that people high in emotional intelligence have more stable marriages and are more productive at work. Furthermore, emotional intelligence is theorized to be a set of skills on which people can be trained. If these relationships are observed, then support for both the test and theory is found. If not, then either the test or the theory, or both, are suspect.

One implication of their construct validity model is that validation is a process that unfolds over time, not from a single study or data collection effort. Cronbach and Meehl (1955) used the concept of nomothetic span to indicate the network of relationships proposed by the theory of the construct. These relationships can be investigated by the traditional methods of science. For a test measuring depression, these could include experimentation (therapy as a treatment group), comparison of groups (high versus low depression subgroups) on various outcomes, and correlational/regression studies (depression predicting outcomes such a job productivity and stability of long-term relationships). As a result, theories are never proven, but enough evidence accumulates over time that the theory is widely accepted. The same goes for tests. Validation is never completed. Any collection of data from the test can be considered validity evidence.

Another implication of Cronbach and Meehl's (1955) model is that the focus of validation is not on the test but on the interpretation of test scores. Any evidence, including content and criterion-related evidence, bears on construct validity. In this regard, Campbell and Fiske (1959) distinguished between convergent and discriminant validity. Convergent validity consists of correlations between tests measuring the same construct, while discriminant validity consists of correlations between tests measuring different constructs. Multitrait-multimethod matrices (MTMMs) became a popular method for investigating construct validity. Table 2.1 shows an MTMM for two constructs, grit and self-concept, and two methods, objectively scored test scores and ratings by the respondents' colleagues. The entries in the matrix are correlations (I should note here that these correlations are fictitious, not from real data). The numbers in parentheses in the diagonal of the matrix are reliability

TABLE 2.1 ⬢ Example of a Multitrait-Multimethod Matrix

		Grit		Self-concept	
		Test	Self-rating	Test	Self-rating
Grit	Test	(.91)			
	Ratings	0.68	(0.85)		
Self-concept	Test	0.66	0.16	(0.79)	
	Ratings	0.35	0.31	0.75	(0.76)

coefficients that indicate the degree of precision of scores (see Chapter 4). The correlations between the same construct by different methods indicate convergent validity. These show relatively strong positive correlations. The correlations between different constructs by the same or different methods show discriminant validity. These are expected to be noticeably weaker than the convergent validity coefficients. Note that this is not the case for the correlation between grit and self-concept test scores. This suggests a correlation due to a common method, using an objectively scored test. Such a correlation undermines the validity of scores on both tests because it suggests that test format can bias scores the same way on both the test and the ratings.

2.5 THE HOLY TRINITY

After the publication of the 1966 *Standards*, validity came to be viewed as a "toolkit." There were three types of validity, each to be used differently for tests with different purposes. For example, typical of many technical reports at the time, the first edition of the *GED Technical Manual* (Auchter, Sireci, & Skaggs, 1993) contained a chapter on validity organized as follows:

> *Content validity:* Showing that the GED Tests measure the typical American high school program of study in each subject area, it's the most important piece of validity evidence. Here, the program shows the steps in developing test specifications and blueprints in consultation with instructional leaders and how the tests reflect high school coursework and workplace skills.

Criterion-related validity: This section shows correlations between GED scores and high school GPA, ACT scores, and other achievement tests. This section also included comparisons of high school seniors and GED candidates on GED scores and analyses showing how GED scores map onto high school letter grades.

Construct validity: This section included any other evidence that did not fall clearly into the previous two sections. This included studies of the relationship of GED passing scores to taking specific high school courses, correlations among the five GED tests, and a comparison of high school seniors and GED candidates on higher education and employment outcomes.

The GED Testing Service planned their validation around these three types of validity; that is, planning to have some evidence to report about each type, even though clearly for this testing program, content and criterion-related validities were the most important.

Not all psychometricians embraced the toolkit approach to validation. Guion (1980) derisively referred to the three types as “something of a holy trinity representing three different roads to psychometric salvation” (p. 386). Cronbach (1971) complained that for some testing programs, construct validity evidence amounted to “haphazard accumulations of data rather than genuine efforts at scientific reasoning” (p. 483).

