Commentary on “Validating the Interpretations and Uses of Test Scores”

Robert L. Brennan
University of Iowa

Kane's paper “Validating the Interpretations and Uses of Test Scores” is the most complete and clearest discussion yet available of the argument-based approach to validation. At its most basic level, validation as formulated by Kane is fundamentally a simply-stated two-step enterprise: (1) specify the claims inherent in a particular interpretation and/or use of test scores (IUA); and (2) provide an evaluation of the claims (validity argument). Kane discusses four types of inferences that provide a scaffolding for addressing these two arguments: scoring, generalization, extrapolation, and decision rules. Decision rules, in particular, are closely related to consequences, which loom large in the argument-based approach to validation. The present commentary on Kane’s paper attempts to simplify some of his discussions, while expanding others. The author suggests that Kane’s argument-based approach to validation offers by far the best current basis for optimism about improvements in validation.

Kane’s article (this issue) “Validating the Interpretations and Uses of Test Scores” is the most recent product of three decades of work that he has devoted to research on validity and validation. Until now, Kane’s essay on validation in the 4th edition of Educational Measurement (2006) was by far the most extensive treatment of the subject. The current article extends the earlier essay largely, I think, by making his perspectives on validation—specifically, the argument-based approach to validation—clearer and more accessible to measurement practitioners and others.

The article is quite long, which is justified in terms of the large number of issues covered and how well they are treated. Still, the length of the paper may impede its accessibility, and it is helpful to recall the five-section, high-level structure of the paper:

1. The Roots of the Argument-Based Approach to Validation, which is essentially an historical perspective on the validity/validation literature.
2. An Argument-Based Approach to Validation, which outlines Kane’s basic ideas.
3. Tailoring Validation to Proposed Interpretations, which deals mainly with types of inferences for observable attributes.
4. Theory-Based Interpretations, which deals mainly with theoretical constructs.
5. Score Uses, which deals mainly with consequences and decision rules.

At its most basic level, validation as formulated and elucidated by Kane is fundamentally a simply-stated two-step enterprise: (1) specify the claims inherent in a particular interpretation and/or use of test scores; and (2) provide an evaluation of the claims based on empirical evidence, logical arguments, etc. Kane calls the first
step the interpretation/use argument (IUA) and the second step the validity argument. Obviously, these steps get repeated for as many claims (or sets of claims) as there may be. Secondarily, Kane suggests that conducting validation studies is usually facilitated by considering various types of inferences (scoring, generalization, extrapolation, and decision rules), although these types of inferences are not apparent in the five-section, high-level structure of the paper.

In a sense, the argument-based approach to validation is not all that novel. The core of this approach “is the basic scientific and social principle that, in public discourse, claims should be supported by appropriate evidence” (Kane, p. 64). What is novel, or at least extraordinarily helpful, is the clarity and persuasiveness of the case Kane makes for adopting the argument-based approach for dealing with the very real practical problems of validation. Using this approach, validation may be challenging, but it is doable; and for many claims, extensive psychometric experience is not required. What is required is clear specifications of IUAs and careful evaluation of them—that is basically it: validation = IUAs + evaluations. There may be devilish details to be considered, but the basic approach is straightforward.

The comments that follow offer somewhat different perspectives on certain issues or expansions of the treatment of other issues. Although I believe that the core of Kane’s argument-based approach to validation is very simple and consistent with scientific inquiry, that does not mean that the measurement community has arrived at consensus on all issues.

**Validity versus Validation**

Early in his article, Kane provides an excellent discussion of “The Roots of the Argument-Based Approach to Validation.” Here I offer a slightly different, and somewhat editorialized, view of this issue. I make no attempt to reference all contributors.

At the risk of oversimplification, the historical literature of validity/validation can be divided into three somewhat overlapping eras: (a) the beginnings to about 1950 in which prediction dominated, along with a recognition of content validity; (b) about 1950 to about 1990 in which discussions about construct validity predominated; and (c) about 1980 to the present which witnessed the development of validation paradigms (most notably the argument-based approach) for addressing practical problems.

In the first edition of *Educational Measurement* Cureton (1951) states: “Validity is therefore defined in terms of the correlation between the actual test scores and the ‘true’ criterion scores” (pp. 622–623). Cureton is faithfully reflecting the fact that the “theory of prediction was very nearly the whole of validity until about 1950” (Cronbach, 1971, p. 443). Afterwards, everything changed.

