DOCUMENT RESUME

ED 446 094 TM 031 783

AUTHOR Kane, Michael
TITLE Current Concerns in Validity Theory.
PUB DATE 2000-04-00
NOTE 32p.; Paper presented at the Annual Meeting of the American Educational Research Association (81st, New Orleans, LA, April 24-28, 2000).
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Scores; *Test Interpretation; Test Results; *Theories; *Validity

ABSTRACT

Validity is concerned with the clarification and justification of the intended interpretations and uses of observed scores. It has not been easy to formulate a general methodology set of principles for validation, but progress has been made, especially as the field has moved from relatively limited criterion-related models to sophisticated construct models. The watershed event in the development of validity theory has been the development of a well-articulated version of construct validity by L. Cronbach and P. Meehl (1955). The general principles they, and others, have articulated, 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 into an argument-based approach to validation. The validity argument evaluated the plausibility of the proposed interpretation by examining the inferences and assumptions in the interpretive argument critically. The validity argument will typically involve different kinds of evidence and is most likely to be effective in improving the measurement procedure and its interpretive argument to the extent that it identifies the weak points in the interpretive argument. A proposed interpretation is most effectively evaluated by challenging its most questionable assumptions and thereby pitting it against the most plausible alternate interpretations of the observed scores. (Contains 44 references.) (SLD)

Current Concerns in Validity Theory

Michael Kane

Presented at the Annual Meeting of the National Council on Measurement in Education, New Orlean, 2000.

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

My backward glance will be just that, a glance. After a brief historical review focusing mainly on the development of the construct model, I will summarize the current state of validity theory, with an emphasis on the role of arguments in validation. I will then examine how an argument-based approach might be applied to two issues in validity theory: the distinction between performance-based and theory-based interpretations, and the role of consequences in validation.

The First Stage: 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 "true" value as accurately as possible. The validity of the measurement would simply be the accuracy of this estimate of the true value.

In practice, this view of validation required some criterion measure which was assumed to provide the true value of the variable of interest, or at least a very close approximation of this true value. Given a criterion, validity could be evaluated in terms of how well the test scores estimate or predict the criterion scores.

The chapter on validity in the first edition of Education Measurement (Cureton, 1950) provided a sophisticated summary of conceptions of validity just before the advent of construct validity. Cureton (1950) takes the essential question of validity to be, "how well a test does the job it is employed to do." (p. 621), and makes the following suggestion for validation:

> 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 validated against the criterion scores.

This criterion-based model could be quite reasonable and useful in many applied contexts, assuming that some suitable "criterion" measure were 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).

Note that under this model, the attribute, represented by the criterion, was assumed to exist *a priori*, and the question of validity could be stated in terms of how well the test estimated this criterion. The criterion measure was taken as the "real" or "true" value of the attribute of interest, and the test was to be validated against the criterion.

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), good criterion measures are not readily available. And where a reasonable criterion is available, questions about the validity of the criterion inevitably 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 quote 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 the 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 liable 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 has argued that content validity evidence is not validity evidence because it does not involve test scores or the performances on which such scores are based (Messick, 1989). He described content considerations as being relevant to validity, but he also tended to downplay their importance.

Nevertheless, a reasonable case can be made for interpreting a direct measure of performance on certain tasks (e.g., the piano) in terms of level of skill in performing that kind of task. And the use of scores on less direct measures to estimate or predict these direct measures can be validated through the criterion model, with the direct measure serving as the criterion. This is a very limited but reasonable methodology, and the basic model is still appropriate in many contexts (e.g., some employment tests). I shall expand on this issue in a later section dealing with the distinction between observable attributes and theoretical constructs.

Second Stage: The Construct Model

In the early 1950s, the APA Committee on Psychological Tests, found it necessary to broaden the existing definition of validity in order to deal with clinical assessment. 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 in Cronbach and Meehl (1955).

Naturally enough, Cronbach and Meehl (1955) adopted the positivist philosophy of science that was dominant in the early 1950s as the framework for their analysis of theoretical constructs. This view, which Suppe has (1977) called the "received view," treats theories as interpreted axiomatic systems. A set of axioms connecting certain primitive, undefined terms constitutes the core of a theory. The axioms are interpreted by connecting some of their terms to observable variables.

