Journal of Educational Measurement
Winter 2001, Vol. 38, No. 4, pp. 319–342

Current Concerns in Validity Theory

Michael T. Kane
National Conference of Bar Examiners

We are at the end of the first century of work on models of educational and psychological measurement and into a new millennium. This certainly seems like an appropriate time for looking backward and looking forward in assessment. Furthermore, a new edition of the Standards for Educational and Psychological Testing *(AERA, APA, & NCME, 1999) has been published, and the previous editions of the* Standards *have served as benchmarks in the development of measurement theory.*

This backward glance will be just that, a glance. After a brief historical review focusing mainly on construct validity, the current state of validity theory will be summarized, with an emphasis on the role of arguments in validation. Then how an argument-based approach might be applied will be examined in regards to two issues in validity theory: the distinction between performance-based and theory-based interpretations, and the role of consequences in validation.

First Stage: The Criterion-Based Model of Validity

Much of the early discussion of validity was couched within a realist philosophy of science, in which the variable of interest was assumed to have a definite value for each person, and the goal of measurement was to estimate this variable's value as accurately as possible. Validity was defined in terms of the accuracy of the estimate.

In practice, this view of validation required some criterion measure which was assumed to provide the "real" value of the variable of interest, or at least a better approximation of this "real" value. Given such a criterion, validity could be evaluated in terms of how well the test scores estimated or predicted the criterion scores. The criterion measure was taken as the value of the attribute of interest, and the *test* was considered valid for any criterion for which it provided accurate estimates (Thorndike, 1918).

The chapter on validity in the first edition of *Educational Measurement* (Cureton, 1950) provided a sophisticated summary of conceptions of validity just before the advent of construct validity. Cureton (1950) took the essential question of validity to be "how well a test does the job it is employed to do" (p. 621) and viewed the criterion model as supplying the best answer:

> A more direct method of investigation, which is always to be preferred wherever feasible, 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. (Cureton, 1950, p. 623)

Basically, the validity of the criterion, defined here in terms of "task performances," was taken for granted, and test scores were to be validated against the criterion scores.

This criterion-based model is quite reasonable and useful in many applied contexts, assuming that some suitable "criterion" measure is available. An employer using a test in hiring or placement wants to know how well each applicant will perform on the job, or in the case of placement, in different jobs, and may have some accepted measure of job performance to use as a criterion. The criterion model led to the development of some very sophisticated analyses of the relationship between test scores and criteria and the relative utility of various decision rules that might be used (Cronbach & Gleser, 1965).

Addendum to the First Stage: Content-Based Validity Models

The trouble with the criterion-based model is the need for a well-defined and demonstrably valid criterion measure. In many cases (e.g., high-school graduation tests), suitable criterion measures are not readily available. And if a criterion measure is available (e.g., first semester GPA in validity studies of college admission tests), questions about the validity of the criterion can always arise.

The criterion model does not provide a good basis for validating the criterion. Even if some second criterion can be identified as a basis for validating the initial criterion, we clearly face either infinite regress or circularity in comparing the test to criterion A, and criterion A to criterion B, etc.

One way out of this dilemma is to employ a criterion measure involving some desired performance (or some desired outcome) and interpret the scores in terms of that kind of performance, as in the Cureton quotation above, so that the validity of the criterion can be accepted without much ado. Ebel (1961) talked about some measures being intrinsically valid. For example, skill in playing the piano can be assessed by having several competent judges evaluate individuals as they play several pieces on the piano. In assessing level of skill in particular kinds of performance (e.g., on the piano, in the backstroke, or in penmanship) claims for intrinsic validity may be quite plausible.

For more broadly defined interpretations (e.g., achievement tests in academic content areas), arguments for validity of the test as a measure of achievement over a content area have generally been based on "a review of the test content by subject-matter experts" (Angoff, 1988, p. 22). This kind of judgment-based validity evidence is open to a number of criticisms (Guion, 1977). In particular, it tends to be highly subjective and has a strong confirmatory bias. The judgments about what a test item measures or the content domain covered by a test are usually made during test development or soon after, by persons involved in test development. Not surprisingly, such persons tend to see the test as a reasonable way to measure the attribute of interest.

Messick (1989) described content-validity evidence as providing support for, "the domain relevance and representativeness of the test instrument", but saw it as playing a limited role in validation because it does not provide direct evidence, "in support of inferences to be made from test scores" (p. 17). Nevertheless, a reasonable case can be made for interpreting a direct measure of performance on certain

tasks (e.g., playing the piano) in terms of level of skill in performing that kind of task (Cronbach, 1971). The scores from less direct measures can then be used to estimate or predict these direct measures and can be validated through the criterion model, with the direct measure serving as the criterion. This is a limited but reasonable methodology, and the basic model is still appropriate in many contexts (e.g., in selection and placement testing).

Second Stage: The Construct Model

In the early 1950s, the American Psychological Association Committee on Psychological Tests found it necessary to broaden the then current definition of validity in order to accommodate the interpretations assigned to clinical assessments. A subcommittee of two members, Paul Meehl and Robert Challman, was asked to identify the kinds of evidence needed to justify the "psychological interpretation that was the stock-in-trade of counselors and clinicians" (Cronbach, 1989, p. 148). They introduced the notion and terminology of construct validity, which was incorporated in the 1954 *Technical Recommendations* (American Psychological Association, 1954), and further developed by Cronbach and Meehl (1955).

Naturally enough, Cronbach and Meehl (1955) adopted the hypothetico-deductive (HD) model of theories, which was dominant in the early 1950s, as the framework for their analysis of theoretical constructs. The HD model (Suppe, 1977) treats theories as interpreted axiomatic systems. A set of axioms connecting certain implicitly defined terms (the theoretical constructs) constitutes the core of the theory. The axioms are interpreted by connecting some of their terms to observable variables, through "correspondence rules." (Note that the HD model presupposes the availability of some observable variables.)

Once interpreted, the axioms can be used to make predictions about observable relationships among variables, and these empirical laws are said to be explained by the theory (Hempel, 1965). The nomological network defining the theory consists of the interpreted axiomatic system plus all of the empirical laws derived from it. The theory is validated by checking the empirical laws against data.

The primitive terms or constructs in the axioms are not explicitly defined by any kind of observation. Rather, they are implicitly defined by their role in the theory. It is necessary, of course, to use some observations to estimate the value of any construct, but the construct is not defined by these observations. The validity of the proposed interpretation of scores in terms of the construct is evaluated in terms of how well the scores satisfy the theory. If the observations are consistent with the theory, the validity of the theory and of the measurement procedures used to estimate the constructs defined by the theory are both supported. If the observations are not consistent with the theory, some part of the network would be rejected, but it would generally not be clear whether the fault is in the axioms, the correspondence rules, or in the details of the measurement procedures.

In the *Technical Recommendations* (APA, 1954) and in Cronbach and Meehl (1955), construct validity was presented as an alternate to the criterion and content models, and as being, at least roughly, on a par with them. Cronbach and Meehl said that "*construct validation* is involved 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). The *Technical Recommendations* (1954) and Cronbach and Meehl (1955) both treated construct validity as an addition to the criterion and content models and not as the overriding concern.

