Pearson, K. (1936). Method of moments and the method of maximum likelihood. *Biometrika*, 28, 34–59.
Quammen, D. (2006). *The reluctant Mr. Darwin: An intimate portrait of Charles Darwin and the making of his theory of evolution*. New York, NY: Atlas Books.
Remmers, H. H., Shock, N. W., & Kelly, E. L. (1927). An empirical study of the validity of the Spearman-Brown formula as applied to the Purdue rating scale. *Journal of Educational Psychology*, 18, 187–195.
Ruch, G. M., Ackerson, L., & Jackson, J. D. (1926). An empirical study of the Spearman-Brown formula as applied to educational test material. *Journal of Educational Psychology*, 17, 309–313.
Sokal, M. M. (1990). James McKeen Cattell and mental anthropometry: Ninetheenth century science and reform and the origins of psychological testing. In M. M. Sokal (Ed.), *Psychological testing and American society, 1890–1930* (pp. 21–45). New Brunswick, NJ: Rutgers University Press.
Spearman, C. E. (1904a). The proofs and measurement of association between two things. *American Journal of Psychology*, 15, 72–101.
Spearman, C. E. (1904b). "General intelligence" objectively determined and measured. *American Journal of Psychology*, 15, 201–293.
Spearman, Charles, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3, 271–295.
Spearman, C. (1927). *The abilities of man*. Oxford, UK: Macmillan.
Spearman, C. (1930a). Autobiography. In C. Murchison (Ed.), *A history of psychology in autobiography*, Vol. 1 (pp. 299–333). Worcester, MA: Clark University Press.
Spearman, C. (1930b). Review of Crossroads in the mind of Man by Truman L. Kelley. *Journal of the American Statistical Association*, 25, 107–110.
Spearman, C. (1930, May 5). Letter to Truman Kelley. Harvard University Archives: Truman Kelley Papers. Cambridge, MA: Harvard University.
Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: The Belknap Press of Harvard University Press.
Thurstone, L. L. (1931). *The reliability and validity of tests*. Ann Arbor, MI: Edwards Brothers, Inc.
Tinker, M. A. (1932). Wundt's Doctorate Students and Their Theses 1875–1920. *The American Journal of Psychology*, 44, 630–637.
Van Wyhe, J. & Rookmaaker, K. (Eds.) (2013). *Alfred Russel Wallace: Letters from the Malay Archipelago*. Oxford, UK: Oxford University Press.
Wallace, A. R. (1858). On the tendency of varieties to depart indefinitely from the original type. *Zoological Journal of the Linnean Society*, 3, 46–50.
Wallace, A. R. (1869). *The Malay Archipelago*, New York, NY: Harper.
Wissler, C. (1901). The correlation of mental and physical tests. *Psychological Review: Monograph Supplements*, 3(6), 1–62.
Wissler, C. (1944). The Contribution of James McKeen Cattell to American anthropology. *Science*, 99, 232–233.
Wolfle, D. (1940). *Factor analysis to 1940*. Psychometric Monograph Number 3. Chicago, IL: University of Chicago Press.
Wood, B. D. (1926). Studies in achievement tests. Part III. Spearman-Brown reliability predictions. *Journal of Educational Psychology*, 17, 263–269.
Yule, G. U. (1897). On the significance of Bravais' formulæ for regression, &c., in the case of skew correlation. *Proceedings of the Royal Society of London*, 60, 359–367.

9

THE EVOLUTION OF THE CONCEPT OF VALIDITY

Michael Kane and Brent Bridgeman[1]

This chapter describes four significant trends in validity theory between 1900 and 2020, but focuses on the period from 1955 to 2010. First, we summarize the development of three traditional models for validity in the first half of the Twentieth Century, the content, criterion, and trait or construct models. Second, we trace the gradual development, between 1955 and 1990, of unified models for validity based on an evolving notion of constructs. Third, we examine the increasing importance of questions about fairness and consequences in the evaluation of testing programs since the 1960s. Fourth, we review the development of general argument-based models that explicitly allow for variability in the kinds of evidence needed for the validation of different kinds of testing programs.

Early Developments, 1890–1950

Kelley (1927) is credited with providing the earliest explicit definition of validity in the context of testing, as the extent to which a test "really measures what it purports to measure" (p. 14) and is appropriate for a "specifically noted purpose" (p. 30), but "validity" had been defined as early as 1921 (Sireci, 2016). However, even before 1921, many of the issues and methods now included under the label, "validity", were invoked. Toward the end of the Nineteenth Century and into the Twentieth Century, researchers sought to develop measures of mental abilities (e.g., general intelligence, memory) that would estimate these abilities more systematically and precisely (i.e., more scientifically) than was ordinarily possible. The abilities were conceptualized mainly in terms of the kinds of tasks assumed to require the mental ability. They were not embedded in an explicitly stated theory, but they were expected to be associated with certain kinds of performance and achievement; for example, students with high intelligence were expected to be

high achievers in school and life. The proposed measures were evaluated in terms of how well they reflected the mental ability of interest (Spearman, 1904).

Early Applications

Tests have always been valued for both the insights they could provide and for their utility in making decisions. By 1920, the possibility of using test scores to predict future performances was recognized. This work on criterion-based predictions focused mainly on applications in selection and placement, with the criteria specified in terms of desired outcomes. Between 1920 and 1950, criterion-related evidence came to be the "gold standard" for validity (Angoff, 1988; Cronbach, 1971; Cureton, 1951; Gulliksen, 1950). The criterion model was supplemented by a content model that focused on the representativeness of the test content. An assessment consisting of a sample of performances from the domain could be used to estimate a test taker's overall domain performance, as in educational achievement tests, and these tests could be validated in terms of their coverage of the domain (Cronbach, 1971: Ebel, 1961; Ryans and Frederiksen, 1951; Rulon, 1946). In both of these models, the "ability" of interest was essentially a given, defined by a criterion to be predicted or a performance domain to be sampled, and the question was whether the test scores provided accurate indications of the ability. Over time, the problems inherent in specifying the ability of interest came to be recognized and more attention was given to explicating the interpretations and uses of scores.

Validity Theory in the Early 1950s

Cureton (1951) began his chapter on validity in the first edition of *Educational Measurement* by associating validity with usefulness for some purpose:

> The essential question of test validity is how well a test does the job it is employed to do. ... Validity is always validity for a particular purpose. It indicates how well the test serves the purpose for which it is used.
>
> *(Cureton, 1951, p. 621)*

Cureton's exposition was highly practical and highly empirical, but he also considered what the observations mean in terms of underlying abilities, suggesting, as an example, that a vocabulary test could be a reasonably valid indicator of verbal intelligence for children with fairly equal opportunities and incentives to learn word meanings, but for children with varied educational backgrounds, "it may be more valid as an indicator of the general quality of previous instruction in reading than as an indicator of verbal intelligence" (Cureton, 1951, p. 621–22). Note that Cureton suggested an interpretation of the scores in terms of a trait, "verbal intelligence", but no theory of intelligence is proposed. He also indicated that validity was not simply a property of the test but depended on the intended interpretation and use of the scores, and the population of test takers.