Once interpreted, the axioms can be used to make predictions about observable relationships among variables, or empirical laws, and these laws are said to be explained by the theory (Hempel, 1965). The nomological network defining the theory consists of the interpreted calculus plus all of the empirical laws derived from it. The theory is tested, or validated, as a whole 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 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 checked by seeing if the scores satisfy the theory. If they do, the validity of the theory and the validity of the measurement procedures used to estimate the constructs defined by the theory are both supported. If the observed relationships among scores were 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 or in the measurement procedures.

In the Technical Recommendations (American Psychological Association, 1954) and in Cronbach and Meehl's (1955) exegesis, construct validity was presented as an alternate to criterion and content methods, and as being 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, p. 282) did go on to say that, "determining what psychological constructs account for test performance is desirable for almost any test." 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 leads to considerations 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 went to some pains to distinguish 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, 1966, p. 13).

So, ten years after Cronbach and Meehl (1955), construct validity was still presented as an alternative to criterion validity 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, 1974, p. 26) continued along this track, listing four kinds of validity associated with "four interdependent kinds of inferential interpretation" (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, 1974, p. 29)

Cronbach (1971, p. 451) clearly distinguished several aspects of validation, including construct validity, and suggested that; "A description that refers to the person's internal processes (anxiety, insight) invariably requires construct validation." In essence, then, validity was presented, even well into the 1970s as involving several parallel approaches to

validation, and it was assumed by publishers and users alike that tests could be validated by any one or more of the three general approaches.

Between the early 1950s and the mid to late 1970s, the practice developed of using the different models as a sort of tool kit to be employed as needed in the validation of educational and psychological tests. The criterion model was used to validate the use of tests for selection and placement decisions. Content validity 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 based on a aptitude construct, but rely heavily on criterion-related evidence, with an achievement test justified by content-related evidence used as the test criterion.

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, 1980; Messick, 1975, 1981). 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 might 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), some aspects of construct validity gradually emerged as general principles of validation applicable to all proposed interpretations. I mention three principles that were proposed by Cronbach and Meehl (1955) for the validation of theoretical constructs, but were then applied to validation in general.

First, Cronbach and Meehl (1955) promoted the notion that validation would involve an extended argument rather than a single coefficient or a single judgment. They 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 requires an extended analysis. The variable of interest is not our there to be estimated; the variable of interest has to be defined or explicated. As a result, the availability of the construct validity model highlighted the inadequacies of most validation efforts based on a single (often dubious) validity coefficients or simply on expert opinion.

Second, by focusing on the role of potentially complex theories in defining attributes, construct validity increased awareness of the need to specify an interpretation before conducting a validation study. Cronbach and Meehl (1955) 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) Within the criterion model, it is relatively easy to develop validity evidence (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 criteria are 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 construct-validity model requires that the proposed interpretation (the network) be specified in some detail.

Third, construct validity's focus on theory testing led to a growing awareness 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, 1980a,b), which gave a lot of attention to the evaluation of competing interpretations. The evaluation of competing interpretation had not been a big issue for the criterion and content models.

These three methodological principles (the need for extended analysis in validation, the need for an explicit statement of the proposed interpretations, and the need to consider alternate interpretations) were introduced in the context of validating theoretical constructs (APA, 1954; Cronbach & Meehl, 1955). However, after 1955 these principles were gradually extended to all serious validation efforts and ultimately transcended the theory-dependent context in which they were introduced. The net result was a broadening of the

methodological concerns in Cronbach and Meehl (1955) into a general methodology for validity.

Construct Validity as the Basis for Unified Validity

By the end of the 1970s, the view initially articulated by Loevinger (1957, p. 636) 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", 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, and reliability, as well as the wide range of methods associated with theory testing (Messick, 1975, 1980; Tenopyr, 1977; Guion, 1977). According to Messick (1988, p. 35):

> 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 interpretive meaningfulness but also the content- and criterion-related inferences specific to applied decisions and actions based on test scores. (Messick, 1988, 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. Loevenger (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, p. 13) has put it: "Questions of construct validity become pertinent the moment a finding is put into words." 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.

Nevertheless, the use of construct validity as the framework for a unified model of validation has also had some drawbacks. The received view 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 received

view involved "logical reconstruction" of existing theories as interpreted axiomatic systems. The proponents of the received view 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 more 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. (p. 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. I will focus on two problems growing out of this linkage.