Unlike criterion-related or content validity, which did not offer a specific set of procedures, construct validity embraced a general scientific approach in which a variety of research methods could be used to provide evidence relevant to test score interpretation. As a result, construct validity came to acquire increasing importance compared to the other two types. This trend led to two ideas. First, a test measures a construct, and any evidence related to that measurement is part of construct validity. Second, there really is only one type of validity, construct validity, and that criterion-related and content validity were types of evidence under the construct validity umbrella. Loevinger (1957) was one of the first psychometricians to argue for considering different types of evidence instead of different types of validity. She divided construct validity into three types: substantive (content validity focused on a theoretical perspective of the construct), structural (internal structure of the test), and external (relationships between test scores and other variables). She was also the first to articulate the view that "construct validity is the whole of validity from a scientific point of view" (p. 636). By the 1980s, this model of construct validity became widely accepted by psychometricians even as many testing programs still used the toolkit model of validity. Then, in 1989, Sam Messick published his landmark chapter on validity in the third edition of *Educational Measurement* (Messick, 1989), in which he introduced an expansion of the unitary model of construct validity, a model that still dominates current thinking.

2.6 MESSICK'S VALIDITY FRAMEWORK

Messick (1989) offered this somewhat dense definition of validity:

> 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, italics in original)

Messick means, first of all, that validity is about the meaning or interpretation of test scores and is not a property of a test. In other words, we would never say that a test is valid. Instead, we try to make the case that a particular interpretation and use of a test score is valid. Second, validity is a matter of degree. Like most theories in education and the social sciences, test score validity is never proven. Evidence is collected over time that supports or undermines a test score interpretation. How evidence is collected is much like how research in general is conducted, that is, using the methods of science, such as experimentation, group comparisons, and correlational/regression studies.

Additionally, because validity has to do with test score interpretations, these interpretations need to be specified before validity can be addressed. Messick stressed the need to investigate alternative interpretations of test scores. For example, does a decrease in reading test scores indicate lower achievement? Or, does that decrease result from contextual factors, such as changes in the order of items or how they are presented (see mention of the NAEP Reading anomaly in Chapter 4)? Probably the most controversial aspect of Messick's unified view of validity is the emphasis given to consequential evidence. In other words, the application of a test may result in value labels, such as retarded, depressed, or suicidal, that have consequences for respondents. There may also be social consequences for groups of respondents. Messick argues that value labels and social consequences can affect score interpretations and the way respondents answer test questions and, as a result, are an important part of the validity framework.

Messick sought a unified framework for construct validity that didn't overly rely on specific types of evidence for a particular use. He considered two interconnected *facets* of validity, as shown in Table 2.2. One facet is the outcome of testing, an interpretation of a test score and/or its use. The other facet is the justification for test interpretation or use, based on evidence or consequences. Test interpretation based on evidence is the construct validity conceptualized by Cronbach and Meehl (1955). However, justification for a particular use of a test may require additional evidence to support that use. For example, an interpretative score report from a test measuring emotional intelligence may provide valuable insight to respondents. But if that same test is used to make a decision about a respondent, such as entry to emotional intelligence training, then that use of the test requires additional supporting evidence.

In this framework, there are two basic types of threats to construct validity. The first is *construct underrepresentation*, in which the test does not include important parts of the construct. An example is a high school science test that does not contain any items about biology, and thus the construct "high school science knowledge" is underrepresented. The other main threat is *construct-irrelevant variance*, where a secondary construct contaminates score interpretation. An example is test preparation strategies for multiple-choice items that lead to higher scores.

TABLE 2.2 ⬢ Messick's Facets of Validity

		Outcome of Testing	
		Test Interpretation	**Test Use**
Justification for Testing	**Evidential Basis**	Construct validity	Construct validity + relevance/ utility
	Consequential Basis	Value implications	Social consequences

To counter these threats, Messick identifies six types of evidence, later called aspects of construct validity (Messick, 1995). These are not types of validity to be chosen as needed for a particular test purpose, but a complete set of evidence types, all of which together answer the two types of validity threats.

2.6.1 Content Evidence

Content validity evidence is similar to but subtly different than the original content validity. Psychometricians have long debated the importance and relevance of content validity for many years. They argued whether judgments of test content were an actual property of a test while other types of validity were the properties of responses to items. In Messick's framework, content evidence is considered to be necessary but not sufficient for a unified evaluation of score validity. That is, content is viewed in conjunction with other types of evidence. Consider, for example, the Iowa Tests of Basic Skills (ITBS) Spelling subtest. The multiple-choice items on this test ask students to identify a misspelled word among four different words. Subject matter experts might not consider this task to be measuring spelling skill, preferring instead to ask students which choice is the correct spelling of a specific word. However, as former ITBS senior author H. D. Hoover once pointed out (1987, personal communication), the item format appearing on its Spelling subtest shows much stronger validity evidence of other types, including stronger correlations with other measures of spelling skill and stronger internal consistency, than the pick-the-correct-spelling format.