Although Cureton recognized that the interpretation of test scores would involve assumptions about the meaning of performance regularities, he adopted a strongly empirical, even operational, stance:

> We must not say that his high score is due to his high ability, but if anything the reverse ... His "ability" is simply a summary statement concerning his actions.
>
> *(Cureton, 1951, p. 641)*

Cureton (1951) did not rely on traits to explain behavior, but rather interpreted them as labels for dispositions to behave in a certain way. As noted below, the notion of a "trait" has been commonly used in this way.

The Criterion Model

The criterion model provided a simple methodology for evaluating validity, and as a bonus, it yielded a quantitative index of validity. Gathering the required data could be difficult, but the basic procedure was simple. Obtain scores on the test and on a criterion assessment, preferably a behavioral measure (Anastasi, 1950), for a fairly large and representative sample of persons, who were not selected on the basis of the test scores, and compute a correlation between the two sets of scores.

By the 1950s, the criterion-based methodology had become very sophisticated (Cureton, 1951; Gulliksen, 1950), and if a good criterion were available, it provided a simple, quantitative approach to validation. The main problem with the criterion model was in identifying an appropriate criterion measure. As Ebel (1961) suggested:

> The ease with which test developers can be induced to accept as criterion measures quantitative data having the slightest appearance of relevance to the trait being measured is one of the scandals of psychometry.
>
> *(p.642)*

It can be difficult to obtain a criterion that is clearly better than the assessment itself, and without some way of validating criteria that does not involve other criteria, we face either infinite regress or circularity. One way out of this dilemma is to base the criterion on direct observations of the performance of interest, or a proxy measure validated in terms of its relevance and reliability (Cureton, 1951; Cronbach, 1971; Ebel, 1961; Guion, 1977; Gulliksen, 1950; Kane, Crooks, and Cohen, 1999; Rulon, 1946).

Criterion-related validity evidence continues to be important in making a case for the validity of any interpretation or use that involves inferences from test

scores to some criterion performance (AERA et al., 2014; Zwick, 2006), but criterion-related evidence is now interpreted in a broader context that requires evaluations of fairness and of the positive and negative consequences of test use. (Cronbach, 1988; Guion, 1998; Messick, 1989a).

The Content Model

The content model assumed that a valid measure of competence in a domain of performances could be developed by including samples of the performance of interest in the assessment, or of the skills needed for the performance, and is widely used in assessments of achievement. Rulon (1946) emphasized that the validity of an assessment would depend on the use to be made of the scores and that validity would be dependent on the relevance and appropriateness of the content included in the assessment. Content related analyses are still widely used in validating achievement tests.

The content-based approach was open to criticism, especially if it was not well defined or carefully implemented (Ebel, 1961; Guion, 1977; Messick, 1989a; Sireci, 1998), but a plausible argument could be made for interpreting scores based on samples of performance of some kind in terms of level of skill in that kind of performance (Cronbach, 1971).

Traits

From the 1890s to the early 1950s, assessment scores were often interpreted in terms of *traits*, which were conceived of as tendencies to perform in certain ways in response to certain kinds of stimuli or tasks. According to Cureton (1951), if the item scores on a test correlate substantially, the sum of the item scores can be taken as a measurement of:

> Whatever ... is invoked in common by the test items as presented in the test situation. This "whatever" may be termed a "trait." ... The existence of the trait is demonstrated by the fact that the item scores possess some considerable degree of homogeneity.
>
> *(p. 647)*

Traits could vary in terms of the behaviors involved, in their generality (e.g., spelling vs. intelligence), and in their stability, but they shared three characteristics (Campbell, 1960; Cureton, 1951). First, they were defined in terms of kinds of performance. Second, no theories or causal mechanisms need be specified. Third, most traits were interpreted as enduring attributes of persons.

Traits have played major roles in educational and psychological measurement. The mental abilities discussed in the last section were traits. The true scores of classical test theory (Lord and Novick, 1968), the universe scores of generalizability theory (Brennan, 2001a, 2001b), the latent traits of item-response theory (Lord, 1980), and the factors in factor analysis (McDonald, 1985) reflect traits.

In general, validity theory in the early 1950s was highly empirical, focusing mainly on the criterion model. The content model provided a model for validating interpretations in terms of performance domains, and in addition, provided a way to develop criteria for criterion-based validation. In addition, trait interpretations were widely used, but the traits were generally defined in terms of kinds of tasks or kinds of behavioral dispositions and were validated in terms of relevance and reliability. Validity addressed both score interpretations and score uses (e.g., in prediction and diagnosis). The requirements for validation depended on "how well a test does the job it is employed to do" (Cureton, 1951, p. 621).

Construct Validity, 1954–55

The model proposed by Cronbach and Meehl (1955) in their landmark paper, *Construct Validity in Psychological Tests*, has not been much used in practice, but its central ideas have had a pervasive influence on validity theory. It was introduced mainly to provide a validation framework for the kinds of trait interpretations used in personality theory and in clinical psychology. The APA Committee on Psychological Testing was charged with outlining the kinds of evidence needed to justify the kind of "psychological interpretation that was the stock-in-trade of counselors and clinicians" (Cronbach, 1989, p. 148). They introduced the basic ideas of construct validity, which were incorporated in the *Technical Recommendations* (American Psychological Association, 1954), and were more fully developed in a subsequent paper (Cronbach and Meehl, 1955). As Cronbach (1971) later summarized its origins:

> The rationale for construct validation (Cronbach & Meehl, 1955) developed out of personality testing. For a measure of, for example, ego strength, there is no uniquely pertinent criterion to predict, nor is there a domain of content to sample. Rather, there is a theory that sketches out the presumed nature of the trait. If the test score is a valid manifestation of ego strength, so conceived, its relations to other variables conform to the theoretical expectations.
>
> *(pp. 462–463)*

The goal was to put the validation of traits on firmer ground.

In the early 1950s, the hypothetico-deductive model of theories provided the dominant framework for evaluating theoretical interpretations (Suppe, 1977). This model treated theories as interpreted sets of axioms. The core of the theory consists of axioms, or hypotheses, connecting abstract terms, or "constructs", which were implicitly defined by their roles in the axioms. The axioms and the constructs were interpreted, by connecting some of the constructs to observations, through "correspondence rules" (Suppe, 1977, p.17). The axioms and correspondence rules, and any conclusions derived from these relationships constituted a "nomological

network" (Cronbach and Meehl, 1955). The validity of the theory and the construct interpretations would be evaluated in terms of how well the theory (with the construct measures) handles empirical challenges. If the empirical predictions derived from the network were confirmed, the construct interpretations and the theory would be validated; otherwise, the construct interpretation or the theory, or both, would be questioned.

For Cronbach and Meehl (1955), construct interpretations were based on a scientific theory, which had to be developed, stated clearly, and validated. Both the score interpretation in terms of the construct and the theory were subject to challenge. In effect, Cronbach and Meehl (1955) shifted the focus from the validity of the test for some intended interpretation or use (e.g., as a predictor of some criterion) to the plausibility of the theory-based interpretation and use of the test scores.

The model proposed by Cronbach and Meehl (1955) was very elegant, but it has been very difficult to apply in the social sciences, and Cronbach later expressed regret that he and Meehl had tied their model to a particular view of theories (Cronbach, 1989). Nevertheless, the 1955 paper shaped the subsequent development of validity theory. In the 1940s, the term "construct" was rarely if ever used in testing; by the 1980s everything was a construct.

The Evolution of "Construct" Validity, 1955–1989

After 1955, construct validity evolved in two directions. First, it was viewed as one of three main validation models, along with the criterion and content models for validity, each associated with a particular interpretation or use of scores, and each involving particular kinds of evidence. This approach was labeled the "Trinitarian" model by Guion (1980), but we will refer to it as the *application-specific* approach. Other specific "kinds" of validity were introduced at various times; Newton and Shaw (2013) identified 32 validity modifier labels that have been proposed at various times for different types of validity.

The second direction involved the development of a general unified framework for validation based on a much-relaxed version of the construct model (Cronbach and Meehl, 1955). In developing the unified models, some aspects of the construct model were dropped (particularly, the need for a formal theory specified in terms of a nomological network), and some aspects got generalized and made more central and explicit (e.g., the expectation that rival hypotheses would be considered). As a result, the general unified model that emerged (Messick, 1975) was quite different from the model proposed by Cronbach and Meehl (1955).

Cronbach and Meehl (1955) presented construct validity as a model to be used, "whenever a test is to be interpreted as a measure of some attribute or quality which is not operationally defined" (1955, p. 282), and for "attributes for which there is no adequate criterion" (1955, p. 299). They presented it as an alternate, specific model, but they also suggested that it involved more fundamental concerns, in that, "determining what psychological constructs account for test

performance is desirable for almost any test" (p. 282). They presented construct validity as a fundamental concern, but not as a general framework for validity. The conflict between the application-specific and unified approaches was there from the beginning.

Construct Validity in Science, 1955–1970

In 1957, Jane Loevinger suggested that "construct validity is the whole of the subject from a systematic, scientific point of view" (p.461), because the other models are *ad hoc* and limited to specific uses. Loevinger was a developmental psychologist and was interested in scientific research, a natural setting for the construct model. It is not clear whether Loevinger was advocating for the adoption of construct validity as a general framework for validity in all contexts, or simply emphasizing its utility in scientific research and downplaying questions about more applied uses like selection and achievement testing.

Campbell and Fiske (1959) suggested multitrait-multimethod analyses, in which several traits are each measured using several assessment methods, as a way to evaluate a number of assumptions that are commonly made about traits and trait measures. For example, correlations between measures of a single trait using different methods should be fairly high (i.e., convergent analyses), and correlations between measures of a different traits using a common method should be relatively low (i.e., discriminant analyses).

According to the 1966 Standards (APA, AERA, NCME 1966):

> Tests are used for several types of judgment, and for each type of judgment, a different type of investigation is required to establish validity. ... The three aspects of validity corresponding to the three aims of testing may be named content validity, criterion-related validity and construct validity.
>
> *(p. 12)*

and

> *Construct validity* is ordinarily studied when the tester wishes to increase his understanding of the psychological qualities being measured by the test.
>
> *(p. 13)*

The 1966 Standards adopted an application-specific approach, with construct validity focused on psychological traits.

Softening the Construct-Validity Model – Cronbach (1971)

In his chapter in the second edition of *Educational Measurement*, Cronbach (1971) continued to associate construct validation with theoretical variables for which

"there is no uniquely pertinent criterion to predict, nor is there a domain of content to sample" (p. 462), and suggested that, "A description that refers to the person's internal processes (anxiety, insight) invariably requires construct validation" (p. 451).

Cronbach (1971) also discussed the need for an overall evaluation of validity, which would include many kinds of evidence, including construct-related evidence:

> Validation of an instrument calls for an integration of many types of evidence. The varieties of investigation are not alternatives any one of which would be adequate. The investigations supplement one another... For purposes of exposition, it is necessary to subdivide *what in the end must be a comprehensive, integrated evaluation of the test.*
>
> *(Cronbach, 1971, p445; italics in original)*

Cronbach (1971) criticized some programs of construct validity as, "haphazard accumulations of data rather than genuine efforts at scientific reasoning" and suggested that:

> Construct validation should start with a reasonably definite statement of the proposed interpretation. The interpretation will suggest what evidence is most worth collecting to demonstrate convergence of indicators. A critical review in the light of competing theories will suggest important counter-hypotheses, and these also will suggest data to collect. Investigations to be used for construct validation, then, should be purposeful rather than haphazard (Campbell, 1960).
>
> *(Cronbach, 1971. p. 483)*

This echoes Cronbach and Meehl, but it is much softer. The talk of theories and nomological networks is replaced with talk of "a reasonably definite statement of the proposed interpretation". Cronbach (1971) envisioned a structured and unified conception of validity that was later more fully elaborated by Messick (1989a) and Kane (2006). But even with his looser and more comprehensive conception of construct validity, Cronbach (1971) maintained that Loevinger's suggestion that claims of content validity be dropped in favor of construct validation was sound in some contexts, but "much too sweeping" (p. 454).

The 1974 Standards

The 1974 Standards (APA, AERA, & NCME 1974) defined validity in terms of "what may properly be inferred from a test score" (p. 25), a general, unified ideal, but it discussed validation in terms of an expanded set of "four interdependent kinds of inferential interpretation" (p.26): *predictive, concurrent, content*; and *construct*

validities. The construct-validity model was to be reserved for measures of theoretical constructs, where the construct is "a dimension understood or inferred from its network of interrelationships" (p. 29).

Meaning and Values in Measurement – Messick (1975)

Messick (1975) quoted Loevinger (1957) to the effect that, from a scientific point of view, construct validity is the whole of the subject, and he maintained that, in contrast with more specific models that focus on specific interpretations and uses, construct validation involves hypothesis testing and "the philosophical and empirical means by which scientific theories are evaluated" (p. 956):

> Construct validation is the process of marshalling evidence in the form of theoretically relevant empirical relations to support the inference that an observed response consistency has a particular meaning. The problem of developing evidence to support an inferential leap from an observed consistency to a construct that accounts for that consistency is a generic concern of all science.
>
> *(Messick, 1975, p. 955)*

Messick (1975) was still treating the construct validity model as the first among other models, as a generic concern in science, rather than a general framework, and he focused on its use in scientific contexts.