Cronbach and Meehl (1955) did go on to say that, "determining what psychological constructs account for test performance is desirable for almost any test" (p. 282). That is, even if the test is initially validated using criterion or content evidence, the development of a deeper understanding of the constructs or processes accounting for test performance requires a consideration of construct validity. So, Cronbach and Meehl (1955) suggested that construct validity was a pervasive concern, but did not present it as a general organizing framework for validity.

The 1966 *Standards* distinguished construct validity from other approaches to validity, particularly criterion validity.

> *Construct validity* is ordinarily studied when the tester wishes to increase his understanding of the psychological qualities being measured by the test.... Construct validity is relevant when the tester accepts no existing measure as a definitive criterion. (APA, AERA, & NCME, 1966, p. 13)

So, ten years after Cronbach and Meehl (1955), the construct model was still presented as an alternative to the criterion model and not as an overriding concern. There was no suggestion that the criterion or content models were to go away or be subsumed under construct validity. Rather construct validity was to focus on the more explanatory, theoretical interpretations.

The 1974 *Standards* (APA, AERA, & NCME, 1974) continued along this track, listing four kinds of validity associated with "four interdependent kinds of inferential interpretation" (p. 26) (predictive and concurrent validities, content validity, and construct validity). The treatment of construct validity in the 1974 *Standards* stuck pretty close to Cronbach and Meehl (1955) in tying construct validity to theoretical constructs:

> A psychological construct is an idea developed or "constructed" as a work of informed, scientific imagination; that is, it is a theoretical idea developed to explain and to organize some aspects of existing knowledge. Terms such as "anxiety," "clerical aptitude," or "reading readiness" refer to such constructs, but the construct is much more than the label; it is a dimension understood or inferred from its network of interrelationships. (APA, AERA, & NCME, 1974, p. 29)

Cronbach (1971) clearly distinguished several approaches to validation, including construct validation:

> 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)

Cronbach goes on to say that, "a description that refers to the person's internal processes (anxiety, insight) invariably requires construct validation" (Cronbach,

1971, p. 451). In essence, then, validity was presented even well into the 1970s as involving several possible approaches.

Between the early 1950s and the mid to late 1970s, the practice developed of using the different models as a sort of toolkit, with each model to be employed as needed in the validation of educational and psychological tests. The criterion model was generally used to validate selection and placement decisions. The content model was used to justify the validity of various achievement tests. And construct validation was to be used for more theory-based, explanatory interpretations. In most cases, more than one model could be pressed into service. For example, a course placement test might be interpreted as a measure of an aptitude construct, but rely heavily on criterion-related validity evidence, with the criterion consisting of an achievement test, which is in turn, justified by content-related evidence. This "toolbox" approach to validation was embedded in the legal system through the Equal Employment Opportunity Commission Guidelines (1979) which were developed by several federal agencies for the implementation of civil rights legislation.

A problem that came to be clearly recognized by the late 1970s was the possibility, even the ease in this context, of being highly opportunistic in the choice of validity evidence (Guion, 1977; Cronbach, 1980a; Messick, 1975, 1981; Tenopyr, 1977). For example, a proposed interpretation stated in theoretical terms might be supported by analyses of test content and/or correlations with various criteria, some of which could be of dubious relevance (correlations of licensure scores with grades in professional school), without ever evaluating the reasonableness of the proposed interpretation (or even stating it clearly).

Development of Construct Validity, 1955–1989

Although construct validity evidence continued to be viewed as one of several types of validity evidence (applicable primarily to theoretical constructs), at least three aspects of the construct-based model gradually emerged as general principles of validation, applicable to all proposed interpretations.

First, Cronbach and Meehl (1955) made it clear that the validation of an interpretation in terms of a theoretical construct would involve an extended effort, including the development of a theory, the development of measurement procedures thought to reflect (directly or indirectly) some of the constructs in the theory, the development of specific hypotheses based on the theory, and the testing of these hypotheses against observations. In the criterion model, the test scores were simply compared to the criterion scores. In the content model, the characteristics of the measurement procedure were evaluated in terms of expert opinion about how the observable variable should be measured. In the construct-validity model, the evaluation of validity always required an extended analysis. As a result, the development of the construct-validity model highlighted the inadequacies of most validation efforts based on a single (often dubious) validity coefficient or simply on expert opinion (Cronbach, 1971).

Second, by focusing on the role of potentially complex theories in defining attributes, Cronbach and Meehl (1955) increased awareness of the need to specify the proposed interpretation before evaluating its validity. They made the point that, "the network defining the construct, and the derivation leading to the predicted

observation, must be reasonably explicit so that validating evidence may be properly interpreted" (p. 300). The variable of interest is not out there to be estimated; the variable of interest has to be defined or explicated. Within the criterion model, it is relatively easy to develop validity evidence based on a preexisting criterion (e.g., a test–criterion correlation) without examining the rationale for the criterion too carefully. In fact, it could be argued that criterion-based validation works best if the criterion can be accepted at face value. To the extent that the criterion requires close examination, the evidence based on it tends to be ambiguous. In marked contrast, the development of construct-related validity evidence requires that the proposed interpretation (the network) be specified in some detail. The emphasis shifts from the validation of the test (as a measure of an existing variable) to the development and validation of a proposed interpretation. It is not the test or the test score that is validated, but a proposed interpretation of the score (Cronbach, 1971).

Third, construct validity's focus on theory testing led to a growing awareness of the need to challenge proposed interpretations and of the importance of considering possible alternate interpretations. Cronbach and Meehl (1955) did not give much direct attention to the evaluation of alternate interpretations, but this notion is implicit in their focus on theory and theory testing, and it was made fully explicit in subsequent work on construct validity (Cronbach, 1971, 1980a, b; Embretson, 1983; Messick, 1989), which gave a lot of attention to the evaluation of competing interpretations. The evaluation of competing interpretations had not been a big issue for the criterion and content models.

The construct-validity model developed three methodological principles (the need for extended analysis in validation, the need for an explicit statement of the proposed interpretation, and the need to consider alternate interpretations) in the context of validating theoretical constructs (APA, 1954; Cronbach & Meehl, 1955). However, after 1955, the three principles were gradually extended to all serious validation efforts and, as a result, transcended the theory-dependent context in which they were introduced. The net result was a broadening of the methodological program initiated by Cronbach and Meehl (1955) into a general methodology for validation.

Construct Validity as the Basis for Unified Validity

By the end of the 1970s, the view initially articulated by Loevinger (1957) that "since predictive, concurrent, and content validities are all essentially ad hoc, construct validity is the whole of validity from a scientific point of view" (p. 636) became widely accepted. The construct-validity model came to be seen, not as one kind of validity evidence, but as a general approach to validity that includes all evidence for validity, including content and criterion evidence, reliability, and the wide range of methods associated with theory testing (Messick, 1975, 1980; Tenopyr, 1977; Guion, 1977; Embretson, 1983; Anastasi, 1986). According to Messick (1988),

> Thus, from the perspective of validity as a unified concept, all educational and psychological measurement should be construct-referenced because construct interpretation undergirds all score-based inferences — not just those related to inter-

> pretive meaningfulness but also the content- and criterion-related inferences specific to applied decisions and actions based on test scores. (p. 35)

As noted earlier, the seeds of this broader conception of construct validity as a general framework for validity were already present in Cronbach and Meehl's (1955) development. Loevinger (1957) made the broader conception explicit. It gradually gained favor in the 1960s and 1970s, and Messick adopted it as a general framework for validity (Messick, 1975, 1988, 1989).

The emphasis on construct validity as a unified framework for validity has been especially useful in emphasizing the pervasive role of assumptions in our interpretations. As Cronbach (1988) has expressed it: "Questions of construct validity become pertinent the moment a finding is put into words" (p. 13). Taking construct validity as the unifying principle for validity puts validation squarely in the long scientific tradition of stating a proposed interpretation (or theory) clearly and subjecting it to empirical and conceptual challenge.

However, the use of construct validity as the framework for a unified model of validation has also had some drawbacks. The hypothetico-deductive model of theories (Suppe, 1977) adopted by Cronbach and Meehl (1955) was concerned mainly with the logical structure of theories and their relationships to experience. Much of the work based on the HD model involved the "logical reconstruction" of existing theories as interpreted axiomatic systems. The proponents of this model explicitly distinguished between the psychology of discovery and the logic of justification, and focused their attention on the logic of justification. According to Feigl (1970), "The rational reconstruction of theories is a highly artificial hindsight operation which has little to do with the work of the creative scientist" (p. 13), and arguably a lot less to do with the work of teachers, policy makers, and others making day-to-day decisions based on test scores.

The basic notion of implicitly defining constructs by their roles in a nomological network assumes that the network is based on a tightly connected set of axioms. Educational research and the social sciences generally have few if any such networks. Cronbach and Meehl (1955) recognized this limitation:

> The idealized picture is one of a tidy set of postulates which jointly entail the desired theorems; since some of the theorems are coordinated to the observation base, the system constitutes an implicit definition of the theoretical primitives and gives them an indirect empirical meaning. In practice, of course, even the most advanced physical sciences only approximate this ideal .... Psychology works with crude, half-explicit formulations. (pp. 293–294)

But they went on to say that the "network still gives the constructs whatever meaning they do have" (p. 294). Cronbach (1988) has pointed out some of the unfortunate consequences of tying construct validity to the hypothetico-deductive model of theories.

Conflict Between the Strong Program and the Weak Program of Construct Validity

The difficulties in applying construct validity to areas in which there is little solid theory (i.e., most of the social sciences) has led to serious ambiguity in the

meaning of construct validity. In particular, Cronbach (1988) distinguished between a strong program and a weak program of construct validity:

> The weak program is sheer exploratory empiricism; any correlation of the test score with another variable is welcomed . . . . The strong program, spelled out in 1955 (Cronbach & Meehl) and restated in 1982, by Meehl and Golden, calls for making one's theoretical ideas as explicit as possible, then devising deliberate challenges. (pp. 12–13)

The strong program is not possible without strong theory, but it is presented as the ideal. The weak program is sufficiently open that any evidence even remotely connected to the test scores is relevant to validity.

The differences between the weak program and the strong program can lead to confusion. It is easy to conclude, using the weak program, that all validity evidence is construct-related evidence and, therefore, that all interpretations are to be validated using "construct validity." The weak program does indeed pull everything under one unified umbrella. In fact, it pulls too much. In the absence of explicit guidelines for identifying the most relevant evidence, the weak program provides essentially no guidance to the validator. On the other hand, it is not so clear that the strong program necessarily includes all kinds of validation efforts. As noted earlier, for two decades the strong form of construct validity was reserved for theory-based, explanatory interpretations (Cronbach & Meehl, 1955; Cronbach, 1971; APA, 1966, 1974), in contrast to descriptive, performance-based interpretations.

In retrospect, the development of two competing versions of construct validity may have been inevitable. The initial formulations of construct validity focused on theoretical constructs implicitly defined in terms of formal theories. The formulation was elegant, but given the dearth of highly developed formal theories in education and the social sciences, the strong program of construct validity was generally not applicable in anything like its pure form. Some progress was made in the development of methods for the implementation of the strong model (Campbell & Fiske, 1959; Cronbach, 1971; Embretson, 1983; Messick, 1989), but presentations of the construct-validity model continued to be relatively abstract. So the definition of construct validity was loosened to make it more applicable, while the label, "construct validity," with its strong associations with formal theory, was retained. As a result, the weak program of construct validity took on much of the abstractness of the strong program, without the support of formal theory to give it teeth, resulting in "sheer exploratory empiricism" (Cronbach, 1988, p. 12).

The implicit adoption of the weak program did not have a positive impact on validation research:

> 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)

The strong program outlined by Cronbach and Meehl (1955) has a narrower focus but it has teeth. One is to lay out theoretical assumptions and conclusions and then subject these to empirical challenges. The approach adopted in the strong program

is essentially that of theory testing in science. The trouble is that this approach has limited utility in the absence of a well-developed theory to test.

Lack of Clear Criteria for the Adequacy of Validation Efforts

The weak program of construct validity is very open ended. It is not clear where to begin or where to stop. Because the weak program invites such an eclectic and possibly unending process, it is not clear that the program does much to discourage an opportunistic strategy based on readily available data rather than more relevant but less accessible evidence. If an essentially infinite number of studies are relevant, where should one start, and how much is enough? If all data are relevant to validity, why not start with the data that is easiest to collect?

The basic principle of construct validity calling for the consideration of alternative interpretations offers one possible source of guidance in designing validity studies and in restraining empirical opportunism, but like many validation guidelines, this principle has been honored more in the breach than in the observance.

> Despite many statements calling for focus on rival hypotheses, most of those who undertake CV have remained confirmationist. Falsification, obviously, is something we prefer to do unto the constructions of others. (Cronbach, 1989, p. 153) [note CV in original refers to construct validity]

As indicated earlier, much validation research is performed by the developers of the assessment instrument, creating a natural confirmationist bias. The weak program of construct validity contains no effective mechanism for controlling such confirmationist tendencies.

Furthermore, construct validity has not provided a unifying influence on an operational level. The 1985 *Standards* (AERA, APA, & NCME, 1985) urged a unified view of validity, but it organized much of its general discussion and specific standards in terms of three kinds of validity evidence (construct, content, and criterion). Messick (1988) criticized the 1985 *Standards* for accepting the idea (in the comment following the first validity standard) that different validation efforts might involve different types of evidence. Messick was concerned that this flexibility in the 1985 *Standards* would encourage reliance on very limited, and perhaps opportunistically chosen, evidence for validity. So, thirty years after Cronbach and Meehl (1955) and almost thirty years after Loevinger's suggestion that all validity is construct validity, the criteria for evaluating validity evidence were still in doubt.