Test content is defined by the 2014 *Standards* as "the themes, wording and format of the items, tasks, or questions on a test" (p. 14). Sources of test content vary considerably according to the intended purpose of the test. For a credentialing exam or job selection test, content may be developed by professional judgments and observations of key behaviors. For an educational achievement test, content may be determined by state curricular standards. For tests measuring personality and other noncognitive constructs, a theory of the construct can guide content. A critical component of content evidence in all tests is evaluating the degree of *alignment* between test content and the sources used to develop the content. Alignment is threatened if the content of the test is not representative of the entire construct, in other words, there is construct underrepresentation (e.g., if a certification test of accounting does not include tax law), or if the content is not directly relevant to the construct (e.g., a test measuring motivation includes items related to self-concept).

Content evidence also includes decisions about test design including item format, time limits and test length, mode of administration, and so forth. The main question here is whether these decisions insert construct-irrelevant variance into scores. For example, many tests use multiple-choice items exclusively. There are logical reasons for this, as enumerated in Chapter 7, but since respondents have a chance of guessing the correct answer, scores may be inflated. Furthermore, some components of the construct, such as some higher-level thinking skills, may be difficult to measure with multiple-choice items.

All of this evidence is often offered in the test specifications that appear in technical manuals or reports. For example, in the current *GED Technical Manual* (GEDTS, 2018), there is an extensive discussion of the rationale for moving the GED Tests in the direction of making the high school credential align with "college and career readiness," the primary focus of the Common Core State Standards that have been widely adopted by state assessment programs. What follows in the manual is an extensive discussion of how this overarching goal drives the content specifications of each GED Test. The manual goes on to describe item formats, time limits, administration technology, and other issues.

Finally, content evidence also refers to item technical quality. As discussed in Chapter 8, items undergo some form of pilot testing. How this is carried out varies widely across testing programs. Messick argues that ambiguous or flawed items elicit construct irrelevant variance that undermines test score validity.

2.6.2 Substantive Evidence

Substantive evidence is "a confrontation between content coverage and response consistency" (Messick, 1989, p. 43). As described above, content is guided by subject matter expert judgment, analysis of behavioral indicators, and/or by a theory of the construct. The question here is the degree to which item responses reflect those content considerations. For example, if an item intended to be high in difficulty in fact turns out to be quite easy, construct validity is undermined. Or, a "Strongly Agree" response is intended to indicate a higher level of the construct than a "Strongly Disagree" response, but item response data suggest the opposite. Is the construct being measured in the way desired by the test developers? Messick's later writing (Messick, 1995) and the 2014 *Standards* described substantive evidence as the degree to which respondents' thought processes while answering questions are consistent with the intent of the test developers.

Compared to the other types of evidence, substantive evidence is one of the least collected types because it can be difficult to access. At this time, substantive evidence can come from three potential sources. First, "think-aloud" and cognitive labs protocols can reveal respondents' thought processes. Think-alouds and cognitive labs are one-on-one interviews with respondents who reveal their thought processes while answering test questions. For example, for a mathematics item thought to be difficult, a respondent might reveal that after two or three of the distractors were obviously incorrect, the answer choice became a guess between two options, thereby making the item seem easier than it really was. Cognitive labs became useful during the aborted efforts to create the Voluntary National Test (VNT) during the late 1990s. When Congress delayed funding of the VNT, the test's developers used cognitive labs to analyze item quality on a very small scale. This method uncovered flaws in many items, thereby avoiding more extensive data collection. Further discussion of these two methods are provided in Chapter 7.

A second potential source of substantive evidence comes from recent technological advances in test administration. Many large-scale tests are administered on the computer. As a result, data such as eye tracking, response time, and log files have the potential to be used to reveal respondents' thought processes. Research on "big data analysis" is a current hot topic, and definitive results are not yet available. A third source comes from recent psychometric developments of tests designed to measure cognitive processes. These are called collectively cognitive diagnostic models and have led to the development of tests targeted at measuring the thought processes that lead to a final answer. These are mentioned briefly in Chapter 5 in the section on Evidence-Centered Design.

2.6.3 Structural Evidence

The conceptual basis for the construct includes an explicit or implied internal structure. This structure can take many forms and is closely aligned with the test scores. Many, if not most, tests report a single score. This implies that the construct can be viewed primarily as a single continuum ranging from low to high; that is, as a *unidimensional* construct. There may be components of the content, but these are viewed as being strongly enough correlated that they are considered to be a single dimension. Examples include many tests measuring psychological constructs, such as grit (Duckworth, Peterson, Matthews, & Kelly, 2007) which has two components (passion and perseverance) but a single score. Alternatively, *multidimensional* constructs contain content components that are different enough to justify multiple scores, such as the Myers–Briggs Type Indicator (MBTI) and NAEP Mathematics. There are points in between these two, such as tests that are primarily unidimensional but also report subscale scores. TIMSS is an example of this approach.