Messick (1975) was also loosening the idea of a construct. He suggested that in order to evaluate an interpretation or use of scores, it is necessary to be clear about the construct meanings and associated values, but he did not require that the construct be embedded in a theory. In broadening thinking about constructs, he drew attention to the importance of values and consequences, and suggested that, in considering any test use, two questions were of central concern:

> First, is the test any good as a measure of the characteristic it is interpreted to assess? Second, should the test be used for the proposed purpose? The first question is a technical and scientific one and may be answered by appraising evidence bearing on the test's psychometric properties, especially construct validity. The second question is an ethical one, and its answer requires an evaluation of the potential consequences of the testing in terms of social values.
>
> *(Messick, 1975, p. 960)*

Messick was strongly committed to the importance of values throughout his career.

Messick (1975) gave the evaluation of plausible rival hypotheses a central role in validation and concluded that, "If repeated challenges from a variety of

plausible rival hypotheses can be systematically discounted, then the original interpretation becomes more firmly grounded" (Messick, 1975, p. 956), and he suggested that convergent and discriminant analyses could be used to rule out alternate hypotheses (Campbell and Fiske, 1959).

Embretson (1983) drew an insightful distinction between two kinds of interpretation: *construct representation* refers to the model-specific processes and structures (i.e., a cognitive theory of performance) that can be used to account for test taker performances, and *nomothetic span* refers to the network of relationships that support inferences to other variables. Both of these theory-based interpretations can provide a basis for construct validation, but the kinds of evidence needed to validate the interpretations differ. There are contexts where one of these two theory types predominates and so any strong version of construct validity may not provide a unified framework for validation.

Applications of Construct Validity in the 1970s

Confirmatory factor analysis (Jöreskog, 1973) can be interpreted in terms of Cronbach and Meehl's (1955) model for construct validation. The confirmatory factor model postulates relationships between latent variables, or constructs, with some theory-based constraints on the factor structure, and the model is checked by fitting it to appropriate empirical data. If the model does not fit the data, either the postulated assumptions or the validity of the assessments must be questioned.

In 1979, the federal agencies responsible for enforcing civil-rights laws published Uniform Guidelines (EEOC et al., 1979), which promoted the use of criterion-related evidence for the validation of employment tests. The Guidelines allowed for the use of content-based and construct-based analyses, but preferred criterion-related analyses, and thus enshrined an application-specific framework in legal analyses of fairness in employment testing.

In practice, construct validity was not treated as a general, unified framework for validity in the 1970s, and when it was used to evaluate testing programs, it was rarely applied in a rigorous way. As Cronbach lamented:

> The great run of test developers have treated construct validity as a wastebasket category. In a test manual, the section with that heading is likely to be an unordered array of correlations with miscellaneous other tests and demographic variables. Some of these facts bear on construct validity, but a coordinated argument is missing.
>
> *(Cronbach, 1980b, p. 44)*

Although Messick, Cronbach, and others were moving toward a more general, unified conception of validity, practice still focused on specific models tied to specific interpretations and uses. According to Angoff (1988):

> In essence then, validity was represented, even well into the 1970s as a three-categorized concept and taken by publishers and users alike to mean that tests could be validated by any one or more of the three general procedures.
>
> *(Angoff, 1988, p. 25)*

Validity theorists (Anastasi, 1986; Cronbach, 1980a; Guion, 1977, 1980; Messick, 1975, 1980) were concerned that the separate models in the application-specific approach did not provide any clear, consistent standards for validity, but practice continued to focus on the application-specific models.

The 1985 Standards – Victory (of Sorts) for the Unified View

The 1985 Standards (AERA, APA, and NCME, 1985) characterized validity as a "unified concept", while accepting that different interpretations and uses would require different kinds of evidence:

> Validity, however, is a unitary concept. Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from the scores.
>
> *(p. 9)*

Validity was taken to be a unitary concept, but the introduction to the chapter was divided into sections for different kinds of evidence: background, construct-related evidence, content-related evidence, criterion-related evidence, validity generalization, and differential prediction.

Evidence in the "construct-related category" would focus on the "psychological characteristic of interest" (AERA et al., 1985, p. 9) and the construct should be embedded in a conceptual framework:

> The conceptual framework specifies the meaning of the construct, distinguishes it from other constructs, and indicates how the measure of the construct should relate to other variables.
>
> *(AERA et al., 1985, pp. 9–10)*

The discussion of construct-related evidence remained close to Cronbach and Meehl (1955), with a "conceptual framework" instead of a formal theory. At about the same time, Anastasi (1986) suggested that

> content analyses and correlations with external criteria fit into particular stages in the process of construct validation, that is, in the process of both determining and demonstrating what a test measures.
>
> *(Anastasi, 1986, p.4)*

The 1985 Standards and much of the subsequent literature on validity theory defined validity as a unitary concept but did not provide much guidance on how to combine different kinds of evidence in validation (Moss, 1995). The evidence was to span the three traditional categories, more evidence would be better than less, quality was important, and the evidence should be chosen in light of intended use, but there was little explicit guidance on how all of this was to be done.

The 1999 *Standards for Educational and Psychological Testing* continued in this vein, defining validity as

> the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests. ... The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations.
>
> *(AERA, APA, NCME, 1999, p. 9)*

Validity theorists wanted a more unified, principled, and consistent approach to validation.

A major development during the 1980s that did provide explicit guidance for validation was an increasing emphasis on empirical challenges to proposed interpretations of test scores. The notion of systematic error has a long history in the physical sciences, but it became especially relevant and explicit in validity theory in the 1980s and 90s (Cook and Campbell, 1979; Messick, 1989a; AERA et al., 1999), in terms of two kinds of systematic errors. Messick (1989a, p. 34) defined *construct-irrelevant variance* as "excess reliable variance that is irrelevant to the interpreted construct," (p. 34), and he defined *construct underrepresentation* as occurring if "the test is too narrow and fails to include important dimensions or facets of the construct" (p. 34). These two types of systematic error reflect construct validity's focus on challenging proposed construct interpretations, empirically and conceptually. If any serious source of construct-irrelevant variance or construct underrepresentation is found, or plausibly suspected, the intended interpretation is undermined.

The Strong and Weak Programs of Construct Validity, 1988–89

In chapters published in 1988 and 1989, Cronbach drew a distinction between the original *strong program* of construct validity and a *weak program* of construct validity:

> Two concepts of CV were intermingled in the 1954 *Standards*: a strong program of hypothesis-dominated research, and a weak program of *Dragnet* empiricism: "just give us the facts, ma'am ... any facts". The CM paper unequivocally sets forth the strong program: a construction made explicit, hypotheses deduced from it, and pointed relevant evidence brought in. This is also the stance of the 1985 *Standards*.
>
> *(Cronbach, 1989, p. 162; italics and abbreviations in original)*

And he favored the strong program:

> The strong program ... calls for making one's theoretical ideas as explicit as possible, then devising deliberate challenges. Popper taught us that an explanation gains credibility chiefly from falsification attempts that fail.
>
> *(Cronbach, 1988, pp. 12–13)*

Cronbach (1988) and Anastasi (1986) explicitly maintained that validity theory had gotten beyond the application-specific approach, but as Moss (1992) noted:

> To this day, most of the popular measurement text books, like the 1985 Standards, continue to organize presentations of validity around the three-part traditional framework of construct-, content-, and criterion-related evidence.
>
> *(p.232)*

The tension between calls for a unified framework, and the diversity inherent in different interpretations and uses had not been resolved.

General Principles Derived from Construct Validity, 1955–1989

Although the original, strong version of construct validity (Cronbach and Meehl, 1955) did not get applied much, it yielded three general principles that shaped the development of validity theory. First, effective validation requires that the proposed score interpretation and use be specified well enough that testable hypotheses can be derived from it. For a test score to have any meaning, it must make testable claims about the test taker.

Second, just as scientific theories are evaluated in terms of their ability to withstand serious challenges, proposed interpretations are to be evaluated against alternate interpretations (Cronbach, 1971, 1980a, 1980b, 1988; Embretson, 1983; Messick, 1989a).

Third, validation requires a program of research that investigates the claims being made and any counterclaims and their supporting assumptions, rather than a single validation study.

Messick's Unified, but Faceted, Model, 1989

In his *Educational Measurement* chapter, Messick (1989a) provided a unified framework for validity based on a broadly defined version of the construct validity model. He defined validity as

> 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)*

Note two significant departures from Cronbach and Meehl's (1955) version of construct validity. First, the definition covers actions as well as inferences, while Cronbach and Meehl focused on theory-based interpretations and not on actions. Messick was a consistent advocate for including the evaluation of consequences in validation as an "integral part of validity" (Messick, 1989a, p. 84).

Second, there is no mention in the definition of a theory that defines the construct, or of nomological networks. Nomological networks are discussed by Messick (1989a), but they do not get a lot of attention. Messick (1989a) defined construct validity much more broadly than Cronbach and Meehl (1955), and took construct validity to be the evidential basis of both score uses and interpretations:

> The construct validity of score interpretation undergirds *all* score-based inferences not just those related to interpretive meaningfulness but also the content- and criterion-related inferences specific to applied decisions and actions based on test scores.
>
> *(pp. 63–64)*

He focused on the need to provide evidence for the "trustworthiness of score interpretation in terms of explanatory concepts that account for both test performance and relationships with other variables" (Messick, 1989a, p. 34).

Messick's (1989a) unified, construct-based framework for validity had at least two significant problems. First, Messick's presentation of his framework is dense making it difficult to apply. Messick organized the different kinds of evidence for validity in terms of a two-by-two matrix with the function of testing (interpretation or use) and the justification for testing (evidence or consequences) as the two dimensions. The table was not used much in theoretical discussion or applications of validity in part because of substantial overlap among the cells (Kane and Bridgeman, 2017).

Second, there is some conflict within the framework. Messick was emphatic about the central role to be played by construct validity, but many examples of more-or-less acceptable validations do not require the strong program of construct validity. For example, Guion (1977) made the case that it would be reasonable to interpret scores on a sample of tasks drawn from a domain, as a measure of skill in the domain. Messick seemed to assume that, within the unified framework, the same construct-based methodology would be applicable to all cases. In this vein, Messick (1988) expressed concern that the comment attached to the first validity standard in the 1985 edition allowed for the use of different types of evidence in the validation of different testing programs. He maintained that under a unified approach based on the construct-validity model, validations should follow a consistent pattern; to the extent that different types of evidence are used for different testing programs, we have an application-specific framework. Messick developed his framework from the late 1950s to the late 1980s, as validity theory moved toward unity, while practice tended to be application-specific.

The Stubborn Conflict between Theoretical Unity and Applied Diversity

The conflict between the theoreticians' desire for a unified framework for validity and the recognition of the great diversity in the goals and contexts of assessment programs continued from 1955 to 1989, and as discussed later, is still with us (Sireci, 2009). Messick made an effort to unify validity under the construct model, but his elegant formulation did not resolve the conflict; in place of lists of kinds of validity, we had lists of kinds of validity evidence, with different mixes of evidence used for different applications. One major problem with basing the unified framework on the construct model is the strong association between "construct validity" and theoretical interpretations and theory-based uses; many testing application focus on practical questions (e.g., how well has the test taker mastered some domain?, how can we predict test taker's performance in some future activity?) that do not rely much on theory.

The Role of Consequences in Validity, 1970

The achievement of the intended outcomes of testing programs has been a fundamental concern in validity theory from the early Twentieth Century, especially in criterion-related applications, but unintended, negative consequences got far less attention.

Adverse Impact as a Negative Consequence

Before 1960, adverse impact across groups (racial, ethnic, gender) was not given much attention in testing, because a test was considered fair if all test takers performed the same tasks under the same conditions and were graded in the same way. Neither "bias" nor "fairness" was listed in the index for the first edition of *Educational Measurement* (American Council on Education, 1951), but by the mid-1970s, bias had become a major concern in assessment.

Messick (1982b) criticized a National Academy of Science report on ability testing, because it "evinces a pervasive institutional bias" (p. 9), by focusing on the intended outcomes of decision rules:

> Our traditional statistics tend to focus on the accepted group and on minimizing the number of poor performers who are accepted, with little or no attention to the rejected group or those rejected individuals who would have performed adequately if given the chance.
>
> *(p. 10)*

The evaluation of how well score-based decisions achieve their intended outcomes was a well-established expectation, but Messick was suggesting that the applicant's welfare also merits attention. Cronbach (1988) made a similar point about narrow inquiries that "concentrate on predicting a criterion the employer cared about" (p. 7), and neglect concerns about the applicants who are rejected.

Once adverse impact got sustained attention, it was natural to think about other potential negative consequences. It was soon recognized that assessment programs can have a substantial impact on educational institutions, curricula, and students (Crooks, 1988; Frederiksen, 1984; Madaus, 1988; Lane & Stone, 2006).