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

The difficulties in applying construct validity to areas in which little solid theory exists (i.e., most of the social sciences) has led to serious ambiguity in the meaning of construct validity. In particular, Cronbach (1988, pp. 12-13) 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 and Meehl) and restated in 1982, by Meehl and Golden, calls for making one's theoretical ideas as explicit as possible, then devising deliberate challenges.

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 and Meehl, 1955; 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. 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)

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, 1980)

The strong program has a narrower focus but it has teeth. One is to lay out theoretical assumptions and conclusions and then subject these to empirical checks. The model is essentially that of theory testing in the sciences.

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 when to stop. Because construct validity (especially the weak program) is such an eclectic and possibly unending process, it is not clear that it does much to discourage an opportunistic strategy that focuses 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 that which is easiest to collect?

The basic principle of construct validity calling for consideration of alternative interpretations offers one possible source of guidance in designing validity studies, 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)

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 effect mechanism for controlling such confirmationist bias.

Furthermore, construct validity has not provided a unifying influence on an operational level. The 1985 Standards 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 provides a very general definition of validity:

> Validity is an integrated evaluative judgement 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](Messick, 1989, p.13)

The new Standards define 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)

There are several aspects of this current view that are worthy of note. First, validity is not a property of the test or of the test scores, but rather an evaluation of the overall plausibility of a proposed interpretation or use of test scores. It is the interpretation that is validated, not the test itself. Note that this presumption is very different from the realism implicit in the early emphasis on criterion validity. The variable is not out there to be found. Interpretations are human creations. Those who propose the interpretation are expected to justify it.

Second, consistent with the general principles growing out of construct validity, these definitions 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 (1989) and the Standards (AERA, APA, NCME, 1999) validation includes the evaluation of the consequences of test uses. 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 this use, and these uses are generally justified by showing that the positive consequences of the proposed use outweigh the anticipated negative consequences. 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 discussion lately. But 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. I will discuss this issue more fully later in this paper.

Fourth, validity is an integrated, or unified, evaluation of the interpretation. It is not simply a collection of techniques to be used as a toolbox. 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 of 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 five perspective on validity in terms of evaluative argument:

> Validation of a test or test use *is* evaluation (Guion, 1980: Messick, 1980), so I propose here to extent 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."

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 validity argument, provides an overall evaluation of the intended interpretation and uses of test scores (Cronbach, 1988). It aims for a persuasive presentation 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 test use. Validation is difficult at best, but it is essentially impossible if the proposed interpretation is left unspecified. 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 links in 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). First, state the proposed interpretive argument as clearly and explicitly as possible. Second, 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). Third, evaluate (empirically and/or logically) the most problematic assumptions. As a result of these evaluations, the interpretive argument may be rejected, or it may be improved by adjusting the interpretive argument and/or the measurement procedure in order to correct any problems identified. Fourth, 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, I will mention five basic inferences: evaluation, generalization, extrapolation, explanation, and decision, each of which requires a different mix of supporting evidence. For example, if the scores on a test consisting of 20 computational problems is 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 their performance on each question. The overall evaluation 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 has been too broad.

Alternatively, if the universe of generalizations is narrowly defined, extrapolation to other kinds of performance may be limited. And, of course, an alternate 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 reasons or has serious unintended negative consequences.

A major strength of this argument-based approach 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 support the main inferences and assumptions in the interpretive argument, particularly those that are most problematic. The weakest link in a chain of inference is 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 paper 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. However, there has also been some tendency to suggest that a unified validity requires that all attributes be validated in the same way, in particular as theoretical constructs.

This kind of uniform approach (as distinct from a unified, but flexible approach) has several disadvantages. First, by eliminating the traditional structure in terms 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 traditional

approach. Granted that the traditional triumvirate of criterion, content, and construct "validities," did not work well, its elimination left a vacuum. Unless, we are willing to assume that all validations are to follow the same pattern, 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?

Second, the elimination of the traditional distinction between theoretical constructs and observable variables makes it very difficult to test theories. If all attributes are implicitly defined by the theory, how 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.