Current Conceptions of Validity

Current definitions of validity reflect the general principles inherent in the construct-validity model, but have dropped the emphasis on formal theories. In his chapter in the most recent edition of *Educational Measurement*, Messick (1989) provides a very general 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. [emphasis in original] (p. 13)

The 1999 *Standards for Educational and Psychological Testing* define validity as the following:

> . . . 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)

Four aspects of this current view are worthy of note. Each has a long history in the theory of validity.

First, validity involves an evaluation of the overall plausibility of a proposed interpretation or use of test scores. It is the interpretation (including inferences and decisions) that is validated, not the test or the test score. The shift from the early, realist models, in which the attribute to be measured was taken as a given to the current emphasis on interpretations is not a recent development (Cureton, 1951; Cronbach & Meehl, 1955; Cronbach, 1971; Messick, 1975), but it has gradually become more explicit and consistent.

Second, consistent with the general principles growing out of construct validity, the current definitions of validity (Messick, 1989; AERA, APA, & NCME, 1999) incorporate the notion that the proposed interpretations will involve an extended analysis of inferences and assumptions and will involve both a rationale for the proposed interpretation and a consideration of possible competing interpretations. The resulting evaluative judgment reflects the *adequacy* and *appropriateness* of the interpretation and the *degree* to which the interpretation is adequately supported by appropriate evidence.

Third, in both Messick's (1989) chapter and the *Standards* (AERA, APA, & NCME, 1999) validation can include the evaluation of the consequences of test uses:

> Tests are commonly administered in the expectation that some benefit will be realized from the intended use of the scores. A few of the many possible benefits are selection of efficacious treatments for therapy, placement of workers in suitable jobs, prevention of unqualified individuals from entering a profession, or improvement of classroom instructional practices. A fundamental purpose of validation is to indicate whether these specific benefits are likely to be realized. (AERA, APA, & NCME, 1999, p. 16)

Those who propose to use a test score in a particular way (e.g., to make a particular kind of decision) are expected to justify the use, and proposed uses are generally justified by showing that the positive consequences outweigh the anticipated negative consequences (e.g., see 1999 *Standards* 1.19, 1.22, 1.25, 1.24, plus comments).

Concerns about consequences are evident in Cureton's (1950) definition of validity in terms of how well a test does what it is designed to do, and in earlier work. It is not a new concern but has been getting more attention lately (Cronbach, 1980a, b; Linn, 1997; Messick, 1975, 1980, 1989; Moss, 1992; Shepard, 1997). However, consensus has not been achieved on what the role of consequences in validation should be, and at least one prominent researcher (Popham, 1997) has suggested that they should not play any role in validity. I will discuss this issue more fully later in this article.

Fourth, validity is an integrated, or unified, evaluation of the interpretation. It is not simply a collection of techniques or tools. The goals of validation, the general approach to validation, and the criteria for judging validation efforts are consistent. The inferences included in the interpretation are to be specified; these inferences and any necessary assumptions are to be supported by evidence; and plausible alternative interpretations are to be examined. The specific components expected in a validation effort may change from one context or application to another, but the general character and structure of what is being done does not change.

Validity as Argument

One way to provide a consistent framework for validation efforts is to structure them in terms of arguments (Cronbach, 1980a, b, 1988; House, 1980). In 1988, Cronbach organized his discussion of validity in terms of evaluative argument:

> Validation of a test or test use *is* evaluation (Guion, 1980; Messick, 1980), so I propose here to extend to all testing the lessons of program evaluation. What House (1977) has called 'the logic of evaluation argument' applies, and I invite you to think of "validity argument" rather than "validation research." (p. 4)

In much of his writing, Cronbach has emphasized the social dimensions and context of validity arguments, in addition to their role in providing structure for the analysis and presentation of validity data (Cronbach, 1980a, b). The 1999 *Standards* suggest that, ". . . validation can be viewed as developing a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed use" (AERA, APA, & NCME, 1999, p. 9).

The *validity argument* provides an overall evaluation of the intended interpretation and uses of test scores (Cronbach, 1988). It aims for a coherent analysis of all of the evidence for and against the proposed interpretation and, to the extent possible, the evidence relevant to plausible alternate interpretations.

In order to evaluate a proposed interpretation of test scores, it is necessary to have a clear and fairly complete statement of the claims included in the interpretation and the goals of any proposed test uses. Validation is difficult at best, but it is essentially impossible if the interpretation to be validated is unclear. The proposed interpretation can be specified in terms of an *interpretive argument* that lays out the network of inferences leading from the test scores to the conclusions to be drawn and any decisions to be based on these conclusions (Kane, 1992, 1994; Shepard, 1993; Crooks, Kane, & Cohen, 1996). The main point of the interpretive argument is to make the assumptions and inferences in the interpretation as clear as possible.

The interpretive argument provides a framework for developing a validity argument. Ideally, we would start with a clear statement of the proposed interpretation in terms of an explicitly stated interpretative argument. Evidence and analysis would then be brought to bear on the inferences and assumptions in the interpretive argument, paying particular attention to the weakest parts of this argument.

A Strategy for Validation Research

The interpretive argument will generally contain a number of inferences and assumptions (as all arguments do), and the studies to be included in the validation

effort are those studies that are most relevant to the inferences and assumptions in the specific interpretive argument under consideration. It is the content of the interpretation that determines the kinds of evidence that are most relevant and, therefore, most important in validation.

An effective strategy for validating the interpretation is easy to outline (but not necessarily easy to implement).

1. State the proposed interpretive argument as clearly and explicitly as possible.
2. Develop a preliminary version of the validity argument by assembling all available evidence relevant to the inferences and assumptions in the interpretive argument. One result of laying out the proposed interpretation in some detail should be the identification of those assumptions that are most problematic (based on critical evaluation of the assumptions, all available evidence, and outside challenges or alternate interpretations).
3. Evaluate (empirically and/or logically) the most problematic assumptions in the interpretive argument. As a result of these evaluations, the interpretive argument may be rejected, or it may be improved by adjusting the interpretation and/or the measurement procedure in order to correct any problems identified.
4. Restate the interpretive argument and the validity argument and repeat Step 3 until all inferences in the interpretive argument are plausible, or the interpretive argument is rejected.

An interpretive argument that survives all reasonable challenges to its assumptions can be provisionally accepted (with the caveat that new challenges may arise in the future).

Each interpretive argument is unique and therefore the associated validity argument will also be unique. Crooks, Kane, and Cohen (1996) have examined many of the inferences commonly found in test-score interpretations. For the sake of simplicity, discussion will be restricted to five basic inferences: evaluation, generalization, extrapolation, explanation, and decision making. Each inference requires a different mix of supporting evidence. For example, if the scores on a test consisting of 20 computational problems are interpreted as a measure of computational skill and used for placement decisions, the interpretation of a student's performance would begin with an evaluation of his or her performance on each question. The resulting score would be generalized beyond the specific performances observed to a universe of possible performances on similar computation problems under similar circumstances. To be useful, the results must usually be extrapolated beyond the testing context to various other contexts (e.g., the classroom, workplace) and to other task formats and performance formats. To the extent that the performances can be explained theoretically, the interpretation is richer and deeper. Finally, the scores can be used to make placement decisions.

The validity argument can make a positive case for the proposed interpretation by providing adequate support for each of the inferences and assumptions in the interpretive argument. The validity argument would also consider any plausible alternative interpretations for the scores and evaluate these alternative interpretations where possible. A fairly easy way to develop alternative interpretations is to

consider changing one or more of the inferences in the interpretive argument. We can challenge the criteria for evaluating performances and suggest different criteria. The existence of large task or rater effects or strong context effects can suggest that generalization is too broad. Alternately, if the universe of generalizations is narrowly defined, extrapolation to other kinds of performance may be limited. And, of course, a competing interpretation can be developed by proposing a different explanation for the observed performances. Finally, critics might claim that the test fails to make appropriate placement decisions for some persons or has serious unintended negative consequences.

A major strength of this argument-based approach to validation is the guidance it provides in allocating research effort and in deciding on the kinds of validity evidence that are needed (Cronbach, 1988). The kinds of validity evidence that are most relevant are those that evaluate the main inferences and assumptions in the interpretive argument, particularly those that are most problematic. The weakest parts of the interpretive argument are to be the focus of the analysis. If some inferences in the argument are found to be inappropriate, the interpretive argument needs to be either revised or abandoned. The structure of the interpretive argument determines the kinds of evidence to collect at each stage of the validation effort and provides a basis for evaluating overall progress.

Issues in Validity Theory

The remainder of this article looks to the future by examining how two issues might be addressed within an argument-based framework for validity. Conceptual approaches like the argument-based framework should be evaluated in terms of the extent to which they help to resolve dilemmas and solve problems, without causing new problems.

Performance-Based, Observable Attributes, and Theoretical Constructs

As noted earlier, the current emphasis on validity as a unified concept arose largely in reaction to the use of the various "kinds" of validity as a sort of toolkit, with only loose criteria for the selection of tools. The unified view emphasized the need for a consistent approach to validation, integrating multiple lines of relevant evidence (Cronbach, 1971; Messick; 1989). However, the suggestion that all validity is construct validity (Loevinger, 1957; Messick, 1988) can also be taken to mean that all interpretations should be validated in the same way, in particular, in terms of theoretical constructs.

This kind of *uniform* approach (as distinct from a *unified*, but flexible approach) has several disadvantages. First, by eliminating the traditional taxonomy of "types" of validity without providing a new structure, the uniform approach can make the choice of research questions for a validation study less clear than it was under the trinitarian model of criterion, content, and construct "validities." The trinitarian model may not have worked very well (Guion, 1980; Cronbach, 1980a; Messick, 1975), but it provided some guidance, and it is still with us, in part because its total elimination would have left a vacuum. Unless we are willing to assume that all validations are to follow the same pattern of inference and evidence, we need some criteria for what to include in each validation. It seems clear that the validation of a

spelling test as a measure of skill in spelling the words in some domain of words need not involve the same level of effort, or the same kinds of evidence, as the validation of a theoretical construct embedded deep in a complex theory. But what is required in each of these two scenarios, and what if anything can be left out?

Second, the elimination of the traditional distinction between theoretical constructs and observable variables makes the empirical evaluation of theories very difficult. If all interpretations are to be treated as constructs defined by a nomological network, validation will always involve the full network. How then can the theory be tested? What can it be tested against? If all variables depend on the theory, any empirical check on the theory must presume the validity of the theory in advance. In order to develop effective empirical checks on the theory, it is necessary to have some variables that can be interpreted without appeal to the theory.

Third, a uniform approach based on the strong program of construct validity can make satisfactory validation especially difficult. The use of the strong program of construct validity is hard even if one has a highly developed theory; it is essentially impossible in the absence of theory. To the extent that the strong program is unattainable, the natural reaction is to slip into the weak program or to ignore the issue of validity altogether.

In contrast to the uniform approach, a unified argument-based approach to validation suggests the need for different kinds of validity arguments to support different kinds of interpretive arguments, involving different patterns of inference. Each interpretive argument will be unique in the sense that it will involve specific inferences and assumptions applied in a specific context. Therefore, the details of the validity argument for each interpretive argument will also be unique. Yet, the general approach, involving the specification of an interpretative argument and the evaluation of its inferences and assumptions, is consistent or unified.

Although every interpretation is unique in some ways, it is possible to distinguish various kinds of interpretations involving certain patterns of inference. One reason for the persistence of the terms, "content validity" and "criterion validity," in spite of repeated attempts to banish them, is the need for some structure and the sense that these terms do reflect (albeit, very loosely) real distinctions among validation problems.

In this section, a distinction is drawn between two kinds of interpretations, which I will refer to as *observable attributes* and *theoretical constructs*. *Observable attributes* are defined in terms of a universe of possible responses or performances, on some range of tasks under some range of conditions (Kane, 1982). These interpretive arguments focus on a limited set of inferences, including the evaluation of specific responses and generalization of the resulting scores to a universe of observations that are of interest (Kane, Crooks, & Cohen, 1996). Cronbach and Meehl (1955) refer to this kind of variable as an "inductive summary" and suggest that such variables can be defined entirely in terms of descriptive dimensions and need involve little or no theory.

The evidence supporting the evaluation of the examinee's performances would involve justifications for scoring rubrics and administration procedures. The evidence for the generalization to the mean over the universe of possible performances defining the observable attribute would involve an estimate of the standard error of

a reliability or generalizability coefficient (Brennan, 1992; Cronbach, Gleser, Nanda, & Rajaratnam, 1972), or of an error/tolerance ratio (Kane, 1996). Explanatory theory may play a background role in these analyses, but it need not be explicitly considered in validating the proposed interpretation as an observable attribute.

Scores on performance assessments can generally be interpreted as observable attributes (Moss, 1992; Linn, Baker, & Dunbar, 1991; Kane, Crooks, & Cohen, 1996). As Cureton (1950) indicated the following a half century ago:

> If we want to find out how well a person can perform a task, we can put him to work at that task, and observe how well he does it and the quality and quantity of the product he turns out. (p. 622)

The observable attribute can be defined in terms of the average level of performance over some universe of possible tasks and, therefore, can be defined without any explicit appeal to theory. The attribute is observable in the sense that its interpretation is specified in terms of a universe of possible observations.

Theoretical constructs are implicitly defined by theories (Cronbach & Meehl, 1955). They are not explicitly defined in terms of any observations, but rather, by their role in the theory from which they derive most of their meaning. The empirical index used to estimate the value of the construct will be defined in terms of observations, but the index does not exhaust the meaning of the theoretical construct. The index actually employed may be one of many possible indices. It is likely to be designed to be consistent with the assumptions in the theory and to yield the results predicted by the theory, but the definition of the observable attribute used as an index does not depend on the theory. Galton's attempt to use reaction times as indices of intelligence failed, but reaction times are still interpretable. The usefulness of the index for a theoretical construct is linked to the usefulness of the theory and its interpretation is determined by the content of the theory.