Structural evidence seeks to uncover support for the underlying structure. This evidence tends to be highly data driven, most commonly through some form of *factor analysis*. Discussed in greater detail in Chapter 11, factor analysis can take two broad forms. Exploratory factor analysis uses the correlation matrix between items or parts of the test to form an internal structure based on the data. Hopefully, this structure is consistent with the intended structure. Confirmatory factor analysis determines how well an item response dataset conforms to the intended internal structure.

An additional piece of structural evidence is an analysis of *differential item functioning* (DIF). DIF analyses of individual items or groups of items are intended to uncover potential item bias for or against target population subgroups. DIF methods are discussed in Chapter 10. DIF is included as structural evidence because its presence signals that the proposed internal structure may be different for different population subgroups.

2.6.4 External Evidence

External evidence is about the relationships between test scores and scores on other variables. Messick (1989) distinguished between two types of external evidence: *trait validity* and *nomological validity*. Trait validity emphasizes convergent and discriminant validity coefficients, which were mentioned above in relation to Cronbach and Meehl's (1955) presentation of construct validity. As shown above, MTMMs are a common way of demonstrating convergent and discriminant relationships.

Nomological validity focuses on a network of theoretical relationships hypothesized to exist between the construct and other variables. As discussed by Cronbach and Meehl (1955), a theory of the construct often includes hypothesized relationships to other variables. For example, in developing her Grit Scale, Duckworth et al. (2007) hypothesized that grit is strongly related to educational achievement but weakly related to IQ. That these relationships were supported by research strengthened score validity and supported the underlying theory. If these relationships had not been supported, then either score validity, construct theory, or both would have been undermined.

Methods for obtaining external evidence can vary as widely as the methods for conducting research. These include predictive or concurrent test/criterion relationships, as used in the old criterion model of validity. These may also include experimentation and group comparisons. For example, Goleman's (1995) theory of emotional intelligence posits that the construct consists of skills that can be taught. To test that hypothesis, an experiment could be designed with emotional intelligence training as a treatment group. Similarly, individuals who are high in emotional intelligence, as measured by the test, are hypothesized to have more stable relationships and to be more productive at work than those with low emotional intelligence. These relationships can be investigated through group comparisons or regression analysis.

2.6.5 Consequential Evidence

The preceding types of evidence deal with the first row of Table 2.2 and are an integration of earlier types of validity (criterion-related, content, and construct). The second row of the table concerns test consequences as a source of validity evidence. The first column then addresses the value implications of test score interpretations. If respondents receive a score report, they will imbue the interpretation of their score(s) with value. For example, a score report that informs respondents that they lack grit, are low in emotional intelligence, are depressed, or that they have a positive self-concept, are high achieving, or have musical or artistic ability conveys values to respondents. As it pertains to validity, the consequential concern is whether the value labels affect respondents' scores. If a test score leads to a decision about a respondent or a label placed on the respondent, the respondent may provide misleading answers to items. Suppose a company offers motivational training to individuals who score low on a test measuring that construct. Low levels of motivation are often attached to undesirable value labels, such as "lazy," "unambitious," and "withdrawn." As a result, a respondent could be motivated to answer in such a way as to avoid those labels. In other words, an invalid score results. Even if a score report is not provided, as in the case of a research study, participants may still anticipate what the researchers are looking for and respond accordingly.

The second column of Table 2.2 points to test use. Here, Messick was concerned with social consequences. Tests have intended uses, and it is important to evaluate whether those outcomes have occurred. Does a professional licensure test actually promote qualified individuals? Does a personnel selection test pick the best candidates? If not, there is a validity problem in that score interpretation (the individual is qualified, competent), and test use is compromised. Additionally, there may be unintended consequences. There has been widespread criticism that state educational assessment programs, while promoting rigorous achievement, have the unintended consequence of narrowing the curriculum ("If it's not on the test, it's not taught."). One particular unintended outcome is an adverse impact on population subgroups. Accusations of cultural test bias have been leveled against standardized tests for decades. And certainly, an investigation by test developers of potential bias is warranted. However, the adverse impact can occur more subtly. Test preparation services are quite popular for tests such as the SAT, GRE, GED tests, and credentialing exams. These can be quite expensive, thereby favoring candidates with higher income, a variable closely related to gender and race/ethnicity. This impact becomes a validity issue: if individuals with and without test preparation achieve the same score, do their scores mean the same thing?