It would be difficult to overstate the influence of Cronbach and Meehl (1955) on the history of validity. That single paper was the proximate cause for construct validity taking center stage in the validity arena for over 30 years, culminating, I think, in the superb treatment of validity qua theory by Messick (1989).

During the second era, phrases such as “construct validity is all of validity” or validity is a “unitary” notion tended to deify construct validity. But our psychometric deities, like the Greek gods, tend to have imperfections. In this case, the laser-like
focus on construct validity too often impeded addressing the practical problems of validation.

Perhaps inevitably, two versions of construct validity emerged: a “strong” program that pretty faithfully mirrored the Cronbach and Meehl (1955) paper, and a “weak” program of “Dragnet empiricism: ‘just give us the facts, ma’am…any facts’” (Cronbach, 1989, p. 156). But the strong program tended to make validity an interminable exercise, and the weak program tended to reinforce the already unfortunate tendency to elevate marginally relevant facts into a validity argument for a vague or poorly stated claim.

The Cronbach and Meehl paper (1955) represented the most important line of validity inquiry in the second era, but evaluation represented another line that was pursued during a nearly coincident time frame. Evaluation became a public, large-scale activity during the 1960s, when many if not most federally sponsored educational and social interventions were required to have an evaluation component. These evaluations tended to be very pragmatic but were often disjointed. Their limitations became evident in the 1970s and led to serious reconsideration of the field of evaluation by researchers such as House (1980) and especially Cronbach (1982).

Important segments of the evaluation literature generated during the 1980s reflect the claims-evidence perspective that characterizes the argument-based approach to validation, although the evaluation literature made little explicit use of this terminology. In this sense, evaluation made contributions to validation and vice-versa. (Note, for example, the substantial role that Cronbach played in both validation and evaluation.)

Kane’s contributions to validation often are viewed as beginning with his 1992 paper entitled “An Argument-Based Approach to Validation” and, up until now, culminating in his 2006 chapter on validation in the 4th edition of Educational Measurement. In my view, however, the genesis of Kane’s current thinking about these matters can be found in two papers published 10 years earlier, in 1982: “A Sampling Model for Validity” (Kane, 1982a) and “The Validity of Licensure Examinations” (Kane, 1982b). Among many other things, the first paper introduced the distinction between generalization and extrapolation (although he did not use the term “extrapolation”) in the context of using generalizability theory to provide an elegant solution to the reliability-validity paradox. The second paper presents a theoretically sound but still practical discussion of numerous validity issues in licensure examinations—a discussion that is still relevant today.

So, at the risk of oversimplification, the development of the argument-based approach to validation was influenced by: (a) the limitations of construct validity for providing a framework for addressing practical validation issues; (b) the reformulation of evaluation along lines consistent with validation, largely due to Cronbach; and (c) the efforts of Kane (and others including, e.g., Cronbach, 1988; Shepard, 1993) to formulate validation paradigms that were faithful to the basic principles of validity qua theory. Although validation and validity qua theory are not the same, the argument-based approach to validation is practical while still being faithful to basic scientific principles.

What Guion (1980) called the trinitarian model of validity (i.e., content, criterion, and construct) has been roundly criticized for decades as being either too narrow
content and criterion) or too abstract (construct). To combat these limitations, in recent years more often than not the measurement field has resorted to addressing validity issues through reference to types of “evidence,” with discussions often being little more than disguised versions of the elements of the trinitarian model. The fundamental problem with the “sources of evidence” approach is that it focuses almost exclusively on evidence without clear specifications of claims (the IUAs). Without clearly specifying claims a priori, it is at best difficult (and often impossible) to evaluate the relevance of evidence. In a sense, the principle virtue of the argument-based approach to validation compared with past approaches is that it puts claims “front and center” where they belong.

Types of Inferences

The claims-evidence duality that characterizes the IUA and validity argument is not unique to Kane. As Kane acknowledges, in one form or another this duality is evident in the writings of others including philosophers of science and measurement experts. Kane’s categorization of types of inferences, however, is a novel and helpful contribution to conducting validation.

Kane discusses four types of inferences: scoring, generalization, extrapolation, and decision rules. This set of four inferences provides a very convenient and theoretically appealing scaffolding for considering the component parts of an IUA and validity argument, although not all types of inferences are necessarily required for each IUA and validity argument.

The comments in this section are intended to expand and/or emphasize certain aspects of Kane’s discussion of inferences. Any inconsistencies with Kane are largely “at the margins” of his arguments.

Scoring

The primary focus of scoring inferences is rules and procedures for obtaining the observed scores that are ultimately used for interpretations and decisions. With respect to scoring inferences, Kane states that:

The process typically begins by scoring the observed performances and combining the task scores in some way, yielding an observed score of some kind...It typically makes assumptions about the appropriateness of the scoring criteria and the rules for combining scores. (p. 10)

This is certainly an accurate description of scoring, but it is, I think, too succinct to provide much guidance for validation.

For most traditional tests, examinee scores are determined by:

1. scoring items (objectively or subjectively);
2. obtaining an examinee’s raw score as an unweighed or weighted sum of item scores;
3. transforming raw scores (often nonlinearily) to scale scores used for reporting and decision making; and
4. employing equating in the previous step if multiple forms exist.
This approach may be traditional but it is not trivial, and reports documenting the actual procedures used for scaling and equating are often nonexistent or superficial, so much so that it is usually impossible for even a qualified psychometrician (not employed by the test developer) to know the full set of details required to replicate reported scores. The problem is made even more difficult by the widespread use of proprietary computer programs that hide important details.

Today, scoring inferences often are even more complicated by the use of item response theory (IRT), which involves at least the following considerations, most of which are in addition to 1–4 above:

- strong assumptions about unidimensionality (usually) and local independence;
- choice of model (e.g., Rasch, 2PL, 3PL, etc.);
- choice of procedure for estimating item parameters (usually marginal maximum likelihood, but not always);
- choice of procedure for estimating proficiency [usually maximum likelihood (ML) or expected a posteriori (EAP)]; and
- transformation of the \((-\infty, \infty)\) proficiency scale to some other scale used for reporting.

Matters are even more complex, of course, when adaptive testing is employed (which is becoming more frequent). Furthermore, scale scores often are grouped into score categories, which adds the additional complexity of standard setting.

It is important to recognize that the “final” observed scores are the scores used to make inferences. Item scores and other intermediate scores are relevant only in the sense that they are part of a chain of computations that lead to the observed scores, which I will call “reported” scores. The characteristics of intermediate scores are at least one step (and often many steps) removed from the scores of principal interest, namely the reported scores that are the basis for score interpretations and uses. Kane does not draw the distinction between observed and reported scores; I do so here because I think it helps clarify various issues involved in scoring inferences, as well as other types of inferences.

The complexity of the manner in which most reported scores are obtained presents a considerable challenge to constructing claims about such scores and providing evidence to support such claims. Indeed, in my opinion, scoring inferences often are among the most difficult and least well understood inferences in measurement—inferences that are camouflaged by the apparent simplicity of the resulting numbers (or score categories with value-laden terms such as “basic,” “proficient,” and “advanced”). For example, just about everyone thinks that a score of 500 on the SAT is “average,” but in what sense is this scoring inference supported by one or more well-formulated claims and accompanying evidence? What about a score of 500 on any one of the many SAT subject tests?

Scoring inferences almost always involve scaling, which is one of the most challenging problems in psychometrics. In part it is challenging largely because it necessitates making a set of judgments that are arbitrary in some sense, and different sets of judgements lead to different characteristics of score scales (see Kolen & Brennan, 2004, especially chap. 9). In most cases, no amount of data can justify the judgments that are made, although data can provide evidence about certain characteristics of
score scales. Note also that scaling almost always involves nonlinear transformations, which complicates matters substantially.

In short, validating scoring inferences is almost always nontrivial and too often overlooked. Reported score-scale numbers or categories may appear simple or obviously easy to interpret, but that apparent simplicity usually hides very complex issues that make implicit claims about test scores—claims that beg to be explicitly specified and validated.

**Generalization and Extrapolation**

Kane provides an extensive and excellent discussion of generalization and extrapolation inferences. Here I comment on only a few matters.

Generalization and extrapolation inferences differ from scoring inferences in one very important respect: scoring inferences usually have little if anything to do with errors of measurement, whereas errors of measurement are intimately involved in generalization and extrapolation.

Generalization involves inferences from observed scores (in the sense of reported scores discussed previously) to universe scores, which are the expected values of observed scores over investigator-specified conditions of measurement (e.g., over tasks, occasions, raters, contexts, etc.). The set of such universe scores is the universe of generalization in the terminology of G theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Brennan, 2001b).

Kane notes: “To simply estimate coefficient alpha or some other measure of internal consistency and assume that the question of reliability/generalizability has been addressed is to beg the question of generalizability” (p. 20). This particularly important point was highlighted by Cronbach himself in the last paper that he wrote (with help from Shavelson) prior to his death (Cronbach, 2004). The essential problem is that the only source of error that affects alpha is associated with sampling of items. Generalization is even narrower when IRT is used in isolation (see Brennan, 2001a, pp. 303–305); indeed, I would argue that when IRT is used, it is more properly associated with scoring inferences than generalization inferences.

The scores that are analyzed for generalization inferences should be consistent with those that result from scoring inferences. Generalizability/reliability analyses are only marginally relevant to generalization inferences when the results are based on intermediate scores rather than the reported scores that are at the heart of inferences.

As noted above, generalization involves inferences from observed scores to a universe of generalization. By contrast, extrapolation involves inferences from observed scores to what Kane calls a “target” domain.

In my view, there are two types of extrapolation. First, what I will call an interpretive extrapolation (IE) involves generalization to a larger set of conditions of measurement than that involved in the generalization inference. For an IE the universe of generalization is a subset of the target domain. This distinction is fundamental to Kane’s (1982a) explanation of the reliability-validity paradox. So, for example, a generalization inference might involve generalizing over tasks and raters only, whereas an extrapolation inference might involve generalizing over tasks, raters, and occasions.
Second, what I will call a loose extrapolation (LE) is what Kane characterizes as a “looser kind of inference found in science, academic discourse, and practical reasoning” (p. 11). A LE often involves inferences from observed scores to criterion scores or other scores that are only loosely aligned to observed scores. A classic example would be the use of ACT or SAT scores to predict first year college grade point averages. For a LE the relationship between the target domain and universe of generalization is usually not as direct as it is for an IE.

Sometimes the distinctions among types of inferences are not quite as clean as terminology would suggest. For example, in my view it seems reasonable to include scaling as part of a scoring inference (as I did in the previous section) when the scaling procedure does not involve explicit consideration of errors of measurement. Sometimes, however, scaling does consider errors of measurement (see, for example, Kolen & Brennan, 2004, chap. 9).

Test Uses, Decision Rules, Consequences, and Responsibility

Kane argues persuasively that, since test score uses can be part of the IUA, consequences necessarily belong in validation (although he appears to leave open the possibility that some consequences might be put into a different category than validation). I think the time has come to assert that any use of test scores falls within the purview of validation, provided there is a clear recognition that responsibility for validation (including decision rules) for any particular use rests with the person or entity that makes the IUA. Uses imply consequences, and users must take responsibility for consequences; otherwise, validation of test score uses is likely to be vacuous.

It follows that test developers are not responsible for validation of claims not made by them, although test developers do have an affirmative responsibility to warn users about reasonably anticipated misuses of test scores. Admittedly, including all test score uses under validation occasionally will present challenges to test developers. However, in an open society such as ours these challenges seem unavoidable. The following two examples may be informative.

In the early 1980s, U.S. Secretary of Education Terrel H. Bell created what came to be known as the “Wall Chart,” which rank-ordered states with respect to their SAT and ACT Assessment scores. Since college admissions tests are taken by self-selected groups of examinees, neither the College Board nor ACT endorsed this use of their test scores. Both organizations made their positions known, although I doubt that the public and most of the media knew about the self-selection problem. Still, I think it can be argued that the College Board and ACT behaved reasonably with respect to their validation responsibilities. The same cannot be said about the Department of Education which, to the best of my knowledge, never made any serious efforts to validate this use of SAT and ACT assessment scores. In short, I would argue that the user (Department of Education) failed in its validation responsibilities, but that failure does not “count” against the College Board or ACT.

The Wall Chart example is relatively straightforward. Other situations are much more challenging. Currently, for example, there is great interest in the use of value-added models (VAMs) for evaluating teachers. VAMs typically use student test
scores as variables, along with other variables. In most cases known to me, test developers do not make claims and provide evidence that would support the use of their test scores for teacher evaluations. Under these circumstances, the responsibility for validating teacher evaluations based on VAMs seems to rest primarily with the users (and perhaps even the developers) of the models. Does this seemingly straightforward logic, however, absolve the test developer from all validation responsibility? I think not. The stakes are so high that it seems to me that test developers have a responsibility to inform users about potential limitations of using their test scores in VAMs for teacher evaluation. Doing so, however, is likely to be challenging, especially when the message is not welcome.