Messick and Cronbach on Consequences

Both Messick (1975, 1989a, 1995) and Cronbach (1971, 1980b, 1988) included the evaluation of consequences within validity, but they saw consequences as playing different roles in validity (Moss, 1992). Messick saw the evaluation of consequences as an aspect of construct validity, because the consequences of score use "both derive from and contribute to the meaning of test scores" (Messick, 1995, p.7), and negative consequences count against validity if they are due to construct-irrelevant variance or construct under-representation. For Messick, unanticipated negative consequences suggest a need for a more thorough analysis of possible sources of construct-irrelevant variance or construct under-representation, but they would not necessarily count against the validity of the scores.

In contrast, Cronbach (1971, 1988) suggested that negative consequences could invalidate score use even if the consequences were not due to any problem with the assessment, because, "tests that impinge on the rights and life chances of individuals are inherently disputable" (Cronbach 1988 p. 6). Cronbach (1988) also maintained that we "may prefer to exclude reflection on consequences from meanings of the word *validation*, but ... cannot deny the obligation" (p. 6). That is, bad consequences do not necessarily invalidate the proposed interpretation of test scores, but they do count against test use, even if the interpretation is well supported.

After referencing Cronbach (1988), Messick (1989b) countered that

> the meaning of validation should not be considered a preference. On what can the legitimacy of the obligation to appraise social consequences be based if not on the only genuine imperative in testing, namely, validity.
>
> *(p. 11)*

Cronbach's insistence on evaluating all consequences probably flowed from his involvement in program evaluation. Cronbach (1982) "advocated investigating what is important, whether or not the questions fit conventional paradigms" (p. xvi). Messick (1989a) developed a general scientific framework for validity, with a primary focus on construct validity and a strong but secondary emphasis on consequences

while Cronbach (1971, 1988) favored a more pragmatic approach, with a more direct focus on consequences. Kane (2006) tends to agree with Cronbach (1988) that negative consequences can invalidate score uses even if they are not due to any flaws in the assessment.

Ironically perhaps, Messick got more criticism than Cronbach for advocating the role of consequences in validation, but arguably, Cronbach gave consequences a stronger role in the evaluation of assessment programs (Moss, 1998). Cronbach's position was less objectionable to critics who were willing to attend to negative consequences but did not want to include them under the heading of validity.

The 1990s Consequences Debates

In the 1990s, several authors (Mehrens, 1997; Popham, 1997) argued against the inclusion of consequence under validity. The critics generally agreed that consequences are relevant to the evaluation of testing programs but wanted to have validity be as objective and value-free as possible. Consequences were to be evaluated but not under the heading of validity. Others (Linn, 1997; Moss, 1998; Shepard, 1997) favored a broader conception of validity, which would include evaluations of positive and negative consequences of score use. Everyone seemed to agree that consequences should be a central concern in deciding whether to use a test in a particular way, but they disagreed about whether this concern should be addressed under the heading of validity. This debate about the role of consequences in validity theory has continued into the 21st century (Bachman and Palmer, 2010; Cizek, 2012).

Unity and Specificity, 2000–2020

Messick's construct-based framework was unified, but it did not provide clear guidance for validation, and a number of general and specific approaches have since been developed to fill the gap between theory and practice. The general frameworks are flexible and conditional and explicitly require different kinds of evidence for different interpretations and uses. The specific models focus on particular interpretations and uses.

General Argument-Based Frameworks, 1988

Cronbach (1988) relied on principles from program evaluation (House, 1980) in developing an argument-based framework for validity:

> I propose here to extend to all testing the lessons from program evaluation. What House ... called "the logic of evaluation argument" applies, and I invite you to think of "validity argument" rather than "validation research".
>
> *(p. 4)*

The *validity argument* would include the evidence for and against the claims inherent in the proposed interpretation and use. The argument was to "make clear, and to the extent possible, persuasive, the construction of reality and the value weightings implicit in a test and its application" (Cronbach, 1988, p.5).

Kane (1992) added the idea of an *interpretative argument*, "with the test score as a premise and the statements and decisions involved in the interpretation as conclusions" (p.527), as a way of specifying the claims that need to be evaluated, and therefore, the kinds of evidence needed for validation. This argument-based approach also provided criteria for deciding when the interpretation and use were adequately supported, that is validated (Crooks, Kane & Cohen, 1996). If the argument were coherent and complete and its inferences were plausible, the interpretation/use could be considered valid. If any part of the argument were not plausible, the interpretation/use would not be considered valid (Haertel, 1999; Kane, 1992; Shepard, 1993).

Bachman and Palmer (2010) proposed an argument-based framework for assessment development and justification that emphasized score uses and the consequences associated with score uses in terms of an Assessment Use Argument (AUA):

> The AUA consists of a set of claims that specify the conceptual links between a test taker's *performance*, ... an *interpretation* about the ability we want to assess, the *decisions* that are to be made, and the *consequences* of using the assessment and of the decisions that are made.
>
> *(p.30)*

Bachman and Palmer based their framework on "the need for a *clearly articulated and coherent Assessment Use Argument* (AUA)" and on "the *provision of evidence* to support the statements in the AUA" (p.31). Following Bachman and Palmer's work, Kane (2013) updated his terminology, by replacing "interpretive argument" by *interpretation/use argument*, or IUA, in order to give more emphasis to the role of uses and consequences.

These argument-based frameworks are intended to retain the rigor in the strong program of construct validity, while making validation more straightforward (Cronbach, 1988; Kane, 2006; Chapelle, Enright & Jamieson, 2008; Bachman and Palmer, 2010), by making the claims to be validated explicit. One specifies the claims being made in some detail and then evaluates these claims. The inferences and assumptions would be subjected to empirical challenges, and if they survive all serious challenges, the interpretations and uses would be considered plausible, or valid. The most questionable parts of the argument should be the focus of the empirical challenges (Cronbach, 1988).

The claims may involve a theory, or they may consist of a more loosely defined set of inferences. At the very least, some assumptions about the generalizability of the scores (over tasks, occasions, contexts, task formats, or time limits?) will be inherent in the interpretation and use of the scores. If the interpretation or use assumes that the scores will be related to other variables, these relationships can be checked. If the scores are to be used to predict some criterion, the accuracy of the predictions can be checked, but note that, if the interpretations and uses under consideration do not involve prediction, then predictive evidence is irrelevant to the validity of these interpretations and uses. Claims that are not inherent in the proposed interpretation and use of the scores can be ignored, and evidence for such irrelevant claims does not strengthen the validity argument.

The argument-based frameworks are quite general and unified in that they impose the same three general requirements for validation on all testing programs: (1) specify the claims being made, (2) verify that the claims accurately represent the interpretation and use of the scores, and (3) verify that the claims are plausible by challenging them empirically.