An interpretation in terms of a theoretical construct generally involves a number of inferences. The observed performances used as indicators of the construct must be evaluated in order to generate an observed score. Usually, this observed score is expected to generalize over various potential sources of irrelevant variance (e.g., raters, occasions, and specific tasks). These first two steps are likely to follow the pattern for any observable attribute; and, focusing just on this part of the interpretive argument, the indicator can be viewed as an observable attribute. In addition, the theory defining the construct will generate empirical hypotheses involving the construct, and any observed relationships among the indices for a set of constructs must be consistent with the hypotheses derived from the theory. This last step may suggest the need for a large number of studies of various kinds.

The observable attribute serving as an index may or may not be of intrinsic interest as an observable attribute, independent of its role as an index. The skills assessed by a math test, used as one indicator of general academic aptitude, could be of great intrinsic educational interest, while the specific skills assessed by another indicator, say a block-sorting task, might be of little interest beyond their potential usefulness in estimating the value of the aptitude for each individual.

The distinction between an observable attribute and a theoretical attribute is in their interpretations and is context dependent. The interpretation of the observed score for an observable attribute involves the evaluation of the observed performance and generalization to some target universe of possible performances. The interpretation of the index for a theoretical attribute goes beyond this kind of inductive summary and seeks to draw conclusions about some construct defined by a theory. The construct interpretation provides an explanation, perhaps a causal explanation (Cook & Campbell, 1979) of observed relationships. The observable variables serving as indicators of the theoretical constructs can be used to check these hypothesized relationships. The distinction here is not among different kinds of validity or even different types of validity evidence, but among different types of interpretations.

The distinction between observable attributes and theoretical constructs is context dependent. A variable can be considered an observable attribute in a particular context as long as it does not rely on theoretical assumptions that are under investigation in that context (Grandy, 1992), and the assumptions that can be taken for granted depend on the context (Cronbach, 1988).

It has long been recognized that the interpretations of observations always rely on prior assumptions (Hanson, 1958; Kuhn, 1962), and therefore the interpretation of an observable attribute always relies on some theory. The terms used to describe the performances are drawn from some language, and languages always incorporate substantive assumptions about how the world functions. In addition, our interest in this particular kind of performance may be based on current theories of learning or performance in this area. We put certain tasks (e.g., arithmetic items) together in a content domain because we think that these tasks require the same or at least overlapping skills or component performances. However, to function as an observable attribute in a particular context, the interpretation of the attribute should not depend on theories under investigation in that context. All of the theories employed in interpreting the observable attribute should be unproblematic in that context.

In addition to suggesting the general content of the observable attributes, theoretical assumptions can also serve as the basis for defining the boundaries of subdomains. For example, rather than specify the task domain for an end-of-unit test on subtraction in terms of performance on subtraction problems, we might choose to define one performance variable for subtraction problems that require "borrowing" and another for subtraction problems without "borrowing." This would make sense if "borrowing" is seen as an important component skill, with high diagnostic value.

Nevertheless, once it is defined, an observable attribute can be interpreted without employing the theory currently under investigation. A universe of tasks can be specified without appeal to cognitive theories of performance for these tasks. To distinguish between the "borrow" and "non-borrow" tasks, it is necessary to know something about arithmetic, but a cognitive model of performance on subtraction problems is not needed.

Once defined, an observable attribute has a relatively simple interpretive argument, with a clear validation strategy. The strategy may not be easy to implement (developing and validating performance tests may be very difficult), particularly

because it may be difficult to supply adequate support for various assumptions (e.g., it may be difficult to establish the generalizability of observed scores because of task specificity), but the strategy is well defined. It is possible to validate the interpretation fairly well in a finite (even a small) number of steps. And because it does not make use of any disputed theoretical assumptions, the resulting validity argument may be convincing to people with different theories about the performance being measured.

Such observable attributes are important for at least three reasons. First, they define goals for theory: They can help to specify the phenomena that theory is called upon to explain. The observable variables can be defined before theory gets highly developed, and arguably some of them have to be defined before the theory gets fully developed. How can we develop a theory of performance in "X" without having some fairly clear idea of what "X" is, and how can we decide whether the theory adequately explains "X" if we cannot measure "X" with some confidence, independent of the theory?

Second, two individuals who hold different theories about a particular kind of performance can often agree on a performance-based interpretation for an observable attribute for which both theories make predictions. One theory might suggest that subtraction items requiring borrowing would be especially difficult for certain students (e.g., those with mild dyslexia) while the other theory might expect to see no differences in performance among the specified groups. To the extent that the adherents of both theories can agree on the definition of observable variables for subtraction with borrowing and for subtraction without borrowing (and on the criteria for categorizing students), they can subject their dispute to empirical tests. Without observable attributes, "critical" experiments would not be possible. Messick (1998) explicitly recognizes this limitation:

> In this synthesis of realism and constructivism, theories can no longer be directly tested against facts because value-neutral data are problematic in the post-modern world. (p. 36)

But Messick (1998) suggests that conjectures can be tested against observations within a specific framework or inquiry system. That is, within a particular context, it is possible to define attributes that are acceptable as observable attributes.

Third, the observable attribute may be of practical importance in a particular context, independent of theory. It may be of importance to an employer to know whether sales clerks can perform the mathematical tasks required of them on the job, independent of how they acquired the skills, or how they perform the tasks.

The distinction being employed here has a long tradition in science, going back at least to Galileo. Low-level inductive summaries, or observable variables, are used to describe observed phenomena and to develop empirical laws. Theoretical constructs and the theories in which they are defined constitute hypotheses or conjectures intended to explain the observed phenomena (Popper, 1965; Lakatos, 1970). The theoretical constructs and the indices used to measure them are validated by examining how well the theory as a whole accounts for the observable phenomena.

Interpretations that do not go much beyond the observations on which they are based (e.g., inferring how well a student can solve geometric analogy items

basedon his or her performance on a sample of 20 geometric analogy items) can be supported by modest validity arguments. More expansive and ambitious interpretations (e.g., from observed scores on geometric analogy items to conclusions about science aptitude or IQ) require more extensive validity arguments. I suggest that we will make more rapid progress in developing and validating our measurement procedures and our theories if we recognize these differences.

The Role of Consequences in Validation

In a recent debate, Popham (1997) argued for a limited, technical definition of validity, involving primarily the descriptive interpretation of scores. He preferred to treat validation as an objective, scientific concern, separate from disputes about the consequences of test use. He acknowledged that consequences were important, but preferred to treat them as a separate issue. Linn (1997) and Shepard (1997) favored a broader conception of validity, which would include the consequences of test use, as well as the descriptive interpretation of test scores. Mehrens (1997) came down closer to Popham's view than to that of Linn and Shephard. More recently, an entire issue of a journal was devoted to an extended discussion of the role of consequences in validity (Yen, 1998).