Consequential evidence has been controversial since Messick introduced it, mainly because it is not clear who should investigate unintended consequences. The 2014 *Standards* offer the following advice:

> *Standard 1.25: When unintended consequences result from test use, an attempt should be made to investigate whether the consequences arise from test's sensitivity to characteristics other than those it is intended to assess or from the test's failure to fully represent the intended construct. (p. 30)*

2.6.6 Generalizability

All tests are samples of items and tasks chosen to be representative of the universe of all possible items and tasks. Data for scaling, scoring, and providing validity evidence come from samples of respondents selected to be representative of the test's target population. It is reasonable to ask then how well scores generalize beyond the specific set of chosen items and samples of respondents.

For items, test reliability is a major piece of evidence, including test-retest coefficients (administration at different times), alternate forms coefficients (different sets of items), and internal consistency coefficients (relationships between items or parts within a test) (see Chapter 4). There is a particular concern for item formats that require human judgment for scoring. Different raters scoring different tasks at different times present different sources of measurement error. As a result, generalizability evidence includes interrater training methods and interrater reliability.

Generalizability to other populations (population generalizability) and settings (ecological generalizability) also come under this type of evidence. Consider the Graduate Record Exam (GRE) whose purpose is to predict academic performance of students in graduate studies. Originally developed for American undergraduates, the GRE is now available internationally in many countries. The generalizability question for the GRE is the degree to which score meaning is consistent across different countries and across different administration conditions.

2.6.7 Integrating Validity Evidence

Messick intended for the six types of validity evidence to be integrated into a rationale for test score validity. He argued that the six types applied to all mental measurements. "Taken together, they provide a way of addressing the multiple and interrelated validity questions that need to be answered to justify score interpretation and use" (Messick, 1995, p. 746). Additionally, alternative interpretations of test scores should be investigated through the possibility of construct underrepresentation and construct irrelevant variance. Together, evidence of all six types and ruling out alternative explanations can provide comprehensive support for test score interpretation.

2.6.8 *Standards* and Messick's Framework

The influence of Messick's unified framework for validity clearly guided the 1999 and 2014 *Standards*. The definition of validity shown at the beginning of this chapter, from the 2014 edition, is an adaptation of Messick's definition from his 1989 chapter and other writings. Furthermore, both the 1999 and 2014 editions discuss types of validity evidence that are similar to Messick's. Table 2.3 compares Messick's six types with the *Standards'* five types of evidence. The primary difference, besides labeling, is that the *Standards* combine external and generalizability evidence into a single type

TABLE 2.3 ⬢ Comparison of Types of Validity Evidence: Messick Versus *Standards*

Messick (1989, 1995)	*Standards* (1999, 2014)
Content relevance	Test content
Substantive theories and process modeling	Response processes
Structural fidelity	Internal structure
External and generalizability	Relations to other variables
Consequences of testing	Consequences of testing

called "Relations to other variables." Within this type, generalizability is called validity generalization.

2.7 KANE'S ARGUMENT-BASED VALIDITY FRAMEWORK

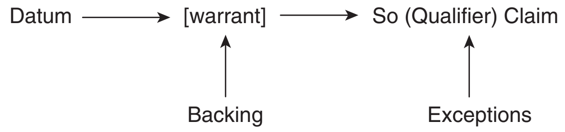
Since Messick's unified framework was introduced and supported in the *Standards*, test developers have expressed some dissatisfaction in applying the framework. This dissatisfaction comes from two primary concerns. First, although integrating validity evidence into a coherent argument is recommended, Messick offered no specific guidance on how to do this. As a result, validation has tended to consist of sorting evidence into each of the distinct types without any prioritization of the evidence. Second, the framework focuses on the construct being measured. Much of the content, structural, and external evidence relies on a coherent theory of the construct. In many applications, such a theory either does not exist or the construct is so complex (e.g., "college and career readiness") that it is difficult to define what evidence is needed.

In recent years, Michael Kane (2001, 2006) has addressed these shortcomings through an argument-based validity framework. Kane defines validity as "the extent to which evidence supports or refutes the proposed interpretations and uses" of test scores and validation as "the process of evaluating the plausibility of proposed interpretations and uses" of those scores (Kane, 2006, p. 17). In this model, much of the same evidence is collected as in the Messick framework, but instead of compiling lists of evidence, validation is organized around the proposed interpretations and uses. In Kane's framework, there are two types of arguments. The *interpretative* argument specifies a network of inferences that lead from responses to test items to the proposed interpretation and use of test scores. The *validity* argument is the evaluation of the interpretative argument, including evidence to support or undermine each of the inferences in the interpretative argument.

In this framework, a test is developed alongside an interpretative argument. First, the interpretative argument is outlined along with test specifications. The test is then developed to be consistent with the interpretative argument. As much as possible, the interpretative argument is evaluated to reveal any weaknesses. Following this procedure necessitates that test score interpretations and intended uses are clearly defined ahead of time, as well as the line of reasoning from test specifications to test scores.

The interpretative argument consists of a series of inferences that result in a claim being made about each inference. This claim then becomes the input for the next inference in the chain. Kane uses the work of Toulmin (1953) to provide a general structure for an inference. This structure is shown in Figure 2.1. As an example, the initial inference for most tests is the scoring inference by Kane. Scoring refers to the process of translating the data, i.e., observed responses to items, to an observed test score. The *warrant* is the set of rules for scoring the observed responses. The *backing* is the evidence supporting the warrant, or in this case the scoring rules. Depending on the item formats, the backing consists of demonstrating that the scoring rules are appropriate and are applied consistently. If the scoring requires human judgment, as is likely needed for performance-based items, then the quality of that scoring would also be a part of the backing. There is room in this structure for exceptions that may qualify the claim. An example of a *qualifier* is the case where a respondent has not answered enough items to produce an accurate observed score.

FIGURE 2.1 ⬢ Toulmin's Structure of an Inference



Source: Adapted from Kane (2006, p. 28).

In his validity chapter in the fourth edition of *Educational Measurement*, Kane (2006) discussed six inferences. The first three—scoring, generalization, and extrapolation—are common to most test development projects. The last three—implication, theory-based interpretation, and decision—can vary depending on the intended interpretation and use of test scores.

2.7.1 Scoring Inference

The scoring inference moves validation "from observed performance to the observed score" (Kane, 2006, p. 34). In other words, observed performance is usually responses to items. The scoring inference asserts that items are scored appropriately and accurately, and that scoring is free of bias. Also, implicit in the scoring inference is that the items themselves have sufficient technical quality.

The validity argument for the scoring inference includes any evidence pertaining to item scoring. Many item formats are scored objectively, including multiple-choice, true-false, Likert, and semantic differential items. For these, subject matter experts' review of item quality and pilot test item analyses can support the scoring inference. For item formats that require human judgment, such as constructed response, essay, and performance items, additional evidence from scoring rubrics, demonstrations of rater training, interrater reliability, and quality checks of rater effects are needed to support the scoring inference. Furthermore, item scores are aggregated to test scores in a way that needs to be justified. For example, if some items are weighted more heavily than others, the rationale for such weighting needs to be presented.

2.7.2 Generalization Inference

The generalization inference moves validation "from observed score to universe score" (Kane, 2006, p. 34). That is, does the observed score estimate the universe score, the hypothetical score that would be obtained if individuals responded to all possible items and tasks? The key question here is the degree to which the items and tasks chosen for the test are a representative sample from the universe of items and tasks. If that is the case, then the observed score interpretation expands beyond the specific set of items and tasks to the broader universe of generalization.

The validity argument to support the generalization inference includes test specifications and the rationale for the balance of content domains and processes. The specifications ensure that alternate forms are as identical as possible. Statistical sampling theory applied to the selection of items also plays a role in the validity argument. Reliability coefficients indicate the degree of consistency of repeated measurements (see Chapter 4). Standard errors of measurement indicate the precision of observed scores.

2.7.3 Extrapolation Inference

The extrapolation inference moves validation "from universe score to target score" (Kane, 2006, p. 34). This inference is about the relationship between the observed score, now the universe score, and the target domain for the construct being measured. In other words, does the universe score really measure the construct? If so, then implications or interpretations associated with the construct apply to test scores.

Evidence to support the extrapolation inference can come from both analytical and empirical sources. Analytically, the congruence between the universe of generalization and the target domain can be examined. Extrapolation could be undermined if there are parts of the target domain that are systematically excluded from test specifications. For example, if the test uses multiple-choice items only, then aspects of the domain that require an alternate format to be measured, such as an essay to measure writing skill, extrapolation would be challenged. That is, construct underrepresentation could be an issue. Another type of supporting analytical evidence is what Messick called substantive evidence. Think-aloud and cognitive lab protocols could be used to evaluate whether individuals are responding in a way consistent with how the target domain is conceptualized. Empirical evidence can also be obtained in the form of test-criterion relationships, or convergent validity correlations between test scores and scores on measures of the same or similar constructs. Studies that collect these data can come from different populations and settings.