Third, the identification of unified validity with the strong program of construct validity has made satisfactory validation especially difficult. The use of the strong program of construct validity is hard even if one has a respectable theory; it is essentially impossible in the absence of theory. Not only are we to validate all attributes in the same way, but we are to validate them all in the hardest possible way. 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.

The argument-based approach to validity 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 involves a specific network of inferences and assumptions applied in a specific context. And therefore the details of the validity argument for each interpretive argument will also be unique. Yet, the general approach is consistent, or unified.

Although every interpretation is unique in some ways, it is possible to distinguish various kinds of interpretations involving certain general 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, I will draw a distinction 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). Their interpretive arguments focus on two inferences: the evaluation of specific responses and generalization of the resulting observations to a universe of observations that are of interest. 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 estimates of variance components, a reliability or generalizability coefficient, or 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 indicated by Cureton half a 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. (Cureton, 1950, p. 622)

The observable variable can be defined in terms of the average level of performance over some universe of possible tasks, and therefore can be stated 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 embedded in theories and derive most of their meaning from their role in the theory (Cronbach & Meehl, 1955). Theoretical constructs are not explicitly defined in terms of any observations, but by their roles in the theory. The empirical index used to estimate the value of the construct does necessarily rely on observations, but this index does not exhaust the meaning of the theoretical construct. The index actually employed may be one of many possible indices, and is likely to be

designed to be consistent with the assumptions in the theory and to yield the results predicted by the theory. The usefulness of the index 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 involves a number of inferences. The observed performances used to estimate 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. In addition, the theory defining the construct will generate empirical hypotheses involving the construct, and any observed relationships involving the indices for various 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 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 that I am drawing 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 state these hypothesized relationships in observable terms. 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 also depends on the context. A variable can be considered an observable attribute in a particular context as long as it does not rely on theoretical assumptions that are in dispute in that context. The interpretation of observable attributes

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.

In addition to defining the general content of the performance variables, 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, the performance-based interpretation can be stated without employing the theory currently under development. A universe of tasks can be specified without appeal to cognitive theories of performance for these tasks. To distinguish between the "carry" and "non-carry" tasks, it is necessary to know something about arithmetic, but a cognitive model of performance on subtraction problems is not needed.

Once defined, the observable attribute has a relatively simple interpretive argument, with a clear validation strategy. The strategy may not be easy to implement (implementing and validating a performance tests may be very difficult), and 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 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. They can be defined before theory gets highly developed, and arguably they have to be defined before the theory gets developed. How can we begin to 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 that both theories are trying to explain. 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 test. Without observable attributes, "critical" experiments would not be possible.

Third, the observable attribute may be of practical importance, independent of theory. It may be of importance to an employer to know whether sales clerks can add and subtract (with or without a calculator), independent of how they acquired the skills, or how they do it.

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 based on their performance on a sample of 20 analogy items) do not require extensive validity evidence. These relatively modest interpretive arguments can be supported by modest validity arguments. More expansive and ambitious interpretations (e.g., from observed scores on geometrical 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 this basic distinction.

The Role of Consequences in Validation

In a recent debate, Popham (1997) has argued for a limited, technical definition of validity, involving primarily the descriptive interpretation of scores. Popham (1997) prefers to treat validation as an objective, scientific concern, separate from disputes about consequences. He sees consequences as important, but would treat them as a separate concern. Linn (1997) and Shepard (1997) favored a broader conception of validity, which included the consequences of test use, as well as the descriptive interpretation of test scores.

As Shepard (1997) noted, 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. 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 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, as for real-world applications of anything, the bottom line involves consequences. We want the desirable consequences of using a measurement procedure to outweigh the negative consequences of such use (Cronbach & Gleser, 1965). If validity is to be "the most fundamental concern 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 test developer or a test user

responsible for every possible consequence of test use. So, the basic question is: who is to be responsible for what consequences of test use. I will not try to suggest a general answer to this question. My goal in this section is to suggest how an argument-based approach to validity might help to state the basic questions 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 comprehension 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 (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 (1980) 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 a perfectly valid measure of reading skills being used badly. It is also easy to conceive of a well-designed decision process that fails because of a poorly developed test (one that does not support the proposed interpretation).

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.

Under this arrangement, the work of validating the interpretation in terms of reading skill 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, 1980).