As Shepard (1997) notes, consequences have always been a part of our conception of validity. Formulation of the basic question of validity in terms of whether a test achieves the purpose for which it was created (Cureton, 1950) immediately raises questions of intended consequences and less directly of unintended consequences (Moss, 1992; Shepard, 1997). Nevertheless, for a long period, consequences were not a major focus in discussions of validity. An emphasis on content and criterion-related questions, as well as the strong program of construct validity, can push consequences to the background, if not off the stage altogether.

It seems clear that some consideration of consequences is essential in any thorough evaluation of the legitimacy of test use (Cronbach, 1988). A highly accurate diagnostic procedure for an untreatable disease would probably not see much use in the clinic, especially if it had serious side effects. And an argument that the diagnostic procedure was perfectly accurate would not save a physician who used it from malpractice suits. The procedure might be employed in research studies, where the potential long-term benefits (identification of promising treatments) could be seen as outweighing any negative short-term effects, but for clinical applications of measurement procedures, as for any clinical applications, the bottom line involves consequences. In real-world applications, we want the desirable consequences of using a measurement procedure to outweigh the negative consequences of such use (Cronbach & Gleser, 1965). Therefore, if validity is to be "the most fundamental consideration in developing and evaluating tests" (AERA, APA, & NCME, 1999, p. 9), it needs to address consequences.

Although the evaluation of consequences seems to be an essential component in the validation of test use, these consequences can be far reaching and hard to determine; and it seems unreasonable and counterproductive to hold a developer or a test user responsible for every possible consequence of test use (Reckase, 1998). So, the basic question is, who is to be responsible for what consequences of test use (Linn, 1998). No general answer to this question is suggested here. The goal in this

section is to suggest how an argument-based approach to validity might help to define the basic issues more clearly.

In discussing the role of consequences in validation, it would probably be useful to separate the interpretive argument into two parts. The *descriptive part* of the argument involves a network of inferences leading from scores to descriptive statements about individuals, and the *prescriptive part* involves the making of decisions based on the descriptive statements. For example, the use of a reading comprehensive test to place students in reading groups involves conclusions about each student's level of reading skill, and then a decision about placement, which may involve additional information or constraints (e.g., group sizes). Messick (1975) made this distinction over a quarter of a century ago:

> 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. (p. 962)

Although they have differed somewhat in emphasis, both Cronbach (1971, 1980b) and Messick (1975, 1980, 1989) have explicitly included both interpretive accuracy and consequences under the heading of validity. Moss (1992) provides a good summary of the literature on this dual focus in validation.

Under the argument-based model, all of the inferences in an interpretive argument leading to a decision would have to be sound for the overall decision to be sound. It is certainly possible to conceive of an accurate measure of reading skills being used badly. It is also easy to conceive of a well-designed decision process that fails because of an inadequate test, one that does not "support the interpretation . . . entailed by proposed uses of tests" (AERA, APA, & NCME, 1999, p. 9).

Given the differences between the descriptive and prescriptive parts of the argument, it might be useful in many cases to evaluate the two parts of the interpretive argument separately. In particular, in cases where an assessment (e.g., a reading test) can be used to make many different kinds of decisions, including for example, admissions decisions, placement decisions, diagnostic decisions, and grading or graduation decisions, it makes sense to separate the descriptive part of the interpretive argument (e.g., level of reading comprehension) from the decision to be made.

The work of validating the interpretation in terms of reading skills could be done by the test developer and would not have to be repeated for each of the decision contexts in which the test might be used. The validation studies for the descriptive part of the argument could be done once and then incorporated, perhaps with some modification, into the interpretive argument for each decision procedure. Test developers seem to be likely candidates to validate the descriptive interpretation of published tests because they generally have the needed resources and because some of these descriptive inferences must in any case be examined as part of the test-development process (e.g., evaluation of scoring keys or rubrics, the conduct of G studies to estimate generalizability).

The two likely candidates to conduct the analysis of consequences of test use are the user and the test publisher/developer. In some cases, the test developer and user are identical and this question is moot. Assuming that they are different, an argument can be made for concluding that the decision makers (i.e., the test users) have the final responsibility for their decisions (the buck stops on their desks), and they are usually in the best position to evaluate the likely consequences in their context, of the decisions being made (Cronbach, 1980a; Taleporos, 1998). They presumably know how they are using the tests, the population being tested, and the intended outcomes/consequences. If the user does not know why and how the test is being used, he or she should probably not be using it. The user is also in the best position to spot any unintended consequences that occur.

An exception to this suggestion might occur if the test developer designs and markets a test for a particular use. In such cases, it would seem reasonable to consider the test developer responsible for providing evidence that supports the proposed use (Shepard, 1997). If the test developer makes a claim explicity or implicity (i.e., by labeling a test as a "placement" or "readiness" test) that a test can be used in some way, it seems incumbent on the developer to back this claim with a validated interpretive argument supporting the use. For example, the developer of a placement testing program has traditionally been expected to report data on how the use of the test scores for placement affected the achievement of students who were placed in different courses.

It also seems reasonable to expect commercial test publishers to anticipate the common uses of certain kinds of tests (Green, 1998; Moss, 1998), and the potential consequences of such use, even if these uses are not explicitly advocated by the publisher. By definition, unanticipated consequences cannot be evaluated in advance (Reckase, 1998; Green, 1998), but they can be monitored after the fact. Much of the initial responsibility for detecting unanticipated consequences necessarily falls on the test user, but publishers can monitor how their tests are being used and the consequences of such use (Green, 1998; Linn, 1998; Moss, 1998).

Note however that there is likely to be some interaction between the descriptive and prescriptive parts of the argument, and in some cases, this interaction may have a major impact on the effectiveness of the testing program. For example, in a high-stakes testing environment, test preparation methods may lead to changes in the meaning of test scores. In particular, the scores may become less predictive of a broad range of achievements to the extent that practice on test questions replaces more general instruction (Heubert & Hauser, 1999).

The evaluation of consequences is likely to be a contentious issue for a long time, and no easy solutions are available. Each application of a measurement procedure will have to be evaluated on its own merits. Many different kinds of evidence may be relevant to the evaluation of the consequences of an assessment system (Lane, Parke, & Stone, 1998), and many individuals, groups, and organizations may be involved (Linn, 1998). But in clarifying the issues involved in assigning responsibility for the overall validation effort, it will be useful to distinguish between interpretive arguments that lead only to descriptions and interpretive arguments that advocate certain actions based on test scores, and to recognize the differences in the kinds of evidence needed to validate these different interpretive

arguments. The validation of decision procedures has always depended on the evaluation of the consequences of the decisions.

Conclusion

Validity is concerned with the clarification and jusitification of the intended interpretations and uses of observed scores. It is notoriously difficult to pin down the interpretation (or meaning) of an observation (hence the popularity of detective novels). It is even more difficult to reach consensus on the appropriate uses of test scores in applied contexts. As a result, it has not been easy to formulate a general methodology for validation.