Often, results based on VAMs are phrased in ways that suggest that teacher performance causes changes in test scores, with that argument used to support awards/sanctions for teachers. Such a causal-inference IUA is challenging to formulate and extraordinarily difficult to validate. Kane considers causal inferences in his section on theory-based interpretations. I would argue that causal inferences can apply to both observable and theoretical attributes, and causal inferences are almost always associated with uses of test scores. From this perspective, I suggest that discussions of causal inference usually belong in discussions of test uses and consequences.

**Concluding Comments**

Over 50 years ago Ebel observed that “Validity has long been one of the major deities in the pantheon of the psychometrician. It is universally praised, but the good works done in its name are remarkably few. Test validation, in fact, is widely regarded as the least satisfactory aspect of test development” (Ebel, 1961, p. 640). If Ebel were alive today, undoubtedly he would acknowledge the substantial amount of work that has been published about validity and validation in the last 50+ years, but I expect he would still bemoan the lack of “good works done in its name.” Why? Where has the field gone wrong—or has it? I think that at least part of the answer is that it took the measurement field a long time to distinguish overtly and clearly between validity qua theory the practice of validation. They are linked, but they are not synonymous or isomorphic.

So, where do we stand now? In my judgment, validity qua theory is in good shape due in large part to Messick and Cronbach (among others) who used their own ingenuity, as well as the writings of numerous philosophers of science, to craft a quite coherent theory of validity. In a sense, validation qua practice is in much better shape than when Ebel bemoaned its good works. With the contributions of Kane (and others) we now have a practical, useful scaffolding that provides ways to frame and address claims about test score interpretations and uses. In another sense, however, validation is still quite impoverished. Relatively few testing programs give validation the attention it deserves. There is a notable lack of a clear listings of the claims made with a corresponding evaluation of each of them. An uncoordinated discussion of a subset of validity matters does not qualify as validation; nor does the mere presence of a chapter or document labeled “Validity” or “Sources of Validity Evidence” constitute validation in a meaningful sense.
At the beginning of this commentary, I emphasized that Kane’s argument-based approach to validation is fundamentally a simply-stated two-step enterprise: validation = IUAs + evaluations. As simple as this is, it is pretty much worthless unless both the IUAs and the evaluations are readily available to users of test scores. Transparency should not be optional, documentation is essential, and black boxes militate against validation. Unfortunately, far too many tests (even well-known, large-scale tests) lack adequate, timely documentation with respect to validation, even validation of the most fundamental claims.

Still, there are reasons to be optimistic. In particular, I believe that Kane’s development and explanation of the argument-based approach to validation offers, by far, the best current basis for optimism about improvements in validation. Hopefully, in the future, there will be a substantial increase in the good works done in the name of validation. Undoubtedly, as measurement professionals gain more experience with the argument-based approach to validation, improvements will be made, but I think the basic claims-evidence paradigm will survive, along with focusing on types of inferences.

Notes

1 Henceforth, any reference to Kane means, more explicitly, Kane (this issue).
2 If there are claims that cannot be validated, they count against the test developer or user.
3 The Cronbach and Meehl paper was published in 1955, but both Cronbach and Meehl were on the committee that wrote the 1954 Standards that considered construct validity. Furthermore, a draft of these Standards was published in 1952 in which the phrase “congruent validity” was used as a kind of precursor to “construct validity.”
4 The author participated in several of these, including evaluations of Head Start, Follow Through, and the National Day Care Study.
5 In these efforts, in my opinion Kane was substantially influenced by his background in physics (B.S. and M.A.), the philosophy of science (Patrick Suppes was his Ph.D. advisor), and generalizability theory, coupled with his practical and extensive experience with several large-scale testing programs.
6 This is a somewhat more unqualified position than I have taken previously (Brennan, 2006, pp. 14–15).
7 I use the phrase “test developers” to include any and all persons or entities that develop or own the testing program.

References


**Author**

ROBERT L. BRENNAN is E. F. Lindquist Chair in Measurement and Testing and Director of the Center of Advanced Studies in Measurement and Assessment (CASMA), University of Iowa, 210D Lindquist, Iowa City, IA 52242; robert-brennan@uiowa.edu. His primary research interests include generalizability theory, classical test theory, equating, and scaling.