These first three inferences are set as a series of steps. Most of the work establishing the validity argument is done during test development. If the first inference, scoring, is not supported, say if the items are not technically sound, then the interpretative argument breaks down and there is no need to move to the generalization inference. If the extrapolation inference is supported, then there is support for the intended meaning of test scores. The inferences that follow focus on the intended uses of test scores.

2.7.4 Implication Inference

The implication inference moves from the target score (from the extrapolation inference) "to any implications suggested by the construct label or description" (Kane, 2006, p. 43). This inference addresses implications beyond a description of the meaning of test scores. These could include expected relationships with other variables, an expectation of score stability over time, and predicted group differences. Evidence to support such implications can vary widely, such as discriminant coefficients for different constructs, and MTMMs, while construct underrepresentation and construct irrelevant variance can undermine predicted implications. Finally, implications can come in the form of intended and unintended consequences. As discussed above with the Messick framework, unintended consequences, including negative value labels and adverse impact, can affect test score validity.

2.7.5 Theory-Based Interpretation Inference

Theory-based interpretations represent an inference moving from the target score to the construct as defined by theory and any claims associated with the theory. For example, the MBTI was originally developed to support Carl Jung's theory of personality types. As a result, not only did a respondent receive a description of their personality type, but Jung's theory made broad predictions of how someone with that personality type would behave in specific situations.

Both analytic and empirical evidence are needed to support theory-based inferences. Analytically, one can examine the relationships between the items and tasks in the test and the theory behind them. That is, the theory should provide guidance to test development. Empirically, evaluating the test is closely tied to evaluating the theory. A theory may make predictions that can be investigated. Jung's theory predicts that people of a certain personality type will not work well with individuals of a different personality type. That prediction can be tested using scores from the MBTI. The theory behind the construct may also hypothesize nomological networks of a collection of variables. It is possible then to investigate whether data support such networks through various methodologies, including multiple regression, experimental manipulations, path analysis, and structural equation modeling.

2.7.6 Decision Inference

Many tests are used to make decisions about respondents. These include personnel selection and credentialing tests, college admissions exams, and psychological diagnostic tests. In addition, to an interpretation of test scores, validity evidence is also needed for the decision process.

The most common method for determining how scores relate to decisions is *standard setting*. Standard setting is a group decision-making process that attempts to determine one or more test scores, called *cut scores*, that form the boundaries between decision points, such as pass/fail, selected/not selected, and referred for intervention/not referred. Also, tests that provide a criterion-reference score interpretation typically use standard setting to recommend the cut scores that divide performance levels. Kane (1994) provides a framework for evaluating the validity of the standard setting process. Standard setting methods and their validation are discussed in Chapter 12.

Additional evidence to support the decision inference can come from an evaluation of consequences. Incorrect decisions can have severe consequences. Tess Neal and her colleagues examined the use of psychological tests in legal proceedings (Neal, Slobogin, Saks, Faigman, & Geisinger, 2019). They found that about 60 percent of the tests used in court cases had unfavorable reviews of their psychometric properties while at the same time legal challenges to test score validity were relatively rare (about 2.5 percent of the cases). There may also be adverse impact on population subgroups. For example, in state educational achievement testing programs, there is social pressure to have more students achieve the "proficient" performance level. That can lead to "teaching to the test" preparation practices that raise test scores, but the validity of those score increases could be suspect.

As discussed above, a major issue regarding consequential evidence is: who should be responsible for collecting it? The 2014 *Standards* (see Standard 1.25 above) are not clear on this question. Kane (2006) suggests that test developers should be responsible for any claims about the interpretation of test scores for its intended uses, while test users who decide to use a test for some other purpose are responsible for evaluating consequential evidence.

2.8 CURRENT STATE OF VALIDITY AND VALIDATION

Psychometricians continue to debate theories of validity, and we have surely not reached an end state. At present, the best advice is to follow the guidelines in the 2014 *Standards* because this framework for validity has achieved the consensus of three prominent professional organizations associated with test development and use. Those *Standards* are largely based on the Messick validity framework that centers validity around five (or six) types of evidence to support test score interpretations and uses. In the Instrument Development and Validation course I teach, I ask students to create a validation plan for a new test and give them the choice of designing their plan around either the Messick/*Standards* or Kane's framework. To date, every student has chosen the Messick/*Standards* framework. I'm not sure why, but it may be that it is easier for them to conceptualize types of evidence than an interpretative argument. At the same time, many new testing programs, particularly ones measuring cognitive constructs, are using Kane's framework for a validation plan. These include the GED Tests (GEDTS, 2018) and the Test of English as a Foreign Language (TOEFL) (Chapelle, Enright, & Jamieson, 2010), and the Tripod Student Survey of teacher effectiveness (Kuhfeld, 2017). These programs have developed interpretative arguments that include inferences in addition to the ones Kane suggests. Further examples of validation plans are provided in Chapter 13.

By contrasts, older tests may still be using the "toolkit" approach described above. I continue to work with professionals and organizations whose views of validity lie at different points along the evolutionary continuum. I suspect that part of the reason is that what individuals learned in their graduate programs is what they use today. A personal anecdote here: my own doctoral program in the early 1980s professed the trinitarian view of validity. When I taught courses in psychometrics for the first time in the early 2000s, Messick's 1989 chapter and the 1999 *Standards* had been published. I hadn't followed these developments, so I had some (embarrassing) catching up to do.

2.9 CHAPTER SUMMARY

This chapter describes the evolution of the concept of validity from "the degree a test measures what it's supposed to measure" to "the extent to which evidence supports or refutes the proposed interpretation and uses." This process moved validity from a criterion model to the content model to the construct model to the three types of validity (criterion-related, content, and construct) to the current unified view of validity to the possibly near-future argument-based approach. It is as important to understand this evolution as different tests use different models of validation. As students, researchers, and practitioners who work with tests measuring educational and psychological constructs, you are likely to confront all of these validity models at some point.

2.10 EXERCISES AND ACTIVITIES

1. Why is validity a property of the test score and not the test itself?
2. Why have some large-scale credentialing and educational achievement testing programs been drawn to Kane's validity framework?
3. What is the major difference between the "Holy Trinity" validity types and Messick's unified construct validity framework?
4. What responsibilities do you think the test developer has to ensure that adverse unintended consequences from test scores do not occur?
5. Locate the technical manual, validity studies, and/or website of a published test. What validity information is provided by the test developer? Compare that to Messick's and Kane's framework. Is there any important evidence that has not been yet collected?

FURTHER READING

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: American Council on Education and MacMillan.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: Praeger.

REFERENCES

American Psychological Association, American Educational Research Association, & National Council on Measurement Used in Education (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, 51(2, Pt. 2), 1–38.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, 37, 1–15.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer, & H. Braun (Eds.), *Test validity* (pp. 19–32). Hillsdale, NJ: Lawrence Erlbaum.

Auchter, J. C., Sireci, S. G., & Skaggs, G. (1993). *The tests of general educational development: Technical manual*. Washington, DC: American Council on Education.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.

Chapelle, C. A., Enright, M. K., & Jamieson, J.(Spring 2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, 29(1), 3–13.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (1st ed., pp. 691–694). New York, NY: American Council on Education and MacMillan.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, 92(6), 1087–1101.

Garrett, H. E. (1937). *Statistics in psychology and education*. New York, NY: Longmans, Green.

GED Testing Service (2018). *Technical manual: GED Test* (Updated 2018 Edition). Washington, DC: Author.

Goleman, D. (1995). *Emotional intelligence*. New York, NY: Bantam Books.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, 6, 427–438.

Guion, R. M. (1980). On trinitarian conceptions of validity. *Professional Psychology*, 11, 385–398.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64(3), 425–461.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), 319–342.

Kelley (1927). *Interpretation of educational measurements*. Yonkers-on-Hudson, NY: World Book Company.

Kuhfeld, M. (2017). When students grade their teachers: A validity analysis of the Tripod Student Survey. *Educational Assessment*, 22(4), 253–274.

Loevinger, J. (1957). Objective tests as instruments in psychological theory. *Psychological Reports*, 30, 635–694 (Monograph Suppl. 9).

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741–749.

Neal, T. M. S., Slobogin, C., Saks, M. J., Faigman, D. L., & Geisinger, K. F. (2019). Psychological assessments in legal contexts: Are courts keeping "Junk Science" out of the courtroom?. *Psychological Science in the Public Interest*, 20(3), 135–164.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. *Philosophical Transactions of the Royal Society of London*, 187, 253–318.

Rulon, P. J. (1946). On the validity of educational tests. *Harvard Educational Review*, 16, 290–296.

Technical recommendations for psychological tests and diagnostic techniques. (1954). *Psychological Bulletin Supplement*, 51(2 Part 2), 1–38.

Toulmin, S. (1953). *The philosophy of science*. London: Hutchinson's Universal Library.