But progress has been made. In particular, we have moved from relatively limited criterion-related models to quite sophisticated construct models. I see the introduction of a well articulated version of construct validity by Cronbach and Meehl (1955) as the watershed event in the development of validity theory. Their formulation of construct validity emphasizsed theoretical constructs, but the general principles introduced in the 1955 paper and subsequently developed by Cronbach, Messick, Guion, Shepard, Linn, Moss, and others, (i.e., that validation requires an extended analysis of evidence, based on an explicit statement of the proposed interpretation, and involving the consideration of competing interpretations) are applicable to all validity arguments.

These principles fit naturally into an argument-based approach to validation (Cronbach, 1988; Kane, 1992; Shepard, 1993). The proposed interpretations and uses of observed scores can be specified in some detail in the form of an interpretive argument. The interpretive argument involves a network of inferences and assumptions leading from the observed scores to the conclusions and decisions based on the observed scores, and provides an explicit and fairly detailed statement of the proposed interpretation. It specifies the interpretation to be evaluated. The validity argument evaluates the plausibility of the proposed interpretation by critically examining the inferences and assumptions in the interpretive argument. It evaluates the proposed interpretation.

The validity argument will typically involve different kinds of evidence relevant to the different parts of the interpretive argument; it is likely to be most effective in suggesting improvements in the measurement procedure and its interpretive argument to the extent that it identifies the weak points in the interpretive argument. In many cases, it may be possible to strengthen a questionable interpretation by improving the measurement procedures or by revising the interpretation. In some cases, it may be necessary to reject a proposed interpretation as untenable. A proposed interpretation is most effectively evaluated by challenging its most questionable assumptions, thereby pitting it against the most plausible alternate interpretations of the observed scores.

In order to be effective in improving the accuracy and effectiveness of measurement procedures, we need a technology as well as a theory of validity. That is, we need a well-defined set of procedures for identifying the questions that need to be addressed in each case and for answering these questions. An argument-based approach to validation provides a framework for such a technology. The argument-based approach suggests that the proposed interpretation be specified in terms of a

network of inferences and assumptions, that these inferences and assumptions be evaluated using all available evidence, and that plausible alternate interpretations be considered. By focusing on the inferences and assumptions in the specific interpretive argument under consideration, the argument-based approach provides detailed guidance in conducting an effective validation.

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 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, & 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, & National Council on Measurement in Education. (1974). *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. 9–13). Hillsdale, NJ: Lawrence Erlbaum.

Brennan, R. L. (1992). *Elements of generalizability theory, Revised edition*. Iowa City, IA: American College Testing.

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

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

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

Cronbach, L. J. (1988). Five perspectives on 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. (1965). *Psychological tests and personnel decisions*. 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 of scores and profiles*. New York: Wiley.

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

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

Cureton, E. E. (1950). 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, Civil Service Commission, Department of Labor, & Department of Justice. (1979). Adoption by four agencies of Uniform Guidelines on Employee Selection Procedures. *Federal Register, 43*, 38290–38315.

Feigl, H. (1970). The "orthodox" view of theories: Remarks in defense as well as critique. In M. Radner & S. Winokur (Eds.), *Analyses of theories and methods of physics and psychology. Vol. 4, Minnesota studies in the philosophy of science*. Minneapolis, MN: University of Minnesota Press.

Grandy, R. E. (1992). Theory of theories: A view from cognitive science. In J. Earman (Ed.), *Inference, explanation, and other frustrations: Essay in the philosophy of science* (pp. 216–233). Berkeley, CA: University of California Press.

Green, D. R. (1998). Consequential aspects of the validity of achievement tests: A publisher's point of view. *Educational Measurement: Issues and Practice, 17*(2), 16–19, 34.

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

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

Hanson, N. R. (1958). *Patterns of discovery*. Cambridge, UK: Cambridge University Press.

Hempel, C. G. (1965). *Aspects of scientific explanation and other essays in the philosophy of science*. Glencoe, IL: Free Press.

Heubert, J. P., & Hauser, M. H. (1999). *High stakes: Testing for tracking, promotion, and graduation*. National Academy Press, DC.

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

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement, 6*, 125–160.

Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin, 112*, 527–535.

Kane, M. T. (1994). Validating interpretive arguments for licensure and certification examinations. *Evaluation and the Health Professions, 17*, 133–159.

Kane, M. T. (1996). The precision of measurements. *Applied Measurement in Education, 9*(4), 355–379

Kane, M. T., Crooks, T. J., & Cohen, A. S. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5–17.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.

Lakatos, I. (1970). Falsification and the methodology of scientific research programs. In I. Lakatos and A. Musgrave (Eds.), *Criticism and the growth of knowledge*. London: Cambridge University Press.

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.

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice, 16*(2), 14–16.

Linn, R. L. (1998). Partitioning responsibility for the evaluation of the consequences of assessment programs. *Educational Measurement: Issues and Practice, 17*(2), 28–30.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex performance assessment: Expectations and validation criteria. *Educational Researcher, 20*, 15–21.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, Monograph Supplement, 3*, 635–694.

Meehl, P. E., & Golden, R. R. (1982). Taxonomic methods. In P. Kendall and J. Butcher (Eds.), *Handbook of Research Methods in Clinical Psychology* (pp. 127–182). New York: Wiley.

Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice, 16*(2), 16–18.

Messick, S. (1975). The standard program: 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. (1981). Evidence and ethics in the evaluation of tests. *Educational Researcher, 10*, 9–20.

Messick, S. (1988). The once and future issues of validity. Assessing the meaning and consequences of measurement. In H. Wainer and H. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale, NJ: Lawrence Erlbaum.

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

Messick, S. (1998). Test validity: A matter of consequences. *Social Indicators Research, 45*, 35–44.

Moss, P. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*, 229–258.

Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice, 17*(2), 6–12.

Popham, W. J. (1997). Consequential validity: Right concern—wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9–13.

Popper, K. R. (1965). *Conjecture and refutation: The growth of scientific knowledge.* New York: Harper & Row.

Reckase, M. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice, 17*(2), 13–16.

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of Research in Education, 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), 5–8, 13, 24.

Suppe, P. (1977). *The structure of scientific theories.* Urbana, IL: University of Illinois Press.

Taleporos, E. (1998). Consequential validity: A practitioner's perspective. *Educational Measurement: Issues and Practice, 17*(2), 20–23, 34.

Tenopyr, M. L. (1977). Content-construct confusion. *Personnel Psychology, 30*, 47–54.

Yen, W. M. (1998). Investigating the consequential aspects of validity: Who is responsible and what should they do? *Educational Measurement: Issues and Practice, 17*(2), 5.

Author

MICHAEL T. KANE is Director of Research in the National Conference of Bar Examiners, 402 West Wilson Street, Madison WI 53703; mkane@ncbex.org. His research interests include psychometric theory, particularly validity, generalizability theory, and standard setting.