The chapter on validity in the most recent Standards (AERA, APA, NCME, 2014) is consistent with these argument-based approaches. It calls for a clear statement of the proposed interpretations and uses of the scores, and the second standard requires that:

> A rationale should be presented for each intended interpretation of test scores for a given use, together with a summary of the evidence and theory bearing on the intended interpretation.
>
> *(AERA, APA, NCME, 2014, p.23)*

However, most of the discussion is organized in terms of five kinds of evidence (evidence based on test content, on response processes, on internal structure, and on relations to other variables, as well as evidence for validity and consequences of testing), and not in terms of the inferences to be evaluated.

Recent Application-specific Models for Validity

As we have noted, application-specific models continued to be popular in practice, even as validity theory became more general and unified. In addition, some application-specific validity models have been put forward as definitions of validity, and in doing so, they propose to restrict the term, "validity" to specific kinds of interpretations.

Borsboom, Mellenbergh & Van Heerden (2004) suggested that:

> a test is valid for measuring an attribute if and only if (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure (p.1016).

They define validity in terms of causal explanations and do not include uses or consequences, or any other variables in their definition (Sireci, 2016). As Holland

(1986) points out, the causal impact of a trait, or construct, on assessment performances cannot be directly demonstrated empirically. The causal inference will generally have to be evaluated indirectly using the strong form of construct validity (Cronbach and Meehl, 1955). In practice, the strong form of construct validity has continued to be popular in contexts like psychological research where theory is of primary interest (e.g. Loevinger 1957; Embretson, 1983), and where explanations of performance are the main objective (Zumbo, 2009).

Lissitz and Samuelsen (2007) raised questions about Messick's (1989a) framework (particularly its complexity) and suggested a framework and terminology that focused on the content and structure of the test as a definition of "the trait of interest" (p. 441). Their examples come particularly from achievement tests and their model makes sense for this kind of interpretation. They suggested a radical simplification of the scope of validity to something like Cureton's (1951) relevance and reliability.

Mislevy and his colleagues (Mislevy, Steinberg, & Almond, 2003) proposed an Evidence Centered Design (ECD) approach to the development and evaluation of assessments that relies on formal, probability-based models (particularly Bayes nets) and reasoning based on such models. Applications of ECD start with an analysis of the construct of interest, and they use student models and task models to develop assessment tasks that would generate the kinds of evidence needed to support the intended inferences (Mislevy et al., 2003). The ECD is akin to the argument-based frameworks in that it starts with a detailed specification of the construct and then seeks to develop evidence for the claims being made. As its name suggests ECD is focused more on assessment design than on validation as such, but it clearly has implications for validity as well. More recently, Mislevy (2018) has proposed a very ambitious sociocognitive approach, which is likely to be applicable mainly in educational contexts, because it assumes fairly rich background knowledge about test takers for its full implementation.

That new validation frameworks tied to particular interpretations or uses continue to be developed (Krupa, Carney, and Bostic, 2019) should not be surprising. If we adopt a model for validating assessments for some kind of use we can prescribe the kinds of evidence needed in some detail. If we adopt a unified framework for validity that is to apply to all cases (e.g., professional licensure examinations, and diagnostic assessments for the subtraction of fractions), it cannot be very prescriptive, because the use cases are so different. The more general the model, the more conditional it is likely to be. For this reason, the various editions of the Standards, which are intended to cover essentially all kinds of educational and psychological assessments, are highly contingent; most of the specific standards start with a "when" or an "if".

The argument-based approach to validity can accommodate various application-specific models, including the traditional content, criterion, and construct models. If the IUA focuses on level of achievement in some content domain, the validity argument would rely on evidence for content coverage and generalizability, as in Lissitz and Samuelsen's (2007) model. If a causal (Borsboom et al., 2004) or explanatory (Zumbo, 2009) model is adopted, the validity argument is likely to include the strong program of construct validity (Cronbach, 1988)

Concluding Remarks

The tools, models, and analytic techniques available to the validator have expanded greatly, as has the range of applications of testing programs. Before 1920, the focus was on mental abilities, but now, a wide range of assessments targeted on a variety of uses need to be validated, and it is assumed that multiple lines of evidence will be involved in the validation.

The four main trends in the history of validity theory that we have traced are: first, the development of several models for validity in the first half of the last century, in particular, the content, criterion, and trait models; second, the gradual development, during the second half of the last century, of unified models that subsumed the specific models under increasingly broad conceptions of construct validity; third, the development of a clear sense of the importance of fairness and consequences in the evaluation of testing programs since the 1960s, and fourth, the development of general argument-based models that explicitly allow for variability in the kinds of evidence needed for the validation of different kinds of testing programs.

Naturally, successive frameworks for validity are shaped in part by the issues that seem most pressing at the time, and by the background and interests of those who propose them. The content model was designed for achievement tests, the criterion model to validate inferences from test scores to other variables (e.g., future performance), and the construct model for measures used in clinical contexts and in scientific research. The focuses on fairness and consequences arose from the need to justify selection, placement, and licensure programs. The unified models were designed to bring order to the large set of application-specific models that were developed to address particular interpretations and uses of test scores.

A second impetus to the development of new models was a dialogue between validity theorists. Messick (1989a) sought to make Cronbach's 1971 formulation less pragmatic, and more scientific. When Michael Zieky asked Messick about the intended audience for his 1989 chapter, Messick replied, "Lee Cronbach" (Kane and Bridgeman, 2017, p.522). Cronbach's (1988) notion of validity argument can be read as advocating a more pragmatic approach than that in Messick's (1975, 1980) unified model, and Kane (1992, 2006, 2013) was if anything even more pragmatic. Borsboom et al. (2004) advocate a radical simplification of validity theory in reaction to the complexity of Messick's (1989a) formulation, and Mislevy (2018) proposed a more structured approach to the development and validation of construct assessments than that provided by Messick.

This conversation between practice and theory and among theorists will go on as new applications and problems arise and as old ones are revisited, and validity theorists are likely to be arguing about issues of bias, fairness, and consequences, as long as test scores are used to make life-altering decisions.

Note

1 The authors wish to thank Suzanne Lane and Stephen Sireci for their review and helpful comments on an earlier draft of this chapter.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

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.

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

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education (1966). *Standards for Educational and Psychological Tests and Manuals*. Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education (1974). *Standards for Educational and Psychological Tests and Manuals*. Washington, DC: American Psychological Association.

Anastasi, A. (1950). The concept of validity in the interpretation of test scores. *Educational and Psychological Measurement*, 10(1), 67–78.

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. 9–13). Hillsdale, NJ: Lawrence Erlbaum.

Bachman, L. & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111, 1061–1071.

Brennan, R. (2001a). An Essay on the history and future of reliability from the perspective of replications. *JEM*, 38(4), 295–317.

Brennan, R. (2001b). *Generalizability theory*. New York, NY: Springer-Verlag.

Campbell, D. T. (1960). Recommendations for APA test standards regarding construct, trait, or discriminant validity. *American Psychologist*, 15(8), 546–553.

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. (Eds.), (2008). *Building a validity argument for the test of English as a foreign language*. New York, NY: Routledge.

Cizek, G. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods*, 17(1), 31–43.

Cook, T. & Campbell, D. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin.

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

Cronbach, L. J. (1980a). Validity on parole: How can we go straight? New directions for testing and measurement: Measuring achievement over a decade. *Proceedings of the 1979 ETS Invitational Conference* (pp. 99–108). San Francisco, CA: Jossey-Bass.

Cronbach, L. J. (1980b). Selection theory for a political world. *Public Personnel Management*, 9(1), 37–50.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco, CA: Jossey-Bass.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. E. Linn (Ed.), *Intelligence: Measurement, theory, and public policy* (pp. 147–171). Urbana, IL: University of Illinois Press.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.

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

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, 58, 438–481.

Crooks, T., Kane, M. & Cohen, A. (1996). Threats to the valid use of assessments. *Assessment in Education*, 3, 265–285.

Cureton, E. E. (1951). Validity. In E.F. Lingquist (Ed.), *Educational measurement*. Washington, DC: American Council on Education.

Ebel, R. (1961). Must all tests be valid? *American Psychologist*, 16, 640–647.

Embretson (Whitely), S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93, 179–197.

Equal Employment Opportunity Commission (EEOC), Civil Service Commission, Department of Labor, and Department of Justice (1979). Adoption by four agencies of Uniform Guidelines on Employee Selection Procedures. *Federal Register*, 43, 38290–38315.

Fredericksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist*, 39, 193–202.

Guion, R. (1977). Content validity: The source of my discontent. *Applied Psychological Measurement*, 1, 1–10.

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

Guion, R. (1998). *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Erlbaum.

Gulliksen H. (1950) *Theory of mental tests*. New York, NY: Wiley. Republished 1987 by Lawrence Erlbaum, Hillsdale, NJ.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice*, 18(4), 5–9.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81, 945–960.

House, E. R. (1980). *Evaluating with validity*. Beverly Hills, CA: Sage Publications.

Jöreskog, K. (1973). A general method for investigating a linear structural equation system. In A. Goldberger & D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 85–112). New York, NY: Academic Press.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64), Westport, CT: American Council on Education and Praeger.

Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin*, 112, 527–535.
Kane, M. (2013). Validating the Interpretations and Uses of Assessment Scores. *Journal of Educational Measurement*, 50, 1–73.
Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement*, 6, 125–160.
Kane, M. T., Crooks T. J., & Cohen, A. S., (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, 18(2), 5–17.
Kane M. & Bridgeman B. (2017). Research on Validity Theory and Practice at ETS. In R. Bennett and M. von Davier (Eds.), *Advancing Human Assessment. Methodology of Educational Measurement and Assessment*. Cham: Springer. https://doi.org/10.1007/978-3-319-58689-2_16.
Kelley, T. (1927). *Interpretation of educational measurements*. Yonkers, NY: World Book.
Krupa, E., Carney, M. & Bostic, J. (2019). Argument-based validation in practice: Examples from mathematics education. *Applied Measurement in Education*, 32, 1–9.
Lane, S., Parke, C., & Stone, C. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement: Issues and Practice*, 17(2), 24–28.
Lane, S. & Stone, C. (2006). Performance assessment. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387–431), Westport, CT: American Council on Education and Praeger.
Linn R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice*, 16(2), 14–16.
Lissitz, R., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36, 437–448.
Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, Monograph Supplement*, 3, 635–694.
Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Lawrence Erlbaum Associates.
Lord, F., & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
McDonald, R. P. (1985). *Factor analysis and related methods*. Hillsdale, NJ: Erlbaum.
Madaus, G. F. (1988). The influences of testing on the curriculum. In L. N. Tanner (Ed.), *Critical issues in curriculum* (pp. 83–121). Chicago, IL: University of Chicago Press.
Markus, K. & Borsboom, D. (2013). *Frontiers of test validity theory; measurement, causation, and meaning*. New York, NY: Routledge.
Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice*, 16(2), 16–18.
Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 30, 955–966.
Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, 35, 1012–1027.
Messick, S. (1982). The values of ability testing: Implications of multiple perspectives about criteria and standards. *Educational Measurement: Issues and Practice*, 1(3), 9–12, 20.
Messick, S. (1988). The once and future issues of validity. Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale, NJ: Erlbaum.
Messick, S. (1989a). Validity. In R. L. Linn (Ed.), *Educational Measurement*, 3rd ed. (pp. 13–103) New York, NY: American Council on Education and Macmillan.

Messick, S. (1989b). Meaning and values in test validation: The science and ethics of assessment. *Educational Reseacher*, 18(2), 5–11.
Messick, S. (1995). Standards of validity and the validity of and standards in performance assessment. *Educational Measurement: Issues and Practice*, 14(4), 5–8.
Mislevy, R., Steinberg, L., & Almond, R. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3–62.
Mislevy, R. (2018). *Sociocognitive foundations of educational measurement*. New York, NY: Rutledge.
Moss, P. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, 62, 229–258.
Moss, P. (1995). Themes and variations in validity theory. *Educational Measurement: Issues and Practice*, 4(2), 5–13.
Moss, P.A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice*, 17(2), 6–12.
Newton, P. E., & Shaw, S. D. (2013). Standards for talking and thinking about validity. *Psychological Methods*, 18(3), 301–319.
Popham, W.J. (1997) Consequential validity: Right concern – wrong concept. *Educational Measurement: Issues and Practice*, 16(2), 9–13.
Ryans, D. G. & Frederiksen, N. (1951). Performance tests of educational achievement. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 455–494). Washington, DC: American Council on Education.
Rulon, P. J. (1946). On the validity of educational tests. *Harvard Educational Review*, 16, 290–296.
Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of Research in Education*, Vol. 19 (pp. 405–450). Washington, DC: American Educational Research Association.
Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16(2), 5–24.
Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research*, 45, 83–117.
Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. Lissitz (Ed.), *The concept of validity* (pp. 19–38). Charlotte, NC: Information Age Publishers.
Sireci, S. G. (2016). On the validity of useless tests. *Assessment in Education: Principles, Policies, and Practice*, 23, 226–235.
Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology*, 15, 201–292.
Suppe, F. (1977). *The structure of scientific theories*. Urbana, IL: University of Illinois Press.
Thorndike, E. L. (1918). Individual differences. *Psychological Bulletin*, 15, 148–159.
Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. Lissitz (Ed.), *The concept of validity* (pp. 65–82). Charlotte, NC: Information Age Publishers.
Zwick, R. (2006). Higher education admission testing. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 647–679), Westport, CT: American Council on Education and Praeger.