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 explicitly or implicitly (i.e., by labeling a test as a "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.

The evaluation of consequences is likely to be a contentious issue for a long time, and I do not mean to suggest easy solutions. Each application of measurement procedure will have to be evaluated on its own merits. But in clarifying the issues involved in assigning responsibility for the overall validation effort, I think that it will be useful to distinguish those parts of the interpretive argument that are likely to be most context dependent from those that are less context dependent.

Concluding Remarks

Validity is concerned with the clarification and justification 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 concensus on the appropriate uses of test scores in applied contexts. As a result, it has not been easy to formulate a general methodology principles 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 emphasized theoretical constructs, but the general principles introduced in the 1955 paper and subsequently developed by Cronbach, Meehl, Messick, Guion, Shepard, 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). 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, and is likely to be most effective in improving 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, and thereby pitting it against the most plausible alternate interpretations of the observed scores.

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, 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.

Brennan, R. L. (1983). Elements of generalizability theory. 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 for 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 validity. Assessment in Education, 3, 265-285.

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.

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. Volume IV. Minnesota studies in the philosophy of science. Minneapolis, MN: University of Minnesota Press.

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

Hacking, I. (1983). Representing and intervening

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

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

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

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

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

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.

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.

Linn, R.L. (1999). Assessments and accountability. Educational Researcher, 29, 2, 4-16.

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. (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.

Moss, P. (1993). Shifting conceptions of validity in educational measurement: Implications for performance assessment. Review of Educational Research, 62, 229-258.

Pellegrino, J. W. (1988). Mental models and mental tests. In H. Wainer & H. Braun (Eds.), Test validity (pp. 49-59). Hillsdale, NJ: Lawrence Erlbaum.

Pellegrino, J., Baxter, G. & Glaser, R (1999) Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. In Iran-Nejad, A. & Pearson, P. (Eds.), Review of research in education, V. 24 (pp. 307-353). Washington, DC: American Educational Research Association.

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

Snow, R. E. & Lohman, D. E. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), Educational Measurement, 3rd ed. (pp. 263-331). New York: American Council on Education and Macmillan.

Sternberg, R. J. (1985). Human abilities: An information processing approach. New York: W. M. Freeman.

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

Willingham, W. (1988). Testing handicapped people - The validity issue. In H. Wainer & H. Braun (Eds.), Test validity, (pp. 89-103). Hillsdale, NJ: Lawrence Erlbaum.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM031783

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Current Concerns in Validity Theory	
Author(s): Michael Kane	
Corporate Source:	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1 ↑ ☑

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A ↑ ☐

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B ↑ ☐

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Michael T. Kane	Printed Name/Position/Title: MICHAEL T. KANE / PROF.	
Organization/Address: U.W. Madison 2000 Observatory Dr. Madison, WI 53706	Telephone: 608-265-2891	FAX: 608-262-1657
	E-Mail Address:	Date: 8/11/00

(over)

Kane

ERIC Clearinghouse on Assessment and Evaluation

University of Maryland
1129 Shriver Laboratory
College Park, MD 20742-5701

Tel: (800) 464-3742
(301) 405-7449
FAX: (301) 405-8134
ericae@ericae.net
http://ericae.net

May 8, 2000

Dear AERA Presenter,

Hopefully, the convention was a productive and rewarding event. As stated in the AERA program, presenters have a responsibility to make their papers readily available. If you haven't done so already, please submit copies of your papers for consideration for inclusion in the ERIC database. We are interested in papers from this year's AERA conference and last year's conference. If you have submitted your paper, you can track its progress at http://ericae.net.

Abstracts of papers accepted by ERIC appear in *Resources in Education (RIE)* and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of *RIE*. Abstracts of your contribution will be accessible through the printed and electronic versions of *RIE*. The paper will be available through the microfiche collections that are housed at libraries around the world and through the ERIC Document Reproduction Service.

We are gathering all the papers from the **2000 and 1999 AERA Conference**. We will route your paper to the appropriate clearinghouse. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality.

Please sign the Reproduction Release Form enclosed with this letter and send **two** copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can mail your paper to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to:

AERA 2000/ERIC Acquisitions
University of Maryland
1129 Shriver Laboratory
College Park, MD 20742

Sincerely,

Lawrence M. Rudner

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE