JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education

Journal of Educational Measurement
Spring 2013, Vol. 50, No. 1, pp. 1–73

Validating the Interpretations and Uses of Test Scores

Michael T. Kane
Educational Testing Service

To validate an interpretation or use of test scores is to evaluate the plausibility of the claims based on the scores. An argument-based approach to validation suggests that the claims based on the test scores be outlined as an argument that specifies the inferences and supporting assumptions needed to get from test responses to score-based interpretations and uses. Validation then can be thought of as an evaluation of the coherence and completeness of this interpretation/use argument and of the plausibility of its inferences and assumptions. In outlining the argument-based approach to validation, this paper makes eight general points. First, it is the proposed score interpretations and uses that are validated and not the test or the test scores. Second, the validity of a proposed interpretation or use depends on how well the evidence supports the claims being made. Third, more-ambitious claims require more support than less-ambitious claims. Fourth, more-ambitious claims (e.g., construct interpretations) tend to be more useful than less-ambitious claims, but they are also harder to validate. Fifth, interpretations and uses can change over time in response to new needs and new understandings leading to changes in the evidence needed for validation. Sixth, the evaluation of score uses requires an evaluation of the consequences of the proposed uses; negative consequences can render a score use unacceptable. Seventh, the rejection of a score use does not necessarily invalidate a prior, underlying score interpretation. Eighth, the validation of the score interpretation on which a score use is based does not validate the score use.

Test scores are of interest because they are used to support claims that go beyond (often far beyond) the observed performances. We generally do not employ test scores simply to report how a test taker performed on certain tasks on a certain occasion and under certain conditions. Rather, the scores are used to support claims that a test taker has, for example, some level of achievement in some domain, some standing on a trait, or some probability of succeeding in an educational program or other activity. These claims are not generally self-evident and merit evaluation.

To validate an interpretation or use of test scores is to evaluate the plausibility of the claims based on the test scores. Validation therefore requires a clear statement of the claims inherent in the proposed interpretations and uses of the test scores. Public claims require public justification.

The argument-based approach to validation (Cronbach, 1988; House, 1980; Kane, 1992, 2006; Shepard, 1993) provides a framework for the evaluation of the claims based on the test scores. The core idea is to state the proposed interpretation and use explicitly and in some detail, and then to evaluate the plausibility of these proposals.

In many cases, tests are specifically designed and developed to support certain decisions about test takers (e.g., selection, diagnosis, placement) by providing information about test-taker attributes that are relevant to the decision. In some cases, tests are designed to assess attributes, which are relevant to a range of decisions in a range

of contexts. For example, a test of communicative competence in some language could be interpreted in terms of the test taker's ability to use the language effectively in a range of contexts, and the test then might be used by various institutions to make various decisions (Chapelle, 1999).

In addition to their use in making decisions about individual test takers, tests also have been used for policy analysis, program evaluation, research, and educational accountability; in all of these cases, the requirements imposed by the intended use shape the design and development (or the selection) of the tests. Concerns about validity have their roots in public concerns about the appropriateness, or fitness, of the test scores for their intended use or uses:

> The question of validity would not be raised so long as one man uses a test or examination of his own devising for his private purposes, but the purposes for which schoolmasters have used tests have been too intimately connected with the weal of their pupils to permit the validity of a test to go unchallenged. The pupil . . . is the dynamic force behind the validity movement. . . . Further, now that the same tests are used in widely scattered places and that many very different tests all going by the same name are gently recommended by their respective authors, even the most complacent schoolmen, the most autocratic, and the least in touch with pupils, are beginning to question the real fitness of a test. (Kelley, 1927, pp. 13–14)

Bachman (2002) and Bachman and Palmer (2010) have developed a validation framework that focuses on test uses. The proposed use is specified in an *assessment use argument*, or AUA, and the use is validated by evaluating the plausibility of the AUA.

I am sympathetic to Bachman and Palmer's emphasis on score use, but I am inclined to give interpretations and uses equal billing. Historically, validity has covered evaluations of interpretations (e.g., see Cronbach & Meehl, 1955) and uses (e.g., see Cronbach & Gleser, 1965). Interpretations involve claims about test takers or other units of analysis (e.g., teachers, schools), and score uses involve decisions about these units of analysis. Interpretations and the uses of the test scores tend to be entwined in practice; tests tend to be developed with some interpretation and some use or uses in mind, and arguments for the appropriateness of a score use typically lean heavily on the relevance of score interpretations.

In the past, I have talked about "interpretive arguments" as explicit statements of the inferences and assumptions inherent in the interpretation and use of test scores (Kane, 1992, 2002b, 2006), but this expression may give too much weight to interpretations and not enough to uses. To rectify this imbalance, I will shift my terminology and talk about an "interpretation/use argument" (or "IUA") where the IUA includes all of the claims based on the test scores (i.e., the network of inferences and assumptions inherent in the proposed interpretation and use). Some IUAs may focus on a particular use, while others may involve an interpretation in terms of a skill or disposition to behave in some way and allow for a range of possible uses.

The validity of a proposed interpretation or use of test scores at any point in time can be defined in terms of the plausibility and appropriateness of the proposed interpretation/use at that time. A proposed interpretation or use can be considered valid to the extent that the IUA is coherent and complete (in the sense that it fully represents

the proposed interpretation or use) and its assumptions are either highly plausible *a priori* or are adequately supported by evidence. The kinds of evidence required for validation are determined by the claims being made, and more-ambitious claims require more evidence than less-ambitious claims.

If the IUA does not claim much (e.g., that students with high scores on the test can generally perform the kinds of tasks included in the test), it does not require much empirical support beyond data supporting the generalizability of the scores. A more-ambitious interpretation (e.g., one involving inferences about some theoretical construct) would require more evidence (e.g., evidence evaluating the theory and the consistency of the test scores with the theory) to support the additional claims being made. If the scores are to be used to predict future outcomes (in an employment- or placement-testing context), evidence indicating that the predictions are accurate is called for, but if no such predictions are anticipated, this kind of predictive evidence is essentially irrelevant. To the extent that claims in the IUA are not adequately supported by the evidence, or worse, are contradicted by the evidence, the proposed interpretations and uses would not be considered valid.

Validity is a matter of degree, and it may change over time as the interpretations/uses develop and as new evidence accumulates. The plausibility of a proposed IUA will increase if ongoing research supports its inferences and assumptions (especially those that are most questionable *a priori*). Validity may decrease if new evidence casts doubt on the proposed IUA. For example, if it is found that the test is coachable or that it is difficult to grade reliably in operational contexts, the plausibility of the proposed IUA may decline.

Validity is not a property of the test. Rather, it is a property of the proposed interpretations and uses of the test scores. Interpretations and uses that make sense and are supported by appropriate evidence are considered to have high validity (or for short, to be valid), and interpretations or uses that are not adequately supported, or worse, are contradicted by the available evidence are taken to have low validity (or for short, to be invalid). The scores generated by a given test can be given different interpretations, and some of these interpretations may be more plausible than others.

Evaluations of precision (e.g., standard errors) play an important role in evaluating claims based on test scores, because almost all test-score interpretations involve generalizations over some conditions of observation (e.g., over tasks, occasions, raters, and/or contexts) and our estimates of precision characterize the dependability of such generalizations. So evidence for the generalizability (or reliability) of scores over conditions of observation is generally necessary in making a case for validity. In addition, most interpretations/uses involve claims (e.g., about relationships to nontest performance) that go beyond generalizations, and these claims also require other kinds of evidence (e.g., correlations with criterion measures). Evidence of generalizability (or reliability) therefore rarely is sufficient for validity.

The assumptions underlying various claims (e.g., that an attribute does not change over time) are hardly ever exactly true. So if we require complete, absolute confirmation of all of the claims in a proposed interpretation/use, every interpretation and use would be shot down immediately. We have to build some slack into the system (like the expansion joints in highways), and typically we do this by postulating the existence of errors (random and systematic) of various kinds. The explicit recognition

of uncertainty makes the interpretations viable, but it also makes interpretations a bit fuzzy and decisions a bit tentative. As is the case in evaluating scientific theories, we never achieve certainty but we can achieve a high degree of confidence in certain interpretations and uses of test scores.

The Roots of the Argument-Based Approach to Validation

In the late 19th and early 20th centuries, test developers seemed to assume that the ability to be assessed existed in some sense, and they sought to develop estimates of each person's standing on the ability. The tests were to assess mental abilities, (e.g., general intelligence, memory, attention), and the test scores were interpreted as indicators of a mental ability on which people differed unidimensionally (more or less like height). The tests were designed to measure these attributes more systematically and precisely (i.e., more scientifically) than was possible in ordinary contexts. So the proposed interpretation (i.e., the ability) was taken as given, and the test would be evaluated in terms of how accurately it seemed to represent the mental ability (Sireci, 1998; Spearman, 1904; von Mayrhauser, 1992). The test developers would use their sense of the mental ability to develop test tasks that seemed to require the mental ability or some more basic ability (e.g., discriminations of various kinds) subsumed under the ability of interest (Spearman, 1904).

Criterion-Based and Content-Based Approaches

By around 1915, the notion of criterion validity was in use. With a criterion measure that was assumed to approximate the "real" value of the attribute of interest, validity could be evaluated in terms of the relationship between test scores and criterion scores (Thorndike, 1918). The early work on criterion-related validation mainly seems to have addressed applied problems in selection and placement with criteria specified in terms of desired outcomes (von Mayrhauser, 1992).

Between 1920 and 1950, statistical models for test-criterion relationships were refined, and criterion validity became the gold standard for validity (Angoff, 1988; Cronbach, 1971; Moss, 1992, 1995; Shepard, 1993; Sireci, 2009). In the first edition of *Educational Measurement*, Cureton (1951) defined validity as "the correlation between the actual test scores and the 'true' criterion score" (p. 623)—the test-criterion correlation corrected for unreliability in the criterion. A test could be considered valid for any criterion for which it provided accurate estimates (Gulliksen, 1950).

Developing a criterion measure is not too difficult in applied settings in which the goal is to promote some desired performance (or some desired outcome); the criterion measure can be defined in terms of assessments of those desired performances (Cureton, 1951; Ebel, 1961). For example, performance in college is routinely evaluated in terms of grade point averages, and performance on a job can be evaluated in terms of supervisor evaluations or some outcome measure.

Developing criterion measures is more challenging where the goal is to measure some ability or attribute that is not defined in terms of a specific behavior or performance. For example, scores on measures of mental abilities can be compared to other measures of the ability of interest, but none of these external measures is necessarily more valid than the test. In some cases, the ability of interest (e.g., the

ability to solve algebra problems) can be defined in terms of some performance domain and a criterion measure can be developed by thoroughly sampling the domain (Cronbach, 1971, 1980; Guion, 1998), but in these cases the criterion might simply look like a longer and less standardized version of the test. The criterion-based model is quite useful and elegant (Cronbach & Gleser, 1965; Cureton, 1951), but coming up with a suitable criterion can be difficult or impossible in many cases.

Most of the early tests of mental ability and many current standardized tests of various kinds (e.g., achievement tests in academic content areas) have been justified primarily in terms of "a review of the test content by subject-matter experts" (Angoff, 1988, p. 22). Assuming that a person's performance is evaluated on a sample of tasks from a domain, it is legitimate to take the observed sample of performance as an estimate of overall performance in the domain if (1) the observed performances can be considered a representative sample from the domain, (2) the performances are evaluated appropriately and fairly, and (3) the sample is large enough to control sampling error (Guion, 1977).

Content-based analyses usually are developed during or soon after test development by people involved in test development; as a result, the content-based approach has a potential confirmatory bias (Guion, 1977) but it is useful in evaluating the relationship between the sample of performance in the test and a larger performance domain that is of interest. Content-based analyses can always be challenged, but a reasonable case can be made for interpreting a direct measure of performance on certain tasks (e.g., playing the piano) in terms of level of skill in performing that kind of task, especially if the performance domain has been carefully specified, the domain has been systematically sampled, and the performances were evaluated appropriately (Cronbach, 1971; Cureton, 1951; Ebel, 1961; Flockton & Crooks, 2002; Kane, 1982, 1996; Kane, Crooks, & Cohen, 1999; Lissitz & Samuelsen, 2007; Sireci, 1998). Criterion measures generally rely on the content model for their justification.

The Construct Model

In their conceptualization of construct validity, Cronbach & Meehl (1955) shifted the focus from the development of a test for a given interpretation to the relationship between the test and a proposed interpretation. They developed their construct validation framework in terms of then-current views in the philosophy of science that theoretical constructs would be implicitly defined by their roles in a theory. For example, the construct of atomic weight depends on an atomic theory of matter, and without such a theory, atomic weight is meaningless. The construct is implicitly defined by the theory, and the construct interpretation and the theory are empirically evaluated together. If predictions derived from the theory do not agree with observations, then either the theory is wrong or the measures for some of the constructs are not appropriate (or some ancillary assumption was violated). If the predictions are confirmed empirically, both the theory and the interpretation of scores in terms of the theoretical constructs are supported.

The rigorous implementation of this approach to validation requires a well-defined theory from which empirical predictions can be derived, and Cronbach and Meehl (1955) recognized that this limited the applicability of the model in its pure form:

> The idealized picture is one of a tidy set of postulates which jointly entail the desired theorems. . . . In practice, of course, even the most advanced physical sciences only approximate this ideal. . . . Psychology works with crude, half-explicit formulations. (Cronbach & Meehl, 1955, pp. 293–294)

Because "the network still gives the constructs whatever meaning they do have" (Cronbach & Meehl, 1955, p. 294), a loose theory yields vaguely defined construct interpretations.

In addition to expanding the definition of validity to include a systematic but complex approach to the validation of implicitly defined constructs, Cronbach and Meehl (1955) shifted the focus from the validation of a test given an interpretation or use to the validity of the proposed interpretation of the test scores (Cronbach, 1971; Loevinger, 1957). Before the advent of construct validity, the score interpretations were treated as if they were relatively simple and well understood. They were either a prediction of some operationally defined criterion, a plausible measure of some mental ability, or a summary measure of performance in some domain. The attribute to be measured was taken as a given, and the focus was on the development and validation of the test as a measure of the attribute. For Cronbach and Meehl, the interpretation of a theoretical construct involved a scientific theory which was obviously not simple and could not be taken for granted but had to be justified in some way.

Development of the Construct Model: 1955–1989

Cronbach and Meehl (1955) presented construct validity as an alternative to the criterion and content models. Construct validity could be used "whenever a test is to be interpreted as a measure of some attribute or quality which is not operationally defined" (p. 282) and "for which there is no adequate criterion" (p. 299), although they also said that "determining what psychological constructs account for test performance is desirable for almost any test" (p. 282).

The 1966 *Standards* (American Psychological Association, American Educational Research Association, & National Council on Measurement in Education, 1966) and the 1974 *Standards* (American Psychological Association, American Educational Research Association, & National Council on Measurement in Education, 1974) considered four kinds of validity associated with four kinds of interpretation (predictive, concurrent, content, and construct). Cronbach (1971) continued to associate construct validation with theoretical variables for which "there is no uniquely pertinent criterion to predict, nor is there a domain of content to sample" (p. 462), but he also suggested that any description "that refers to the person's internal processes (anxiety, insight) invariably requires construct validation" (p. 451). He also emphasized the need for an overall evaluation of validity involving multiple kinds of evidence and the need for "an integration of many types of evidence" (p. 445).

By the late 1970s, validity theorists (Cronbach, 1980; Guion, 1977, 1980; Messick, 1975, 1981; Tenopyr, 1977) became concerned about a tendency to treat the different validation models as a toolkit, with different methods to be employed in different cases. In the absence of a general framework for validation, the choices made in developing validity evidence often were dictated mainly by the availability of data. Gradually, Loevinger's (1957, p. 636) suggestion 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, and the construct model was adopted as a general framework for validation (Anastasi, 1986; Embretson, 1983; Guion, 1977; Messick, 1988, 1989). In this vein, the 1985 *Standards* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985) treated validation as an overall evaluation of score interpretations and uses while treating the content, criterion, and construct models as providing different kinds of evidence for the "interpretation of test scores entailed by proposed uses" of the test scores.

However, the version of construct validation that was adopted as the general framework for validity was a mix of the very formal, theory-dependent version outlined by Cronbach and Meehl (1955) and a much looser and more general version of the model, which Cronbach (1989) referred to as the "weak program":

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

Cronbach (1989) recognized that the strong program is "most appropriate in a scientific perspective that reaches centuries into the future . . . " (p. 163), while the weak program can generate a miscellaneous collection of marginally relevant findings and lead to test manuals that hand readers "a do-it-yourself kit of disjoint facts" (Cronbach, 1989, p. 156).

Although the strong program of construct validation had limited applicability, three aspects of Cronbach and Meehl's (1955) formulation developed into widely accepted principles of validation that transcended the positivist, theory-dependent strong program while not embracing the anything-goes aspects of the weak program.

First, test-score interpretations could not generally be taken for granted. The development of construct-related validity evidence requires that the proposed construct interpretation (i.e., the theory) be elaborated in some detail. As a result, the emphasis shifted from the validation of a test, given an interpretation (as a measure of an existing criterion or a content domain), to the development and validation of a proposed interpretation of test scores (Cronbach, 1971, 1982).

Second, validation requires the evaluation of the proposed interpretation in terms of claims implicit in some interpretive framework (e.g., in a defining theory), and this typically involves an extended research program rather than a single empirical study (Cronbach, 1971).

Third, the focus on the evaluation of a construct's defining theory suggested the need to challenge the theory-based interpretation, and the most effective way to challenge a theory is to evaluate how well it predicts observed phenomena and in particular to compare its performance to that of alternate theories (Cronbach, 1971, 1980, 1982; Embretson, 1983; Messick, 1989). The proposed interpretations are to be subjected to critical inquiry.

By the mid-1980s, the model introduced by Cronbach and Meehl (1955) had developed into a general framework for validation based on these three principles: the need for a clear specification of the proposed interpretation, the need for conceptual and empirical evaluation of the proposed interpretation, and the need to consider challenges to the proposed interpretation based on alternate interpretations.

Messick (1988, 1989) adopted a broadly defined version of the construct model as a unifying framework for validation, and defined validity as:

> an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment. [italics in original] (Messick, 1989, p. 13)

The proposed score interpretations and uses are to be evaluated in terms of their overall adequacy and appropriateness, within a general construct framework.

The general construct-based framework was conceptually rich and inclusive, but it did not provide clear guidance for the validation of particular interpretations and uses. In the absence of strong theories, construct validity can be very open ended. For example, Cronbach (1971) had suggested that construct validation be viewed as "an ever-extending inquiry into the processes that produce a high or low test score and into the other effects of those processes" (p. 452), and he subsequently characterized construct validation as "a lengthy, even endless process" (Cronbach, 1989, p. 151). Anastasi (1986) suggested that "almost any information gathered in the process of developing or using a test is relevant to its validity" (p. 3). If all data are potentially relevant to validity, where should one start, and how much evidence is needed to adequately support a proposed interpretation or use?

Argument-Based Approach to Validation

Cronbach (1982, 1988) and House (1980) proposed that the logic of evaluation argument could provide an effective framework for validation, and Cronbach (1988) suggested that a *validity argument* could provide an overall evaluation of the intended interpretations and uses of test scores by examining the evidence for and against the claims being made, including any evidence relevant to plausible alternate interpretations and uses. The analysis "should make clear, and to the extent possible, persuasive, the construction of reality and the value weightings implicit in a test and its application" (Cronbach, 1988, p. 5).

To make validation more manageable, it is useful to have a fairly clear and complete statement of the claims being made so that what it is to be evaluated is clear. One way to do this is to develop an IUA that lays out the reasoning inherent in the proposed interpretations and uses of the scores.[1] The IUA can be specified as a network of inferences and assumptions leading from the test performances to the conclusions to be drawn and to any decisions based on these conclusions (Crooks, Kane, & Cohen, 1996; Kane, 1992, 2006; Shepard, 1993). The IUA provides an explicit statement of what is being claimed and thereby provides a framework for validation.

The argument-based approach was intended to avoid the need for a fully developed, formal theory required by the strong program of construct validity, and at the

same time to avoid the open-endedness and ambiguity of the weak form of construct validity in which any data on any relationship involving the attribute being assessed can be considered grist for the mill (Bachman, 2005; Cronbach, 1988; Haertel, 1999; Kane, 1992). By specifying the claims being made, it provides guidance on the kinds of evidence needed for validation. Once the IUA is developed, it provides a framework for validation and it provides criteria for evaluating whether the proposed interpretation and use have been adequately validated. If the argument is coherent and complete and all of its inferences and assumptions are highly plausible (either *a priori* or because of the evidence provided), the interpretation/use would be considered plausible, or valid. If any part of the argument is not plausible, the interpretation/use would not be considered valid.

Under the argument-based approach, it is not the case that "almost any information gathered in the process of developing or using a test is relevant to its validity" (Anastasi, 1986, p. 3) or that validation is "a lengthy, even endless process" (Cronbach, 1989, p. 151). The evidence needed for validation is that needed to evaluate the inferences and assumptions in the IUA. Validation therefore can be demanding, but it is not more open-ended than other basic scientific inquiries. Validation can proceed systematically by specifying an IUA that adequately represents the proposed interpretation and use and by checking its inferences and assumptions. It is clear where to begin (by specifying the IUA), how to proceed (by evaluating the coherence and completeness of the IUA and the plausibility of its inferences and assumptions), and when to stop (when the inferences and assumptions have been evaluated).

The argument-based approach to validation was developed mainly as a way of facilitating the process of validation. By the 1990s, a general unified model of construct validation had been developed (Cronbach, 1971; Cronbach & Meehl, 1955; Loevinger, 1957; Messick 1989) but the model did not yield a specific strategy for conducting validations. The argument-based approach was designed to retain the generality inherent in the unified model (Messick, 1989) while proposing a more straightforward approach to validation efforts (Bachman, 2005; Bachman & Palmer, 2010; Chapelle, 1999; Chapelle, Enright, & Jamieson, 2008, 2010; Cronbach, 1988; Kane, 1992, 2006; Xi, 2010). The process outlined within the argument-based approach is basically quite simple. First, state the claims that are being made in a proposed interpretation or use (the IUA), and second, evaluate these claims (the validity argument).

An Argument-Based Approach to Validation

The argument-based approach to validity reflects the general principles of construct validity without requiring formal theories. The IUA plays the role that a formal theory plays in the strong program of construct validity. In cases where the attribute being assessed is a theoretical construct, the theory defining the construct would constitute the core of the IUA and the argument-based approach would mimic the approach proposed by Cronbach and Meehl (1955). However, in most cases, the IUA will simply lay out a rationale for whatever claims are being made by the interpretation and use.

The argument-based approach to validation (Kane, 1992, 2006) makes use of two kinds of arguments: a validity argument of the kind suggested by Cronbach (1988),

and an IUA that specifies what is being claimed in the interpretation and use and thereby provides a framework for the *validity argument*, which provides an evaluation of the proposed IUA. To claim that a proposed interpretation or use is valid is to claim that the IUA is clear, coherent, and complete, that its inferences are reasonable, and that its assumptions are plausible.

IUA

The purpose of the IUA is to make the reasoning inherent in proposed interpretations and uses explicit so that it can be better understood and evaluated. The IUA is to specify the proposed interpretations and uses of the scores generated by the testing program as applied to some population over the range of contexts in which it is to be used. The interpretations and uses made of test scores can be quite varied, and the contexts in which the scores are used also can be quite varied; to accommodate this variability, it is necessary to be flexible in the development of the corresponding IUAs. The IUA should reflect the proposed interpretation and use; it should not be constrained to fit some prespecified structure.

In developing the notion of an IUA in this section, I will suggest that some inferences (e.g., scoring, generalization) are likely to occur in most if not all IUAs and that many others are commonly employed, but I do not want to suggest that IUAs have to follow any particular pattern. The inferences discussed in this section and in later sections are intended as examples and not as a checklist. The IUA is to specify the proposed interpretation and use as it is to apply to the populations and contexts in which it will be applied, and it is to do so in enough detail to provide a framework for the evaluation of its most critical and questionable inferences and assumptions.

A typical test requires test takers to perform some tasks (or respond to some stimuli), and the results are used to draw conclusions and make decisions. The process typically begins by scoring the observed performances and combining the task scores in some way, yielding an observed score of some kind. A *scoring inference* takes us from the observed performances to an observed score. It typically makes assumptions about the appropriateness of the scoring criteria and the rules for combining scores (which are based on judgment and/or on statistical criteria).

We usually are not interested in the observed performances and observed scores for their own sake but rather as estimates of some more general attribute and as a basis for decisions. That is, we are not so much interested in making claims about how well the test taker did on a particular set of tasks on a particular day in a particular year; that is a matter of fact. Typically, we want to make claims about how well the test taker can perform in some larger domain of tasks over some range of occasions and conditions of observation, and this involves a *generalization inference* from the observed sample of performances to claims about expected performance in a universe of possible observations (most of which were not made) or to an estimated trait value that can be used to draw conclusions about the future performances. The generalization inference relies on assumptions about the sampling of the observed performances from the larger universe of possible performances that are of interest (e.g., that the sample is random, or representative) or on assumptions about the relationship between trait values and performance in the universe (e.g., based on an

item-response model). The generalization inference may or may not change the numerical value of the score, but it greatly expands its implications.

We could stop at this point and have an interpretation in terms of expected performance over a universe of tasks like those included in the test, but typically we extend the interpretation further. We might be interested in using the scores to predict future performance in some other context (in college or on the job) or to draw inferences about how well the test taker would be likely perform on different kinds of tasks in different contexts. These *extrapolation inferences* extend the interpretation into new performance domains. In some cases, these extrapolations may be fairly informal and rest mainly on experience (e.g., that a student who can solve algebraic equations on an algebra test can also do so in a science class). In other cases, predictions may be based on regression equations or other empirical results. In general, the assumptions supporting the extrapolation inference will involve relationships between the test performances and the performances in some larger domain.

Alternately, a test taker's observed score might be used to draw conclusions about the test taker's standing on some trait or construct that is assumed to explain the observed performances. Such causal inferences extend the interpretation to a trait or construct that is assumed to account for the person's performance, and they generally rely on theories to support the causal claims.

In almost all cases, the test scores are used to make decisions (Bachman, 2002; Bachman & Palmer, 2010). The decision rule might specify, for example, that if a test taker's score is in a certain score range then certain actions will be taken. The decision inference takes us from a person's score to a decision about the person (or about an educational program or teacher). Decision inferences generally rely on assumptions about the consequences of different decisions for individuals with different scores.

These examples of the kinds of inferences that can appear in IUAs are just that: examples. They are not intended as a general template for specifying the arguments or as an exhaustive list of possible inferences, but they do have a place in many interpretations and uses, and they indicate what I mean by an inference.

Note that the inferences take the general form of "if-then" rules. For example, the scoring inference says that if the observed performance has certain characteristics, it gets a certain score. The extrapolation inference could indicate that if the observed score has a particular value, then the criterion variable is expected to have some value (e.g., as indicated by a regression equation). That is, they are rules for making a claim based on an available datum, which may be a previously established claim (e.g., an observed score based on a sample of performance), but they are not formal inferences of the kind found in logic or mathematics. Instead, they are the looser kind of inference found in science, academic discourse, and practical reasoning in which the initial conditions are uncertain and the inferences are contingent and somewhat tentative.

Most of the inferences in IUAs are presumptive in the sense that they can establish a presumption in favor of the conclusion but do not establish it definitively. A presumptive inference can establish presumption in favor of its conclusion (but it is not mechanical or certain), and presumptive inferences can be challenged in particular cases (Blair, 1995; Pinto 2001; Toulmin, 1958, 2001; Walton, 1989). By establishing

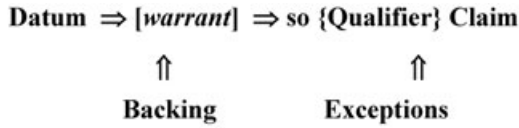

Figure 1. Toulmin's Model of Inference.
Note. Backing provides the evidence for a warrant. Exceptions indicate conditions under which an otherwise sound inference may fail.

a presumption in favor of certain claims, presumptive arguments shift the "burden of proof" onto those who would challenge these claims. Toulmin (1958) proposed a general framework and terminology for analyzing presumptive inferences and suggested that those who make a claim have a responsibility to "make it good and show that it was justifiable" (p. 97).

In Toulmin's model (Figure 1), each inference starts from a *datum* and makes a *claim*. The inference relies on a *warrant*, which is a general rule for inferring claims of a certain kind from data of a certain kind. For example, the warrant for predicting a test taker's performance in some future context from a current test score could be a regression equation. Warrants generally require support, which is referred to as the *backing* for the warrant. The backing for a regression equation typically would consist of an empirical study of the relationship between test scores and some criterion measure of the future performance. Some warrants may be highly plausible *a priori* and therefore not require much backing, but most warrants require some evidence as backing. Warrants that authorize strong claims (e.g., regression equations which make point estimates of future criterion scores or claims about causal explanations) require extensive empirical evidence.

The warrants in presumptive arguments are not infallible and often have a *qualifier* which indicates the strength of the relationship expressed by the warrant. For some inferences, the qualifier may simply qualitatively indicate the likelihood of the claim (e.g., using words like "sometimes," "usually," or "almost always"). Many of the inferences in IUAs for test scores can have well-defined quantitative qualifiers. For example, regression-based estimates of future criterion scores generally are accompanied by standard errors of estimate, and generalization inferences often are qualified by standard errors of measurement that indicate the uncertainty in the generalization.

Finally, a presumptive inference establishes a presumption in favor of the claim but does not prove the claim, and the claim may fail in a particular case because of special circumstances. To accommodate this complication, Toulmin (1958) explicitly allows for exceptions, or *conditions of rebuttal*, indicating conditions under which the warrant would not apply. For example, if a regression equation were based on data for one population (e.g., third graders), its applicability generally would be restricted to that population. In addition to such explicitly stated limits in applicability, every IUA includes a general assumption to the effect that nothing has interfered with the proposed interpretation. For example, an inference from scores on a reading test to conclusions about a student's reading level may hold ordinarily but not apply to a

farsighted student with broken glasses. Similarly, a prediction about some outcome based on past performance may be undermined for a particular student who has a broken leg (Meehl, 1954).

An important class of exceptions in the IUAs for standardized testing programs involves test takers with disabilities. In the accommodations provided for individuals with certain disabilities, the testing materials or procedures are changed and the IUA may have to be adapted to reflect the accommodations. In many cases, the accommodations (e.g., large-type editions for test takers with impaired vision) are designed to make assumptions implicit in the interpretation or use (e.g., that test takers can read the questions) more plausible than they would be otherwise. Such changes require additional assumptions to the effect that the accommodated form of the test is largely equivalent to the original form. The goal is to reach the same kind of conclusions for all students, and the testing accommodations are designed to achieve this goal (Sireci, Scarpati, & Li, 2005).

The IUA is likely to include a number of linked inferences, with the data for some inferences consisting of the claims resulting from former inferences. Although some of these inferences may be purely mathematical or logical, most of the inferences will be presumptive in the sense that their warrants are justified by a preponderance of the evidence, and even the most technical parts of the IUA rely on assumptions about the appropriateness and fit of statistical models. As a result, the IUA is a presumptive argument; if validated, it can generate a strong presumption in favor of the proposed interpretation and use and therefore can justify the proposed interpretation and use of the test scores.

The inferences in the IUA generally will be qualified to various degrees, and therefore the argument as a whole will be qualified. The qualifier to be applied to the IUA as a whole will depend in a complicated way on the qualifiers of the different inferences and on how these inferences fit into the argument as a whole. The qualifiers do not average out, and a serious weakness in any core inference tends to undermine the argument as a whole, even if the other inferences are strongly supported. The inferences can be envisioned as the spans of a bridge leading from the test performances to the conclusions and decisions included in the proposed interpretation and use; if one span falls, the bridge is out, even if the other spans are strongly supported.

The IUA is not an end in itself but is developed to make the evaluation of the proposed interpretation and use as rigorous as possible. It should be stated in enough detail to guide an effective validation, but it does not need to be spelled out exhaustively; that would be deadening. It is particularly important to identify the inferences and assumptions that are most questionable *a priori*.

Although they may not be explicitly mentioned in discussing scores, the warrants for various inferences are an integral part of the IUA and could presumably be supplied if called for. When an inference is drawn (e.g., in scoring the observed performances), the warrant (e.g., the scoring rule) and its backing (e.g., expert judgment and scaling and equating models) are typically relied on implicitly. Similarly, possible exceptions to the proposed interpretation or use are not necessarily mentioned in reporting and interpreting scores unless the exception applies to the case under consideration. However, the main warrants (as they would apply to most test takers)

and their supporting assumptions would be spelled out in the IUA and are evaluated by the validity argument.

The IUA provides an explicit statement of the reasoning inherent in the interpretations and uses of test scores, and it specifies the steps involved in getting from the observed test performances to the claims based on test scores. The generic form of the IUA is applied every time test results are used to draw conclusions or make decisions, and it does not need to be developed anew for each performance (although unusual circumstances can trigger exceptions). The IUA plays the role that a scientific theory or nomological network plays in Cronbach and Meehl (1955) by laying out the claims being made, but it does so in a more general way. It allows for a wide range of possible interpretations ranging from simple claims about expected performance in some domain to the kind of extended theoretical interpretations embodied in a complex theory or nomological network (Marion & Perie, 2009). Perhaps more importantly, it includes an exposition of the intended uses of the test scores.

The Validity Argument

The *validity argument* provides an overall evaluation of the claims in the IUA. The proposed interpretations and uses are valid to the extent that the IUA is complete and coherent, that its inferences are reasonable, and that the assumptions supporting the warrants for these inferences are either inherently plausible or are supported by adequate evidence.

Although they cannot be proven, IUAs can be evaluated in terms of their clarity, coherence, and plausibility. The first step in developing the validity argument is a conceptual analysis of the IUA. The IUA should be coherent in the sense that it provides a plausible rationale for the proposed interpretation and uses and no essential inferences or assumptions are left out. It is particularly important to ensure that questionable inferences or assumptions are recognized and investigated.

The validity argument then can evaluate the warrants in the IUA and the assumptions on which they depend. Some assumptions may be accepted *a priori* or be based on analyses of procedures (e.g., sampling assumptions). Some assumptions (e.g., that time limits are adequate for most students) may be accepted on the basis of experience, but any questionable assumptions will require new empirical evidence to be considered plausible. Strong claims (e.g., causal inferences or predictions of future performance in different contexts) typically would require extensive empirical support. The most questionable assumptions should get the most attention in the validity argument. For highly questionable assumptions, it is useful to consider several parallel lines of evidence.

Different kinds of warrants require different kinds of backing. Scoring rules that take us from observed performances to a score generally rely on expert judgment about the criteria to be used in scoring and on quality control of scoring procedures (and possibly on data regarding rater accuracy and consistency). Generalizations from a sample of observations to expected performance over a universe of possible observations rely on evidence that the sampling was consistent with the statistical model being employed and on generalizability (or reliability) analyses (or item response theory [IRT]–based analyses) indicating that the sample was large enough

to control sampling errors. Extrapolations to different kinds of performance in various contexts rely on empirical evidence (e.g., from a regression analysis) and/or on analyses of the overlap in the skills required for the different kinds of performances in different contexts. Theory-based inferences rely on evidence for the theory and for the appropriateness of the test scores as indicators of constructs in the theory. Score-based decision procedures require evidence that the procedure achieves its goals without unacceptable negative consequences.

Cronbach (1989) proposed four criteria for deciding on the empirical studies to be pursued by the test evaluator:

1. Prior uncertainty: Is the issue genuinely in doubt?
2. Information yield: How much uncertainty will remain at the end of a feasible study?
3. Cost: How expensive is the investigation in time and dollars?
4. Leverage: How critical is the information for achieving consensus in the relevant audience? (p. 165)

Studies of the most questionable assumptions in the IUA are potentially most informative because they address the weakest links in the argument, but all of the inferences merit some attention (Cronbach, 1982, 1988). If the proposed IUA survives serious challenges, its plausibility increases; if one or more inferences fails, then the proposed IUA is undermined.

If it were necessary to support every inference and assumption in the IUA with empirical studies, validation would be a never-ending process because most IUAs involve a number of inferences each of which relies on multiple assumptions, and empirical studies developed to investigate these assumptions often will bring in other assumptions. Fortunately, many inferences and assumptions are sufficiently plausible *a priori* to be accepted without additional evidence, unless there is some reason to doubt them in a particular case.

Like scientific theories, IUAs can be challenged in various ways; one of the most effective ways to challenge an IUA (or theory) is to propose an alternative argument (or theory) that is more plausible given the evidence. The identification and evaluation of plausible competing interpretations is therefore a particularly useful strategy in the evaluation of IUAs.

The desirability of paying most attention to the inferences that are most in doubt has implications for the specification of the IUA. In particular, it is important to specify in some detail those parts of the IUA that are central to the proposed interpretation and use and that are in serious doubt, even if this requires that the more routine parts of the argument are not specified in as much detail. For some interpretations, generalization over some universe of possible observations is the goal, and for these interpretations, the specification of the universe would be a central concern in the IUA. Other interpretations may focus on the prediction of some future outcome, and in these cases generalizability might not be a major concern if the predictions have been shown to be accurate. There is no need to belabor the obvious, and the validation effort should focus on the most questionable parts of the IUA (Crooks et al., 1996).

The argument-based approach is straightforward: state the claims being made and then evaluate the plausibility of these claims. The details will vary from test to test and from context to context depending on the structure and content of the IUA, but the general process is simple. For example, an interpretation of scores on a math test in terms of ability to solve the kinds of problems on the test is quite limited, involves few inferences and assumptions, and therefore would require modest support for its validation; an interpretation of scores on the same test in terms of mathematics aptitude and readiness for an educational program would involve more inferences and assumptions and therefore would require more evidence and more kinds of evidence for its validation. The second of these IUAs includes predictions of future performance in its interpretation of the scores, and the validity argument therefore should evaluate the accuracy of these predictions. If the interpretation had not included such predictions, evaluations of the accuracy of the predictions would be irrelevant. The claims being made vary from one case to another; the evidence required to support these claims therefore will vary, but validation always involves the specification (the IUA) and evaluation (the validity argument) of the proposed interpretations and uses of the scores.

Note that the evaluation of evidence in the validity argument is not symmetric. To make a positive case for the proposed interpretations and uses of scores, the validity argument needs to provide adequate backing for all of the inferences in the IUA and to rule out challenges based on plausible alternative interpretations. However, a refutation of any core warrant can be decisive in undermining an IUA.

Developing the Test, the IUA, and the Validity Argument

Although it is useful to distinguish between the IUA and the validity argument for conceptual reasons (particularly to make the point that the interpretation and use that are to be validated should be clearly defined as a framework for validation), these two arguments tend to be intertwined in practice and are not neatly sequential.

In developing a test, we typically have some purpose in mind, and this purpose guides the development of both the test and the IUA (Bachman & Palmer, 2010). The general intent and context of the testing program provides some sense of the attributes that we want to assess and some of the constraints on the testing procedures. It indicates what we want to achieve with the test scores and provides a rationale for thinking that the scores will serve the purpose. For example, if the purpose is to develop a selection procedure for an educational program, it would be reasonable to begin by identifying some of the core skills and aptitudes associated with success in the program and then to develop a test that seems to measure these attributes. As the test is developed, the IUA can be fleshed out and made more specific, and as the requirements and limitations that need to be satisfied become clearer, the test design may be modified. In developing the test and the IUA, the properties of the scores and the legitimacy of the assumptions in the IUA are investigated, and in doing so, evidence relevant to the validity of the proposed interpretation and use is accumulated.

The development of the test and the validation of a proposed interpretation and use may require an extended effort, and the focus of the inquiry typically changes over time. In the *development stage*, which typically predominates at the beginning of the

process, the goal is to develop (or adopt) a testing program and to develop an IUA that represents the proposed interpretation and use of the scores and is consistent with the characteristics of the test. If any assumptions are found to be untenable, the test, the IUA, or both can be modified to resolve the discrepancies. This iterative process of development and revision continues until the fit between the test and IUA is considered acceptable. Mislevy (2009) uses the term "assessment design argument" to draw attention to the design choices to be made during test development, all of which are influenced by the proposed interpretation and use.

The development stage tends to produce evidence that supports proposed interpretations and uses, because any indication of a flaw in the assessment design or a weakness in the IUA triggers an effort to fix the problem. For example, if the generalizability of the scores is found to be inadequate, the test might be lengthened, thereby improving the generalizability of the scores. The development stage is expected to deliver a clear statement of what is being claimed (e.g., in the form of an explicit IUA linking test performances to the proposed interpretations and uses of the scores).

Once the test and the IUA are developed, the focus shifts (especially for high-stakes applications) and a more critical and arm's length evaluation of the proposed interpretation and use can be adopted. In the *appraisal stage*, the IUA should be challenged, preferably by a neutral or skeptical evaluator. If the validity of the proposed IUA is to be evaluated by the assessment developers, as is often the case, they should seek to identify and examine the challenges that might be posed by a skeptical critic.

The appraisal stage would begin with a critical review of the IUA, with particular attention to identifying any hidden assumptions or alternative plausible interpretations of the test scores, and it also would provide a critical review of the assumptions built into the IUA. The appraisal stage generally would include empirical investigations of the most questionable assumptions in the IUA. As Cronbach (1980) suggested:

> The job of validation is not to support an interpretation, but to find out what might be wrong with it. A proposition deserves some degree of trust only when it has survived serious attempts to falsify it. (p. 103)

In developing the test and IUA during the development stage, it is appropriate to adopt a confirmationist attitude, but at some point it is necessary to shift to a more critical attitude. In the development stage, the focus is on the development of the test and the IUA, but much of the evidence needed for the validity argument will be developed during this stage (e.g., content-related evidence, generalizability analyses, studies of model fit, differential item functioning (DIF) analyses, interrater reliability analyses, think-aloud studies). It is useful to anticipate potential challenges to the IUA during test development so that relevant data can be collected for later use in evaluating the IUA.

In the appraisal stage, the focus is on the development of an adequate validity argument, with the IUA serving as a framework for this validation effort. The coherence and completeness of the IUA for the proposed interpretation and use, as indicated in the test label and description, is evaluated. The inferences included in the

proposed interpretation and use then can be evaluated, employing sources of evidence that address the assumptions supporting the warrants for these inferences.

The requirement that the inferences and assumptions be explicitly stated and evaluated provides some protection against inappropriate interpretations and uses of test scores. To the extent that the IUA is clearly stated, gaps and inconsistencies are harder to ignore and overstated claims can be easier to recognize. A proposed interpretation or use that has undergone a critical appraisal both of its coherence and the plausibility of its inferences and assumptions can be provisionally accepted as being valid, with the understanding that new evidence could lead to a reconsideration of this conclusion.

Fallacies

The specification of an IUA clarifies what is being claimed and a carefully developed validity argument can provide a reasonable basis for accepting or rejecting these claims, but the process is not automatic or algorithmic; it requires judgment, and it can go wrong. Some mistakes in reasoning, or "fallacies," are common enough to have acquired names (Hansen & Pinto, 1995), and IUAs and validity arguments are as prone to going off the rails as any presumptive argument (Kane, 2006).

The *begging-the-question fallacy* is said to occur when some critical inference or assumption in an argument is simply taken for granted, or "begged" (Walton, 1989). Most IUAs depend on a number of inferences and assumptions, and a validity argument that fails to recognize questionable inferences or assumptions "begs" us to accept these inferences and assumptions without serious evaluation. For example, analyses of "content validity" that rely on the relevance and representativeness of test content, the accuracy of scoring, and the reliability of the scores to justify trait or construct interpretations beg a number of serious questions (Messick, 1989). As Cronbach, Gleser, Nanda, and Rajaratnam (1972) pointed out, reliability studies often implicitly define the universe of generalization too narrowly (e.g., over items, but not over item types, occasions, or contexts) given the proposed interpretation and use, and in doing so, "they underestimate the 'error' of measurement, that is, the error of generalization" (p. 352). In some cases, it may be quite reasonable to ignore potential sources of error because they are expected to be small (e.g., the room used for a multiple choice test), or because they are included in the residual error. However, it can be risky to make such assumptions without at least thinking about their plausibility in the case at hand. In the context of validation, the begging-the-question fallacy typically involves the specification of an IUA that does not fully represent the proposed interpretation and use of the test scores.

The *straw man fallacy* goes in the opposite direction and adopts an IUA that is more-ambitious than it needs to be, given the proposed interpretations and uses. The IUA entailed by a proposed interpretation or use is not necessarily obvious, and there is a natural tendency for test developers, test vendors, and test users to want to get as much out of test scores as possible. On the other side, it can be tempting for hostile critics to overstate assumptions inherent in the testing program's claims (e.g., by asking licensure programs to provide predictive validity evidence); more-ambitious IUAs, like "straw men," are easier to knock down than less-ambitious IUAs.

The *reification fallacy* (Hansen & Pinto, 1995) involves an inference from an observed regularity to the existence of some "thing" that is the source of the regularity. For example, it can be reasonable to associate observed consistencies in performance over some kind of task with a trait (e.g., "analytic ability") that accounts for the observed performance patterns, but it is risky to assume that the trait explains or causes the regularities without developing the evidence needed to support such claims or any additional assumptions associated with the "trait" (e.g., that its effect generalizes over contexts or over long periods of time).

Another fallacy, *gilding the lily*, involves the accumulation of additional evidence for claims that already are well established. For example, it can be easy to generate some kinds of evidence which are relevant to the validity of the proposed IUA (e.g., internal-homogeneity reliability estimates, like coefficient alpha), and accumulating such evidence is reassuring and clearly not harmful in itself. The gilding-the-lily fallacy comes into play if the accumulation of a lot of one kind of evidence masks the fact that support for some other part of the IUA is weak.

Scriven (1987) has labeled the "use of a correlate . . . as if it were an explanation of, or a substitute for, or a valid evaluative criterion of, another variable" as the *fallacy of statistical surrogation* (p. 11). The fallacy involves a "substitution of a statistical notion for a concept of a more sophisticated kind such as causation or identity" (Scriven, 1987, p. 11). Instances of this fallacy (e.g., assuming that test scores are the desired outcomes of education rather than indicators of the desired outcomes) can be particularly dangerous in high-stakes contexts, where any indicators that can be manipulated will be manipulated, and as a result the usefulness of the indicator is likely to decline over time. In addition, the manipulations (e.g., an emphasis on test preparation in the context of educational accountability) can have a negative impact on the intended outcomes of a score-based decision program.

The Need for Two Kinds of Arguments

The deployment of two kinds of arguments, IUAs and validity arguments, tends to produce some repetition; the interpretation and use are first laid out in the IUA, and they then are revisited in the validity argument. So why not combine the two and simply evaluate the claims as they are being laid out?

Basically, it is helpful to separate the specification of the claims being made from the evaluation of these claims for a practical reason: The specification of the interpretation and use in an IUA is intended to encourage a careful and relatively complete specification of the proposed interpretation and use, and it is desirable to get a good sense of what is being claimed before getting into the potentially complex task of evaluating these claims.

The fallacies listed above have well-established names because they are common and easy to make. For example, in describing a proposed interpretation or use of test scores it is easy to overlook some implicit assumptions (and thereby beg various questions) or to include unnecessary assumptions (and thereby set up a straw man). I will sketch two examples.

Generalizations. In interpreting test scores, we typically generalize over the tasks included in the test or over test forms. That is, we do not typically make statements

like, "Mary did well on form 26B of the reading test:" rather we are likely to make statements like, "Mary did well on the reading test." That is, we assume that the particular test form used does not matter much and that the results would generalize over forms (i.e., the score would not vary much over forms). Variability over tasks or items, and by extension over forms, is typically evaluated using measures of internal homogeneity, like coefficient alpha.

For achievement tests and trait measures, we also typically generalize over settings and time of day. In using the simple statement, "Mary did well on the reading test," we are implicitly generalizing over a host of factors (including testing settings, time of day, day of the week, test administrator, etc.), any of which could possibly have an impact on Mary's score. If the interpretation is not to be generalized over possible testing dates and rather is tied to a certain date (e.g., as posttest scores in an experiment might be), it generally would be important to mention the testing date in the report (e.g., "Mary did well on the posttest given on the last day of the program").

If the interpretation generalizes over a number of conditions of observation (e.g., settings, time of day, day of the week, test administrator), the interpretation assumes that scores do not vary too much over variations in these conditions and it would be prudent to include these assumptions in the IUA and to consider their plausibility. For many testing programs and many kinds of conditions, these assumptions may be quite plausible *a priori*; for example, if a student can solve linear equations in a math class on one day, experience suggests that the student probably will be able to solve similar equations a few days later. But in other cases, the assumptions may not be so plausible *a priori*.

In developing the IUA, it would be worthwhile to indicate how widely the interpretation is to be generalized. 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.

Licensure testing programs. Licensure tests are designed to provide some protection to the public by requiring candidates to demonstrate some level of competence in some domain of knowledge, skills, and judgment (KSJs) that are needed in practice (Clauser, Margolis, & Case, 2006). The intent is to exclude individuals who lack the KSJs to a substantial extent from practice, and thereby to protect the public and to improve the overall level of practice. The focus is on a limited set of critical KSJs (mainly cognitive skills), and the goal is to improve the safety and quality of practice by excluding candidates who lack the KSJs to a substantial extent.

It is necessary to assume that the KSJs are needed in practice, that a serious deficiency in a practitioner's mastery of the KSJs would pose a threat to the public, and that a low score on the test (one below the passing score) indicates a serious deficiency in the KSJs. It is not necessary to assume that those who pass the test will be good practitioners; they may lack other essential characteristics (e.g., conscientiousness, interpersonal skills). It also is not necessary to assume that the scores of passing candidates will be positively correlated with their performance in practice. Licensure programs are not designed to rank order candidates from best to worse, and for the passing candidates, they are not designed to predict individual performance in practice.

Nevertheless, it is easy to assume that licensure test scores should be positively correlated with performance in practice and that they should be validated in terms of how well they predict performance in practice. It would be nice if this were the case, and it would be useful for some purposes (e.g., in subsequent employment decisions).

However, the claims made in the use of test scores in licensure programs do not entail an assumption that the scores be positively correlated with performance in practice, and imposing such a requirement for validation is an example of the straw man fallacy. It is an unnecessary requirement that would be very difficult, if not impossible, to meet in practice (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Clauser et al., 2006; Kane, 1986).

Tailoring Validation to Proposed Interpretations

In this section, I will make three basic points about test-score interpretations and uses and their validation:

1. Test scores can have multiple possible interpretations/uses, and it is the proposed interpretation/use that is validated, not the test itself or the test scores.
2. The validity of a proposed interpretation/use depends on how well the evidence supports the proposed interpretation/use of the test scores.
3. More-ambitious interpretations and uses require more backing (i.e., evidence) than less-ambitious interpretations and uses.

To make the discussion of these points simpler and more concrete, I will frame it in terms of attributes that are defined as tendencies to perform or behave in some way (e.g., a tendency to act aggressively in some contexts or to solve algebraic equations correctly) under some circumstances. I will refer to these attributes as "observable attributes;" they are "observable" in the sense that they are defined in terms of observable performances. Such interpretations make no strong claims about any underlying explanatory traits or constructs. They do generally make a weak claim to the effect that the performances of interest reflect some characteristic or cluster of characteristics of the test taker, but they do not make any claims about what these underlying characteristics are or how they function. In particular, they do not generally assume that the observed performances are explained by a single, underlying trait.

Observable attributes can be useful in many contexts. For example, in certification and employment contexts, a test of skill in servicing computers which required candidates to diagnose and correct various kinds of problems might be quite useful without making assumptions about an underlying "computer-servicing" trait, and a test requiring students to read and interpret passages in a language can provide an indication of level of literacy in that language without developing and evaluating a theory of literacy. In both of these cases, performance is likely to depend on knowledge (e.g., of computer architecture, vocabulary), specific skills (e.g., in diagnosing hardware and software problems, in parsing sentences and paragraphs), and more general cognitive strategies (identifying problems by systematically ruling out competing possibilities, and using structural characteristics of text to help in extracting meaning). There is no need to assume that the performances are associated with a

single trait or with a specific theory that accounts for, explains, or causes differences in performance.

Observable Attributes

An *observable attribute* is defined in terms of how well a test taker will perform on average over a target domain of possible observations. The *target domain* specifies the kinds of tasks and the ranges of contexts and conditions of observation associated with the observable attribute. The observable attribute summarizes how well people can perform some kind of task or activity or how they respond to some kind of situation or stimulus over some range of conditions. The target domains for observable attributes typically include a variety of "real-world" tasks and performances in non-test contexts; they are not restricted to test tasks or test-like tasks in standardized testing situations. The value assigned to an observable attribute for a person is basically a claim about what the person can do or about how the person is likely to react to some kind of stimulus or situation.

Target domains are defined in terms of the kinds of tasks or stimuli that they include and in terms of conditions of observations, contexts, and occasions that are allowed. Some observable attributes are defined broadly (e.g., literacy or proficiency in algebra), and some are defined more narrowly (e.g., vocabulary knowledge or skill in solving a particular kind of algebra problem). They can be limited to a specific context (e.g., a particular workplace or academic setting) or include a wide range of contexts (e.g., wherever a skill might be exercised). They can be associated with specific occasions (e.g., moods, which can change from moment to moment) or include all occasions over some extended period (e.g., aptitudes or skills, which are assumed to be stable over extended periods).

The range of observations to be included in the target domain may be defined in terms of the goals of an instructional program or of a specific lesson, in terms of the requirements of a job, or in terms of the skills thought to be required by the tasks in some domain. For example, an employer who uses the scores as indications of the ability to apply basic arithmetic in the work setting (e.g., in making change, measuring lengths and volumes, figuring discounts) would adopt a target domain covering the workplace performances that are considered particularly important.

In most cases, observable attributes are assumed to reflect some trait or skill or collection of skills of the test taker, and the performances included in the target domain are intended to reflect or require these skills. However, once the domain is specified, it defines the observable attribute. General assumptions about underlying traits or abilities do not need to be included in the validation of an interpretation of the test scores in terms of the observable attribute unless these assumptions play a substantial role in the interpretation or use of the scores. It is recognized that theories of some kind may be able to explain the performances covered by the observable attributes (now or in the future), but such theories are not necessary for the interpretation and use of test scores as measures of observable attributes.

The observable attributes that are of most interest in education and other applied contexts (e.g., literacy, proficiency in algebra) tend to be quite general, and they include target domains that are broadly defined. For example, when we think of a

student's achievement in arithmetic, we are likely to have in mind his or her ability to solve a range of problems in a variety of contexts (e.g., in math and science classes and in various activities in work and life). Similarly, literacy as an observable attribute can be defined as the ability to perform a variety of tasks in a variety of contexts ranging from the reading of simple street signs to the reading of a newspaper editorial or a book in a library or on a bus.

The target domain specifies the proposed interpretation of the observable attribute. As indicated later in this section, we may embellish this meaning in various ways, but the observable attribute is defined by its target domain and the usage of the label assigned to an observable attribute also is determined by its target domain (Kane, 2011). For example, for some observable attributes (e.g., literacy) the target domain would include a broad range of tasks, contexts, and occasions and a test taker's standing on such attributes can be discussed without mentioning specific tasks, situations, or occasions. For more narrowly defined attributes, the target domain might involve specific kinds of tasks (e.g., map reading), situations (e.g., context-specific anxieties), or occasions (e.g., moods), and claims about such attributes would need to reflect the appropriate restrictions.

Assessments of Observable Attributes

For practical reasons, generally it is not possible for testing programs to employ random or representative samples from the target domain (but some exceptions are discussed later). For example, the target domain of performances for literacy probably would include the reading of a newspaper or a novel, but it generally would not be feasible to include either of these tasks in a test of literacy. The test might include a task involving the reading of part of a newspaper or part of a novel, but the more general activities are too open ended and too time consuming. The test tasks and testing conditions are standardized in various ways, and the observations included in the test are drawn from a part of the target domain (Kane, 1982).

Using the terminology of generalizability theory, the domain from which the observations are actually sampled for a test will be called the *universe of generalization* for the testing procedure, and a person's expected score over the universe of generalization will be called the person's *universe score* (Brennan, 2001b; Cronbach et al., 1972; Shavelson & Webb, 1991). As a result of standardization, the universe of generalization for a test is a subset of the target domain for the observable attribute, and inferences from the observed score to the target score are not simple generalizations from a sample to the domain from which the sample is drawn.

It can be reasonable to take the performances included in standardized tests as a random or representative sample from the universe of generalization for a test, but it would be much less reasonable to take these observations as a random or representative sample from the target domain for an observable attribute and it is not legitimate simply to generalize from the observed score to the target score. Instead, the interpretation of observed scores on standardized tests in terms of expected performance over the target domain requires at least three main inferences: a scoring inference, a generalization from the observed score to the universe score, and an extrapolation from the universe score to the target score.

The target domain is meant to reflect the attribute of interest (e.g., a competence or a disposition to behave in some way), and it is not restricted to test items or test-like tasks. The range of observations included in the target domain should be consistent with the label assigned to the observable attribute. If the observable attribute is given an expansive interpretation (explicitly or implicitly in a label like "English literacy"), the target domain would include a wide range of tasks in a wide range of settings. If the observable attribute is given a narrow description (e.g., skill in solving quadratic equations), the domain would be fairly limited.

Performance tests and operational definitions. Although generally it is not possible to draw random or representative samples from target domains, it is possible to do so in at least two cases: for operationally defined attributes and for performance tests.

An operational interpretation defines the attribute in terms of a measurement or testing procedure (Bridgeman, 1927), and as a result the target domain by definition is the same as the universe of generalization. The operational interpretation generalizes over replications of the testing procedure (i.e., over the kinds of tasks included in the test forms, over testing contexts, and over some range of occasions, all of which are consistent with the definition of the testing program), but it does not extrapolate to any broader domain that includes real-world performances in real-world contexts. For operationally defined attributes, the target domain is taken to be the same as the universe of generalization.

Operational interpretations tend to be narrowly defined. This is not necessarily a problem, because narrowly defined attributes can have important uses; as discussed more fully later, operationally defined attributes are used as indicators of theoretical constructs. More generally, it is often useful to be able to analyze and talk about the properties of a testing procedure as such, and operationally defined attributes facilitate such discussion.

Testing procedures are defined operationally; they need to specify how the data on test-taker performance is to be collected and the variations (in terms of tasks, contexts, occasions, and other conditions of observation) that are allowed. Operational interpretations become problematic when the operational interpretation is used in validating the tests scores, and then a more-ambitious interpretation (e.g., in terms of a much broader target domain or a construct) is adopted in practice to draw conclusions or make decisions (Ennis, 1973; Feest, 2005). For example, the scores on a particular test might be taken as defining "intelligence." This kind of implicit slide from one interpretation to a more-ambitious interpretation is an example of the begging-the-question fallacy.

In designing performance tests, we seek to draw representative samples from the target domain (Kane et al., 1999; Lane & Stone, 2006). For example, skill in playing the piano would be assessed by having the test takers play a number of pieces before a panel of judges; if the sample is large enough and varied enough, it could be quite reasonable to generalize from the observed performances to the target domain for "skill in playing the piano." For performance tests, the universe of generalization is designed to cover much if not all of the target domain.

The analysis of performance tests is similar to that for operational definitions in that the universe of generalization and the target domain are fused into a single

universe/domain. They are distinct in how this is achieved, however. For operationally defined attributes, the test is taken as a given and the attribute is defined in terms of scores on the test. For performance tests, the attribute is taken as a given and the test is designed to cover the target domain for the attribute (Mislevy, 2006).

Test Scores, Interpretations, and Uses

Note that the testing procedures and the test content do not in themselves lock in the interpretation or use of the scores. Starting with a given set of test performances, we can interpret the resulting score as an estimate of expected performance over replications of the testing procedure (i.e., the operational interpretation) or as an estimate of the expected performance over some larger target domain (i.e., as an observable attribute), and we can consider different target domains, some broader than others. The target domain will depend on the proposed interpretation and use of the scores and can include performances on a wide range of tasks and in a wide range of contexts. The scores derived from a particular sample of performances can be given different interpretations and can have different uses.

In general, it is not possible to evaluate the validity of test scores without adopting, explicitly or implicitly, some proposed interpretation or use. If someone showed us an unlabeled test booklet and a set of test administration guidelines and asked us to validate the test, what would we do? The first thing I would do would be to ask how the test scores are to be interpreted and used, the population from which the test takers will come, and the contexts in which all this is to happen. With a proposed interpretation and use spelled out, the claims being made can be evaluated.

Validating Interpretations in Terms of Observable Attributes

The IUA for an observable attribute typically includes at least three main inferences: scoring, generalization, and some kind of extrapolation to nontest performances or competencies. Scoring evaluates the observed performance, which yields an observed score (a raw score or scaled score of some kind), the observed score is generalized to the universe score, and the universe score is extrapolated to a target score.

Scoring. The warrant for the scoring inference is a scoring rule or rubric that assigns scores to a test taker's observed performance. Using Toulmin's model, the performances are the data and the score is the claim. The scoring inference depends on a number of assumptions: that the scoring procedures are appropriate, are applied as intended, and are free of overt bias. The backing for the scoring inference typically involves both the judgment of panels of experts who develop and review the scoring criteria and evidence that the scoring procedures are implemented consistently and correctly (Clauser, 2000). The scoring rules and most of their backing typically are developed as the test is developed.

Generalization. The generalization inference treats the observed score as an estimate of the universe score (the mean over the universe of generalization for the test taker). The value of the score does not necessarily change, but the interpretation is extended from a claim about observed performances to a claim about expected

performance over the universe of generalization. The generalization warrant takes the observed score as the datum and the universe score as the claim.

A test taker's scores will vary over replications drawn from the universe of generalization, and the precision of the generalization from the observed score to the universe score is limited by this sampling variability which is evaluated in reliability studies (Feldt & Brennan, 1989; Haertel, 2006), generalizability studies (Brennan, 2001a, 2001b; Cronbach et al., 1972; Kane, 1996) or in terms of IRT-based information functions (Yen & Fitzpatrick, 2006). Universes of generalization include observations that can vary in a number of ways, involving, for example, samples of tasks, testing contexts, occasions in which the test is administered, and possibly raters who score the responses. In generalizability theory, these different kinds of conditions of observation are referred to as *facets*. For example, we can have an item facet and an occasion facet and various items would be conditions of the item facet, and various occasions would be conditions of the occasion facet. A particular observation from the universe of generalization would involve some condition of each facet (e.g., a particular item, occasion, context, rater).

Statistical sampling theory provides the framework for analyses of generalization inferences, and it depends on assumptions about the sampling of performances (e.g., random sampling) from the universe of generalization, but these sampling assumptions are hardly ever realized in practice; test tasks and testing contexts are created by the test developers, raters are recruited and trained, and testing occasions are dictated by scheduling concerns (Brennan, 2001b; Kane, 2002a; Loevinger, 1957). So a case has to be made that the sample, which is not random, is representative enough in some sense that the generalization inference will work as intended, even though, strictly speaking, the sample of observations is not random. Typically, the representativeness of the sample is supported by two kinds of evidence. First, an effort is made to make the sampling of tasks and conditions of observation as representative of the universe of generalization as possible. Second, an effort is made to identify and eliminate any effects that would make the sample substantially unrepresentative of the universe of generalization and therefore statistically biased. For example, if the test is to be used in an international assessment of world geography, it would be inappropriate to have all of the tasks be specific to North America; such a skewed sample would be biased in the statistical sense and in the more general sense of putting students from other parts of the world at a disadvantage. This example is sufficiently gross that it would probably not occur, but it takes a serious effort by test developers to avoid more subtle potential sources of bias (e.g., in terms of relative emphasis on physical, cultural, and political aspects of geography). If test developers have made a serious effort to make the sampling representative of the universe of generalization (i.e., by defining the universe clearly and systematically sampling from it, as in stratified samples) and no serious source of sampling bias can be identified, it would be reasonable to assume that the sample is representative (Kane, 2002a).

Statistical sampling theory also provides a qualifier for the generalization inference in the form of an estimate of the sampling error associated with this generalization inference. The qualification or uncertainty in the generalization is determined by the sampling error in generalizing from the observed score on the test to the

expected value over the universe of generalization, as estimated in reliability studies, generalizability studies, or through IRT information functions.

The sampling error generally involves multiple sources of error associated with sampling from different facets (e.g., tasks, contexts, occasions, raters). The fact that it is the squares of the errors rather than the errors themselves that are additive is significant, because it implies that relatively small random errors can be safely ignored (Kane, 2010, 2011). For example, assume that we have two sources of error and that the larger error is about 10 times as large as the smaller error, say 10 and 1. The total error would be the square root of $(10)^2 + (1)^2 = 101$, or 10.05. The addition of a second error one-tenth as large as the first adds only one-half of 1% to the initial error. Given that the overall error does not need to be known with great precision and generally is not known with much precision, we need to be concerned only with the largest sources of error (in particular, with the largest source of error and with other sources of error of the same order of magnitude).

We have at least two options if the sampling errors associated with a facet are large. First, during the development phase we can modify the testing procedure so that sampling variability is reduced (e.g., by increasing sample sizes for the facet). Second, we can modify the definition of the universe of generalization so that it does not involve generalization over the facet in question. For example, if the scores are highly variable across different testing conditions, the testing procedure (i.e., the universe of generalization) can be tightened up (i.e., standardized), or if scores are variable across raters, the selection and training of raters may be tightened and the scoring rules may be made more specific and less judgmental. Such efforts to tighten the definition of the universe of generalization can improve the generalizability of the scores across replications of the procedure, but they also narrow the universe of generalization (Kane, 1982).

IRT addresses questions about the generalizability of scores across replications of the testing procedures using information functions and standard errors based on these information functions (Yen & Fitzpatrick, 2006). The standard errors for estimates of the values of a latent trait depend on assumptions built into the model, the characteristics of the test tasks, and the number of tasks. The estimated IRT standard error provides an indication of the expected variability in scores over hypothetical replications of the testing procedure using a specific set of test tasks. Or, assuming that the task scores fit the model, it provides an indication of the expected variability of replications with different sets of tasks. The IRT standard errors do not generally reflect variability over occasions, contexts, or any other facets. This is not a limitation in the IRT models but rather is a limitation in the data. With results from a single testing session, it is possible to evaluate internal homogeneity (comparable to coefficient alpha) but it is not possible to evaluate variability over facets like occasions or contexts.

The estimates of standard errors of measurement derived from generalizability or reliability studies or IRT analyses put limits on the precision of estimates of the universe score (Brennan, 2001a, 2001b), and confidence intervals derived from the standard errors can provide quantitative qualifiers on the generalization inference. Large standard errors and broad confidence intervals imply weak conclusions about the universe score.

Extrapolation to the target domain. The extrapolation inference extends the interpretation from the universe of generalization to the target domain. The score does not necessarily change, but its interpretation is broadened to include "real-world" performances. The extrapolation inference does not involve a simple statistical generalization, but rather a more-ambitious leap from claims about test performances to claims about the full range of performance in the target domain, including nontest performances in nontest contexts (Bachman & Palmer, 2010; Haertel, 1999; Kane, 2006).

The backing for the extrapolation warrant can be based on two kinds of evidence: analytic and empirical. The analytic evidence relies on analyses of the relationship between the universe of generalization and the target domain and tends to be generated during test development. The empirical evidence examines relationships between observed scores on the test and other scores based on observations from other parts of the target domain.

Our confidence in the extrapolation inference depends, at least in part, on the relationship between the test design (i.e., how items are created and selected and how the test is administered and scored) and the definition of the target domain for the observable attribute. To the extent that the test covers most of the target domain, the extrapolation is likely to be considered plausible and the extrapolation inference might not require much empirical support. On the other hand, if the target domain is very broadly conceived (e.g., literacy) and the test is highly standardized, the extrapolation inference would be a major concern in validation. For tests that employ task presentations or response modes that are substantially different from those in the target domain (e.g., the uses of objective questions in cases where the performances of interest are extended), the possibility that irrelevant method variance will limit the validity of the proposed interpretation is a natural concern (Messick, 1989, 1994).

If the processes involved in responding to the test tasks and to other tasks in the target domain are understood to some extent and if the test tasks seem to involve the same processes as most tasks in the target domain, extrapolation is likely to seem reasonable. The processes (e.g., those involved in reading English prose or arithmetic problems) do not have to be understood in any detail for these analyses to be plausible. In practice, the analyses supporting the extrapolation inference often are based mainly on experience; for example, students who can solve arithmetic problems on a test generally can also do so in other contexts. "Think-aloud" protocols in which test takers report how they are addressing test tasks can provide more detailed analyses of the processes being applied to test tasks (Bonner, 2005; Cronbach, 1971). If the processes applied to test tasks are similar to those applied to other tasks in the target domain, extrapolation is supported. If the processes are substantially different, extrapolation gets less analytic support and would require more empirical support.

To the extent that the test performances are restricted to a small subset of the target domain for the observable attribute, the extrapolation inference might rely on backing in the form of empirical studies of the relationship between test scores and scores based on more thorough and representative assessments of overall performance in the target domain. In some cases, it may be possible to obtain more representative samples from the target domain than those provided by the test; if so, the test scores can be compared to scores on this criterion measure. For example, a test of proficiency

in a foreign language might consist of short printed passages in the foreign language and taped conversations in the foreign language followed by sets of questions about each passage or conversation, while the criterion assessment might include actual one-on-one conversations and group discussions based on texts, all in the foreign language. The criterion measure might not be feasible in a testing program but would be feasible in a validity study with a modest sample size.

Reliability as a Necessary Condition for Validity?

In classical test theory, and in subsequent psychometric models, reliability is taken to be a necessary but not sufficient condition for validity. The discussion presented above indicates why this assumption generally makes sense, but it also indicates some caveats to be considered in making this assumption (Mislevy, 2004; Moss, 1994).

The general rule that reliability is a necessary but not sufficient condition for validity can be explicitly derived in the context of criterion-related validation studies, for which "the validity of a test with respect to any criterion cannot exceed the index of reliability" (Lord & Novick, 1968, p. 72). More generally, one can make a case that if test scores vary substantially over replications of the testing procedure, it is hard to give the scores any consistent interpretation.

Under the argument-based approach to validation, generalizability is a necessary condition for the validity of almost all test score interpretations because almost all of these interpretations involve generalization over some universe of generalization that goes beyond the observations actually made. Assuming that we intend to generalize over some range of occasions, some level of generalizability over occasions is necessary, and assuming that we are going to generalize over some universe of tasks, some level of generalizability over tasks is necessary (Brennan, 2001a, 2001b). In the context of classical discussions of reliability, generalizability over occasions is evaluated in terms of "stability coefficients" and generalizability over tasks is evaluated in terms of "parallel forms reliability" or measures of internal homogeneity (e.g., coefficient alpha).

Under the argument-based approach, generalizability/reliability is not a sufficient condition for the validity of most test score interpretations because these interpretations involve additional inferences (e.g., extrapolation) that are not addressed by reliability or generalizability evidence; to establish the validity of an interpretation, we need adequate backing for all of the inferences in the interpretation.

However, some interpretations/uses do not assume generalization over some facets, and therefore their validation does not require generalizability over these facets. For example, "state" variables (e.g., moods, state anxiety) are expected to vary from moment to moment and do not assume generalization over any extended period. A performance test may focus on a particular task or kind of task and not assume generalization over tasks or kinds of tasks. In extreme cases, the focus may be on specific performances in specific contexts and there is no need to generalize over any facets. Moss (1994) provides several examples of score uses in which little or no generalization is assumed, but for almost all IUAs some generalization is assumed and some evidence for generalizability is appropriate.

Standardization and the Reliability/Validity Paradox

Standardization of the test format, content, and procedures creates a universe of generalization that is distinct from (usually narrower than) the target domain. For example, we might standardize the task format in a test of literacy to objective questions following short passages. This kind of task is a legitimate example of what we mean by literacy, but the resulting universe of generalization is a narrow slice of the target domain for literacy.

If we based the test on a representative sample from the target domain for literacy, we would have a performance test (Kane et al., 1999) and we would need to generalize from the observed score to the universe score, but we would not need to extrapolate from the universe score to a distinct target score because the universe of generalization would be the target domain. This kind of performance assessment tends to be cumbersome and expensive (if not impossible) for broadly defined observable attributes like literacy. In addition, because the tasks and conditions of observation tend to be so variable in the target domain, generalizability from the observed score to the universe/target score is likely to be poor. So the tests for almost all observable attributes employ standardized task and response formats and standardized conditions of observation, and the universes of generalization are narrowly defined subsets of the target domain.

Standardization has two main advantages. It promotes fairness by ensuring that all test takers are asked to perform more or less the same tasks under the same conditions, and it tends to improve generalization from the observed score to the universe score by narrowing the universe of generalization. If we standardize the test by fixing the task format to a particular kind of question, we eliminate any random error that might have resulted from the sampling of task formats across replications of the testing procedure. Furthermore, we can concentrate all of the test tasks in this smaller universe of generalization and therefore get a fairly large sample of tasks from this relatively homogeneous universe of generalization. Similarly, by standardizing the testing conditions (e.g., settings, time limits, instructions), we narrow the universe of generalization and enhance generalizability.

However, standardization of any aspect of the testing procedure that is not also fixed in the target domain introduces a source of systematic error (Kane, 1982). If performance does vary across task formats and we fix the task format in the test, then this difference between the universe of generalization and the target domain will introduce systematic error into estimates of the target score. Furthermore, these systematic errors will not wash out as the sample of observations gets larger as is the case for random errors. The errors introduced by standardization constitute systematic errors associated with the underrepresentation of the target domain in the universe of generalization.

Standardization tends to improve generalization from the observed score to the universe scores, and it tends to interfere with extrapolation from the universe score to the target score. Using generalizability theory models, it can be shown that, depending on how standardization is implemented, it can increase the accuracy of inferences from the observed score to the domain score, it can leave the accuracy unchanged, or it can decrease the accuracy of these inferences (Brennan, 2001a, 2001b; Kane,

1982); basically, standardization can increase the expected correlation of the observed score with the universe score, but it tends to decrease the expected correlation of the universe score with the target score.

The goal in designing testing procedures is to standardize in ways that control random error effectively while introducing as little systematic error as possible. For example, we can standardize the time limit for a test to a particular value, thus eliminating the random error (and a potentially significant threat to fairness) that would be introduced if different test takers had different time limits, and we can limit the potential for large systematic errors by making the time limit long enough that it is sufficient for all (or almost all) test takers.

Some Additional Claims that Can be Attached to Observable Attributes

As indicated above, the IUAs for observable attributes generally involve three basic inferences (scoring, generalization, and extrapolation), but some additional inferences also are common in the IUAs for these attributes.

Inferences about different variables. In some cases, the test performances may not be sampled from a subset of the target domain but rather from a universe of generalization that is distinct from the target domain. We might, for instance, want to use test scores obtained at one time and in one context to predict performance on a different variable, in a different context, and at some time in the future.

The warrant for this predictive inference typically would be a regression equation that would yield an expected score on a criterion variable for the target domain based on an observed score on the test. The backing for this warrant typically would be drawn from a "predictive validity" study in which the relationship between test scores and the variable of interest is investigated for some sample of individuals. The standard error of estimate for the regression equation provides a qualifier for this predictive warrant.

Scale-based inferences. Various scaling models can be used to transform observed scores to scaled scores that add to the interpretation in some way (Kolen, 2006; Petersen, 2007). The scaling models add warrants to the IUA which allow for additional inferences from the scores and which need to be evaluated as part of the validity argument. In some cases, the scaled scores may be reported as such, and in other cases, they may be reported separately in the form of interpretive guidelines for the reported scores (e.g., as tables or charts of norm-referenced percentiles). Generally, the scaling models add to the interpretation of the scores by providing additional information about what the scores mean.

For example, observed scores may be given a norm-referenced interpretation in terms of the percentages of some population (e.g., third graders in U.S. public schools, freshman majoring in engineering at a particular college) that have scores at or below each observed score. The warrant for this inference would associate each test score with the corresponding percentage in the relevant population, and the backing for this warrant would be derived from data on the distribution of scores in the population. For the norm-referenced interpretation to be most useful, it is necessary

that the score distributions be based on appropriate sampling plans, that the samples be large, and that the data be fairly recent.

Criterion-referenced interpretations of various kinds also can add to the IUA by indicating what test takers with various observed scores can be expected to be able to do. In some cases, these criterion-referenced interpretations may be specified at a very general level (e.g., in terms of whether test takers with scores in various score ranges can be considered advanced, proficient, basic, or below basic in a content area, according to some criteria). In other cases, test takers with scores in various score ranges can be characterized in terms of specific tasks they are expected to be able to perform (e.g., claims that individuals with scores at a particular point on the score scale would have a certain probability of answering certain items correctly) or in terms of learning progressions indicating that test takers with certain scores can perform at some level in a progression of increasing skill levels (Kolen, 2006).

Item-response functions represent the probability that a test taker will get an item right as a function both of the test taker's standing on a latent ability scale and certain item characteristics (or parameters). To the extent that the model fits the data and the item parameters are known, we can predict how a person with a particular ability level will respond to any item or any set of items that fit the model. The IRT model provides a more detailed, item-specific analysis of the relationship between an estimated ability level (based on performance on some items) and expected performance on other items in the universe of generalization for a testing procedure.

> Fitting an IRT model is an empirical exercise, capturing and quantifying the patterns that some people tend to answer more items correctly than others, and some items tend to be answered correctly less often than others. The conception of . . . competence embodied by the IRT model is simply the tendency to perform well in the domain of tasks. (Mislevy, 1996, p. 393)

That is, the IRT model summarizes the relationships between test taker ability, as indicated by performance on a sample of items, and expected performance on other items that also fit the model and for which item parameters are known (Zumbo, 2007).

Simple generalization relies on sampling models to warrant generalization to a universe score; IRT models warrant inferences from item performances to an estimated ability level on the latent trait scale. In both cases, the warrant justifies an inference from the observed item scores to expected performance on the items in some universe of possible items (those in the universe of generalization for simple generalizations, or those that fit the IRT model).

Trait hypotheses. As noted at the beginning of this section, target domains are not defined arbitrarily but may reflect the content of instruction, the skills assumed to be required in some context (e.g., a job or educational program), or some assumptions about underlying traits. The observable attribute reflects a general disposition of test takers to perform or behave in some way on some kinds of tasks under some conditions. Individuals who perform in a certain way on the tasks get the corresponding scores; for example, students who can generally solve algebra problems get high scores in algebra and those who cannot solve most of the problems get low scores.

We may think of the test scores as reflecting standing on some trait that accounts for performance on the test, but we do not necessarily make any assumptions about the nature of the trait or how it works or about any implications of trait values other than those specified by the target domain. Test performance may be due to a trait or it may be accounted for by a cluster of specific knowledge elements, skills, and habits, which are called upon in various combinations by the test tasks.

There is no *a priori* reason to assume that the performances summarized by an observable attribute are accounted for, or due to, a single, unidimensional characteristic of the test takers. As Lord and Novick (1968) put it:

> Much of psychological theory is based on trait orientation, but nowhere is there any necessary implication that traits exist in any physical or physiological sense. It is sufficient that a person behave as if he were in possession of a certain amount of each of a number of relevant traits and that he behave as if these amounts substantially determined his behavior. (p. 358)

My use of the term "trait" is consistent with Lord and Novick's usage. The trait does not account for or explain the observed performances; rather, the target domain serves as a definition for the observable attribute, or "trait." The use of trait language can be harmless if the trait label is used as shorthand for the observable attributes or as an expression of belief that the observed regularities in performance in the target domain are due to some characteristics of the person, even though we do not know much about this characteristic. However, the adoption of trait language can also be misleading if the trait is taken as an explanation for observed regularities in performance without specifying the causal mechanisms or providing evidence for the claims being made (Meehl, 1950). Unfortunately, it is easy and somewhat natural to assign trait interpretations to observable attributes by assuming the existence of an underlying trait or construct that accounts for the observed regularities in performance (Loevinger, 1957). For example, Boorstin (1983) quotes Galen, a second century physician and philosopher, on the tendency to do so in medicine:

> "So long as we are ignorant of the true essence of the cause which is operating," Galen explained, "we call it a faculty. Thus we say that there exists in the veins a blood-making faculty, as also a digestive faculty in the stomach, a pulsative faculty in the heart, and in each of the other parts a special faculty corresponding to the function or activity of that part." (p. 361)

In a lighter vein, in *The Imaginary Invalid,* Molière made fun of the physicians who "explained" the sleep-inducing properties of opium in terms of a "virtus dormitiva," or "dormative principle." In these cases, the effect is attributed to a "cause" that is essentially a more general label for the effect itself. Again, this kind of vague causal claim that some otherwise unspecified trait accounts for observed consistencies in performance is not necessarily misleading as long as it is not taken too seriously, but if it is taken seriously as an adequate causal explanation it can close off the search for more informative explanations.

It clearly is desirable to develop explanations for observed performances and for relationships among observable attributes (Zumbo, 2007, 2009), but we do not have

to wait for full explanations of observed regularities in performance before making use of these regularities (e.g., in the form of observable attributes).

The Evidence Needed for Validation

The validation of a proposed interpretation begins with an evaluation of the coherence and completeness of the IUA followed by an evaluation of the plausibility of its inferences and assumptions. The different kinds of observable attributes considered in this section involve different sets of claims and therefore they require different mixes of evidence for their validation.

The argument-based approach to validation does not specify any particular kind of interpretation for scores, but it does require that the claims being made be clearly stated and adequately supported. Naturally more-ambitious interpretations require more justification. As Humpty Dumpty put it to Alice:

> Humpty Dumpty . . . "I meant by 'impenetrability' that we've had enough of that subject, and it would be just as well if you'd mention what you mean to do next, as I suppose you don't mean to stop here all the rest of your life."
>
> "That's a great deal to make one word mean," Alice said in a thoughtful tone.
>
> "When I make a word do a lot of work like that," said Humpty Dumpty, "I always pay it extra." (Carroll, 1871/2010)

If scores are to do more work, we have to pay them more by supplying the evidence needed to support the additional claims being made.

More-ambitious interpretations require more backing. This can be burdensome, but a clear specification of the proposed interpretation and use also puts limits on the evidence required for validation. If the interpretation does not include a particular inference, there is no need to provide backing for the corresponding warrant, and in fact, providing such evidence would muddy the waters. If the interpretation and use being proposed does not involve generalization over a facet, there is no need to evaluate generalizability over that facet. If the IUA does not involve extrapolation to a broadly defined target domain, there is no need to investigate whether such extrapolations are warranted.

As noted earlier, in developing a test and an IUA, we generally start with a proposed interpretation or use and work backwards to develop an IUA and a test that would support the proposed interpretation or use. Then in validation, we evaluate how well the evidence supports the inferences and assumptions in the IUA.

This had seemed to be a settled issue (Cronbach, 1971; Kane, 2006; Messick, 1989; Moss, 2007), but recently, some authors have suggested that it is the test that is to be validated (Borsboom, Mellenbergh, & van Heerden, 2004; Lissitz and Samuelsen, 2007). For example, Lissitz and Samuelsen argue that a test's "claim to be a seventh-grade mathematics test . . . is met or refuted by the process used to create the test and an analysis of the test items and their relationship to socially constructed definitions of what we mean by saying that a test is a measure of such material"

(pp. 442–443). For me, "a socially constructed definition of what we mean" by seventh-grade mathematics seems like a general statement of an interpretation. So as I see it, Lissitz and Samuelsen are not treating validity as a property of tests *per se* but are treating it as a property of tests for a particular kind of interpretation.

Lissitz and Samuelsen (2007) object to using correlations with other variables as significant parts of validation as tends to occur under the weak form of construct validity. I share their concern about encumbering relatively simple interpretations with validation requirements that are appropriate for more-ambitious interpretations. Under the argument-based approach, if the scores are interpreted in terms of an observable attribute, correlations with other variables generally would be irrelevant to their validation. It does not seem necessary to reduce the scope of validity theory to one kind of interpretation and to reduce the methodology of validation to content-based methods to solve the problem that Lissitz and Samuelsen seem to be most concerned about (i.e., not burdening the validation of observable attribute with extraneous requirements).

Borsboom et al. (2004) also take validity as a property of tests *per se* but add a strong theoretical, causal component to the interpretation that is adopted as the standard interpretation:

> Validity is a property of tests: A valid test can convey the effect of variation in the attribute one intends to measure. . . . A test is valid for measuring an attribute if variation in the attribute causes variation in the test scores. (p. 1067)

Borsboom et al. (2004) are emphatic about the claim that validity is a property of the test, but they also are very emphatic about specifying a particular kind of interpretation, an interpretation in terms of an underlying, causal attribute. This approach can provide a rich interpretation of test scores if the causal theory is justified, but if the causal theory is not specified and evaluated we have an observable attribute with an added trait assumption.

Under the argument-based approach, it is perfectly legitimate to develop a test to reflect some underlying attribute and to interpret the scores in terms of that underlying causal attribute. In the next section, I will discuss IUAs for such interpretations. However, I see no reason to restrict our conception of validity to this particular kind of interpretation. To restrict ourselves to any one kind of interpretation would be inconsistent with trends over the last century and would limit our ability to respond to most needs in applied contexts and in research. For example, observable attributes describe consistencies in performance, but they do not explain these consistencies. They can be quite useful (especially in applied contexts) in defining competencies and in specifying observable phenomena, but their use does not require the specification of causal explanations for the performances. If we cannot measure anything without an explanatory theory, we may not be able to collect the kinds of data that would help us to develop theories.

As noted earlier, it is hard to imagine validating a test as such without having some idea of the proposed interpretation and use. How do we determine whether a test is measuring what it is supposed to measure if we do not know what it is supposed to measure? Suggesting that it is tests that are validated make sense only if some interpretation or use is implicitly assumed.

Recap of Section "Tailoring Validation to Proposed Interpretations"

The discussion in this section has made three main points and several ancillary points. First, test scores can have multiple possible interpretations, and it is the proposed IUA that is validated. The IUAs for seemingly similar interpretations can differ substantially if considered carefully, and therefore the proposed IUA needs to be specified with care. An IUA that understates the intended interpretation and use (by omitting some inferences or assumptions) begs at least some questions, and as a result the validation effort will not adequately evaluate the actual interpretation and use. An IUA that overstates the interpretation and use (by including some inferences or assumptions that are not required for the actual interpretation and use) will make validation more difficult and may lead to an erroneous conclusion that the scores are not valid for the interpretation and use. If we embark on a validation effort without being clear about the proposed interpretation and use, it is easy to ignore implicit assumptions and thereby to beg questions or to make unnecessary assumptions and thereby to set up straw men.

Second, the validity of a proposed interpretation depends on how well the evidence supports the interpretation. The proposed IUA is valid to the extent that it is coherent and complete and its inferences and assumptions are adequately supported. As a result, the evidence needed for validation may be extensive, but it is limited and fairly well defined. The only properties of test scores or relationships with other variables that need to be investigated are those inherent in the proposed IUA.

Third, more-ambitious interpretations require more support (i.e., backing) than less-ambitious interpretations. In most cases, generalizability/reliability is a necessary condition for validity because generalization is an integral part of the interpretation/use for almost all observable attributes; usually we are interested in drawing conclusions about expected performance over replications of the test rather than in a specific replication of the testing procedure (Brennan, 2001a). Because testing does not generally involve random sampling from the full target domain defining the observable attribute, most test-score interpretations require an extrapolation inference from the universe of generalization to the target domain. Additional claims (e.g., norm- or criterion-referenced scaling, inferences to other variables, trait interpretations) can enhance the usefulness of observable attributes, but they also add to the evidential requirements for the interpretation and use.

There tends to be a trade-off between the generalization and extrapolation inferences. If the test observations are highly standardized and very similar in content and format, generalizability is likely to be secure but extrapolation to a broadly defined target domain may be relatively shaky. To the extent that we make the test more representative of the target domain, generalizability might decrease but the extrapolation inference would be easier to justify. The trade-off between generalization and extrapolation leads to the traditional reliability/validity paradox.

Theory-Based Interpretations

In this section, I will focus on two additional points about interpretations and their argument-based validation and will use theory-based construct interpretations to illustrate the points being made:

4. More-ambitious interpretations (e.g., a construct interpretation or a causal claim) tend to be more useful than less-ambitious interpretations, but they are harder to validate.
5. Interpretations and uses can change over time in response to new needs and new understandings.

These points apply to validation in general, but theory-based interpretations, which build on the observable attributes discussed in the last section, provide a convenient context for the discussion.

Observable attributes serve two important functions in the development and evaluation of theories. First, the phenomena of interest, which typically are described in terms of observable attributes and relationships among estimates of the observable attributes, specify the phenomena that the theory is to explain and thereby shape the development of the theory. Second, the observable attributes provide the grist for empirical checks on the theory. The relationships among observable attributes provide the "stubborn, dependable, replicable puzzles" (Cook & Campbell, 1979, p. 24) that the theory is to solve, and therefore they provide criteria for evaluating the theories (Popper, 1962).

Developing a theory that accounts for test scores can greatly enhance the usefulness of the scores. Assuming that the theory makes predictions about a variety of phenomena, the scores can be used, in conjunction with the theory, to make predictions about these phenomena (Embretson, 1983). For example, blood pressure readings are useful in medicine mainly because they are known to be directly related to the functioning (and malfunctioning) of the heart and circulatory system. Note, however, that the substantial utility of blood-pressure measurements does not come cheap; it is based on well-confirmed physiological theories and depends on many decades of empirical research and theory development.

Theory-Based Interpretations: Constructs and Indicators

Once we have observable attributes and relationships between observable attributes, we are likely to develop theories that explain the observed regularities. These theories commonly posit theoretical constructs which are assumed to be related to each other and to the observable performances in some way.

Cronbach and Meehl (1955) developed their model of construct validity in terms of nomological theories which consist of networks of nomological (or "law-like") relationships among theoretical constructs. These constructs are defined in terms of their roles in the theory. In addition, the nomological nets contain links that connect some of the constructs to observations. A construct is defined as "some postulated attribute of people assumed to be reflected in test performance" (Cronbach & Meehl, 1955, p. 283); the constructs are not directly observable, but they are tied, through the theory, to observable attributes that can be used to estimate the constructs. Empirical checks on how well the theory predicts relationships between observable attributes provide evaluations of the theory and also provide an empirical check on the validity of the indicators of the theoretical constructs.

Process models explain observed performances in terms of cognitive processes (Embretson, 1983; Mislevy, Steinberg, & Almond, 2003; National Research Council,

2001; Pellegrino, Baxter, & Glaser, 1999). In addition to observable attributes specifying the performances to be explained, process models generally employ parameters (constructs) that characterize latent abilities that account for performance and can be estimated from aspects of the observed performances.

For example, a process model might assume that some cognitive skills are required to perform some kinds of tasks. A test taker's performance on these tasks depends on ability parameters representing a test taker's mastery of the relevant skills as well as task parameters indicating the level of each skill required by the task (Embretson, 1984; Embretson & McCollam, 2000). Estimates of the ability and task parameters can be developed using assumptions in the process model that relate these parameters to the test takers' scores on the tasks included in a test.

An *indicator* of a theoretical construct can be defined in terms of a domain of possible observations (e.g., certain kinds of tasks administered under certain conditions, certain stimuli presented in certain contexts, etc.) that can be used to estimate the theoretical construct. For example, a process model for subtraction probably would include "borrowing" from an adjacent column as a skill, and the model would predict that students who lack this skill generally would get wrong subtraction problems that required borrowing and might or might not get right most other subtraction problems. An indicator for skill in "borrowing" could be based mainly on performance on subtraction problems that require "borrowing" (an observable attribute) or on the difference between performance on subtraction problems that require "borrowing" and similar problems that do not require "borrowing." Note that the theoretical construct, skill in employing "borrowing" when needed, is not directly observable, but performance on subtraction tasks that do or do not require this skill is observable.

In this context, there is no need to extrapolate to a broadly defined target domain that goes beyond the universe of generalization for the indicator. That is, the indicators tend to be defined operationally. It is in the context of theory testing that operational definitions were first introduced (Bridgeman, 1927).

The indicator does not define the construct, and a construct can have multiple indicators. The indicator is specified as an operationally defined attribute, but in its role as an indicator it is interpreted mainly as an estimate of the construct defined by the theory. For the interpretation of the scores as estimates of the construct to be plausible, the theory must be plausible, and for the theory to be accepted, its predictions about observed phenomena must be reasonably accurate.

If a theory is rejected, the interpretations of scores as indicators of the theoretical constructs defined by the theory also would be rejected, but the interpretations of the operationally defined indicators as observable attributes need not be rejected. The failure of a theory to account for certain phenomena generally leads to a rejection of the theory, but it does not necessarily lead to a rejection of prior descriptions of the phenomena in terms of observable attributes.

IUAs for indicators of theoretical constructs. As is usually the case, the IUAs for indicators of theoretical constructs can be developed backwards. The theory indicates the kinds of observations that would be relevant to a particular construct and the kind of scoring rule that would be appropriate. For example, if the theory predicts that individuals with high values on a construct are likely to perform well on some tasks

and that individuals with low values of the construct are likely to perform poorly on these tasks, performance on these tasks could be used to develop an indicator for the construct.

In general, it is desirable that the indicator includes a fairly broad sample of relevant performances so that it fully represents the construct as intended in the theory and is not unduly affected by idiosyncratic characteristics of particular observations (Cook & Campbell, 1979; Messick, 1989). Defining the indicator broadly helps to avoid construct underrepresentation and construct-irrelevant variance. The theory also may indicate various conditions of observation that could have a substantial impact on the observations. In developing an indicator, efforts would be made to control any potential source of irrelevant variation. Basically, indicators are designed to provide plausible estimates of the construct and to be relatively free of systematic and random error.

The IUA for an indicator of a theoretical construct generally would include at least three inferences: scoring, generalization, and an inference from the value of the indicator to the value of the construct. The first two inferences support the indicator as an operationally defined observable attribute.

The interpretation of indicator scores as estimates of a theoretical construct extends the interpretation to a claim about a construct as defined by the theory. A theory-based interpretation of scores on an indicator assumes that the theory is plausible and that the indicator provides an appropriate estimate of the construct, as defined by the theory. The warrant for this inference is a rule that takes the value of the indicator as an estimate of the construct. The backing for the warrant would involve analytic and empirical evidence supporting the theory and the appropriateness of the indicator. The analytic evidence relies on analyses of the relevance of each indicator to its construct and is produced during the development stage, as the indicators are designed to fit the theory. The empirical evidence examines how well the theory's predictions (with the constructs estimated by their indicators) agree with observable phenomena.

It is useful to maintain a distinction between theoretical constructs and observable attributes and the corresponding distinction between indicators of theoretical constructs and measures of observable attributes. The constructs are defined in terms of their roles in the theory, while the observable attributes are defined in terms of target domains of observations. Observable attributes can be used as indicators of theoretical constructs, but this use generally does not eliminate any prior interpretation of the observable attribute; rather, it adds another theory-based layer of interpretation to the original interpretation in terms of the observable attribute. As a result, the observable attribute plays a dual role: as an observable attribute defined in terms of some target domain, and as an indicator of the theoretical construct.

Validating theory-based interpretations. Once the indicator is defined in terms of a universe of generalization and scoring rule, the scoring and generalization inferences would be evaluated in much the same way as they would be evaluated for other operationally defined attributes. Most of the additional evidence needed to evaluate a theory-based inference would be that needed to evaluate plausibility of the theory, which includes the use of the indicator as an estimate of the construct.

To the extent that a theory's predictions are supported, both the theory and the proposed construct interpretations of the indicators are supported. If the predictions do not agree with the observed relationships, some part of the theory is called into question. The plausibility of the theory and of interpretations based on the theory could be evaluated over a number of studies. A theory that survives a range of serious challenges can be accepted, at least presumptively (Cronbach, 1980; Lakatos, 1970; Loevinger, 1957; Popper, 1962).

Theory-based construct interpretations can be challenged by claiming that the theory is not plausible. Even if the theory as a whole is considered credible, the appropriateness of a particular indicator for a theoretical construct still can be challenged in terms of either construct underrepresentation or construct-irrelevant variance (Cook & Campbell, 1979; Messick, 1989). For indicators of theoretical constructs, questions about representativeness focus on the extent to which the indicator elicits the full range of processes associated with the construct (Bachman, 2002; Borsboom et al., 2004; Embretson, 1983; Loevinger, 1957). Any factor other than the construct that has an impact on the indicator (e.g., the method or context of observation) is a source of construct-irrelevant variance, or systematic error.

Messick (1989, p. 34) defines construct-irrelevant variance in terms of "excess reliable variance that is irrelevant to the interpreted construct." Limiting the test to one method of assessment for a broadly defined trait can lead to both underrepresentation of the trait and irrelevant method variance. These problems can be alleviated by including multiple task formats (Messick, 1989, p. 35). If the observations are sufficiently diverse, "the errors are uncorrelated and more or less cancel each other out" (Loevinger, 1957, p. 648).

In addition to posing a general threat to validity, irrelevant variance can create an exception for certain individuals or groups. For example, a test may require some competencies (e.g., computer skills) that are not associated with the construct being estimated. The extra competencies required by the test might not be a barrier for most students but could be a serious source of systematic error for some students (e.g., those with limited experience with computers).

Under the strong form of construct validity, the validity of an interpretation of scores in terms of the value of a theoretical construct depends mainly on the plausibility of the defining theory and the plausibility of the assumed relationship between the indicator and the construct (based mainly on empirical checks). On a fundamental level, if we have more than one indicator for a construct, these indicators should agree (convergent evidence; Campbell & Fiske, 1959). If the theory assumes that two constructs are independent of each other, their indicators also should be independent (divergent evidence). If the theory postulates that two constructs are related in some way (directly, inversely, through a latent trait model), the observed empirical relationships between the corresponding indicators should be consistent with the theory's predictions. However, the validity of an indicator does not depend on the construct indicators' relationships with other variables that are attached to the defining theory.

In the context of process models, Embretson (1983, 1998) distinguished two aspects of validity: construct representation and nomothetic span. *Construct representation* is defined in terms of "the processes, strategies, and knowledge that persons

use to solve items" (Embretson, 1998, p. 382), and *nomothetic span* is defined in terms of the relationships of test scores with other variables. The validity of the construct representation, or construct meaning in terms of the process model, is supported by evidence that supports the proposed cognitive model.

The nomothetic span goes beyond the interpretation of scores in terms of a theoretical construct. Within the framework adopted here, construct representation can be associated with the meaning of the construct, and the validity of the proposed construct interpretation depends on the evidence for construct representation. Nomothetic span, on the other hand, describes additional implications associated with the construct label that may or may not be relevant to validation. Additional implications that are not included in the proposed interpretation/use generally would not be relevant to validation. However, implications inherent in the proposed IUA would be evaluated as part of the validity argument. For example, if scores on an indicator of a theoretical construct are used in selecting candidates for a job, the IUA generally would include (explicitly or implicitly) claims that higher scores on the test are related to better performance on the job; this assumption might not be included in the theory defining the construct, but it is highly relevant to the proposed IUA for the selection procedure.

The argument-based approach to validation generally imposes serious demands for evidence to serve as backing for the warrants included in the IUA, but it also limits the need for evidence in important ways. In particular, claims about relationships to other variables that are not relevant to the proposed interpretation do not have to be examined as part of validation.

Construct validity and theory testing. In theory testing, predictions are derived from the theory and compared to observations, which are to "serve as the universal, neutral arbiter among alternative hypotheses" (Galison, 1987, p. 7). The observations are taken as given, and the theory is evaluated in terms of how well the predictions fit the data (Lakatos, 1970; Popper, 1962). Theories are developed to account for phenomena and the phenomena can be specified in terms of observable attributes without relying on the theory (Galison, 1987; Guion, 1977). The specification of the phenomena in terms of observable attributes typically occurs prior to the development of the theory.

If the theory's predictions are accurate, confidence in the theory and in the validity of the indicators as measures of their theoretical constructs increases. If the theory's predictions are not accurate, the theory would be rejected and estimating the values of constructs in the abandoned theory would be of little interest. However, the descriptions of the phenomena in terms of observable attributes do not necessarily change.

Theories are likely to develop and evolve over time, and since the interpretation of a theoretical construct depends on its defining theory, the interpretation of indicators of these constructs will develop and evolve with the theory. If the evidence for the theory increases, confidence in the theory as a whole, including confidence in the appropriateness of the indicators, is likely to increase. If the theory is extended to account for a wider range of phenomena, the interpretation also expands. The extension of the theory to new types of phenomena will necessarily require additional evidence to support the broader application of the theory.

The theory may need to be modified over time, especially if it runs into empirical results that contradict some part of the theory (Lakatos, 1970). On a more drastic level, the theory may run into empirical results that contradict the core assumptions of the theory. In such situations, the theory and all of the theory-based construct interpretations based on the theory may be abandoned, especially if an alternate, better supported theory is available. The construct interpretations rely on their defining theory, and when the theory goes, these theory-based interpretations go with it. Note, however, that if a previously existing observable attribute had been used as an indicator of a construct, its role as an observable attribute, defined in terms of its target domain, is not invalidated by evidence against the theory.

Causal Interpretations

Messick (1989) defined a trait as "a relatively stable characteristic of a person . . . which is consistently manifested to some degree when relevant, despite considerable variation in the range of settings and circumstances" (p. 15). Trait language tends to be implicitly causal, with the trait accounting for performance on tasks in the target domain. For observable attributes with some attached trait assumptions, the nature of the trait is not specified and causal mechanisms are not specified.

Meehl (1950) cautioned that to infer the existence of a trait from observed regularities in performance and then to use this "trait" to explain these regularities is to reason in a tight circle. Assuming that we have no additional evidence to support the existence of the trait, the use of a trait to explain observed regularities in performance adds nothing to our understanding. Cureton (1951) also pointed out the potential for circular reasoning in causal interpretations:

> We must not say that his high score is due to his high ability, but if anything the reverse. We say he has high ability because his performance has yielded a high criterion score. His "ability" is simply a summary statement concerning his actions. (p. 641)

Causal explanations of observed regularities in performance in terms of traits can be immensely powerful if the traits can be shown to exist in some sense and to be largely or exclusively responsible for the observed performances, but establishing a causal connection between the proposed trait and observed performance generally requires strong theory supported by extensive empirical evidence.

Causal inferences can be justified in at least two ways: empirically (by showing that the introduction of the "cause" consistently leads to the effect and that the removal of the cause removes the effect) and/or theoretically (by establishing a well-confirmed theory that implies that the cause generally will yield the effect). In many cases where it is not possible to perform well-controlled experiments (e.g., in establishing that cigarette smoking causes cancer in humans), both approaches play a major role.

The empirical approach outlined by John Stuart Mill in 1843 in his *System of Logic* provides the conceptual foundation for the experimental method; we assign the "cause" to the experimental group and withhold it from an otherwise-equivalent control group to see if the effect is found for the experimental group and not for the

control group, or we apply the cause to an existing group and see if the effect ensues (see Mill, 2002). For example, if a light goes on every time we flip a switch to the "on" position and goes off every time we flip it to the "off" position, we have a reasonable basis for claiming that the switch turns the light on and off. We have not specified a causal mechanism, but we have strong evidence that, under the conditions prevailing during the experiment, the "cause" (flipping the switch) produced the "effect" (the light going on and off). This direct empirical approach would be, at best, very difficult to employ in establishing a causal connection between a trait (which cannot be turned on and off at will) and test performance.

Alternately, a solid case can be made for a causal inference by deploying a well-established theory that implies that, under certain conditions, the cause will produce the effect. Under this approach, claims about the causal connection between the switch and the light can be justified in terms of certain well-established theories (mainly circuit theory). If we have confidence in the theory, we can have confidence in the claim that flipping the switch turns the light on and off.

> To explain a test score, one must bring to bear some sort of theory about the causes of the test performance and about its implications. Validation of test interpretations is similar, therefore, to the evaluation of any scientific theory. (Cronbach, 1971, p. 443)

Strong causal claims typically are based in large part on a theory that specifies the causal mechanism through which the causal agent produces its effects. Given such a theory (Embretson, 1998) and adequate evidence to support the theory, causal claims can be made with some confidence.

Borsboom et al. (2004) have argued for a general causal framework for validity which claims that validity requires

> that an attribute, designated by a theoretical term like intelligence, exists and that measurement of the attribute can be performed with a given test because scores are causally affected by variation in the attribute. (p. 1067)

This is a very strong kind of interpretation if it can be made good, but it is a very difficult kind of interpretation to justify. As Borsboom, Cramer, Kievit, Zand Scholten, and Franic (2009) put it,

> the price paid for these semantics is that one has to make reference to the property measured as a causal force that steers the direction of the measurement outcomes. This is no small matter. It requires a very strong assumption about what the world is like, namely that it contains some property that exists independently of the researcher measuring it. ... It also obliges the researcher to explicate what the property's structure or underlying process is and how this structure or process influences the measurement instrument to result in variations in the measurement outcomes. This seems to be a very daunting task indeed for many psychological properties that researchers claim to measure. (p. 148)

Under this approach, validation would require the deployment of a well-established theory of test performance that specifies how the property produces the observed test performance. This is a stronger interpretation than that required for the

strong program of construct validity (Cronbach, 1989); it requires a causal, explanatory theory (and not just a nomological net).

As noted above, this approach would provide a very effective form of validation in those cases where it is applicable, but at present it would seem to have a very limited range of applicability. For essentially all applied testing programs (i.e., academic achievement tests, licensure and certification tests, employment tests, and selection tests) and for most psychological attributes (e.g., anxiety, intelligence), we have no serious candidates for the kind of detailed process model required by this kind of strict causal interpretation. Note that it is not that we have models that are not fully vetted but that we do not even have a sketch of the required model. We have some models for some well-defined activities and for some elementary skills in mathematics, science, and reading, but that is about it.

Furthermore, in many—if not most—cases where validity is an important issue, it seems unlikely that any single property causes the observed regularity of performance. For example, Borsboom et al. (2009) associate general intelligence with "a single linear ordered property, if there is one, that causes individual differences on IQ-tests" (p. 144). It is of course possible that eventually we will isolate such a property for intelligence, but it is by no means a given, as Borsboom et al. recognize, and current research in cognitive psychology does not support the idea that test performances are explainable in terms of such simple traits (Mislevy, 2006). Furthermore, it seems extremely unlikely that we will identify any such single linearly ordered property for achievement in academic subjects (history, biology) or for competence in the broadly defined domains of KSJs required in professional licensure, for occupational certification, or for educational selection or placement. In all of these cases, the intent is not to measure a single linearly ordered property but to assess the overall level of competence in a cluster of skills that are needed in a particular context.

So it would seem that, in most contexts, interpretations of test scores that rely on strong causal models are a long way off and that the validation of such causal interpretations is even further off. Nevertheless, the validation of a strong causal interpretation of test scores in terms of a construct or property that accounts for the variation in test scores is currently feasible in a few narrowly defined areas (Mislevy, 2006; National Research Council, 2001) and it may have wider application in the future.

The argument-based approach to validation allows for a wide range of possible interpretations, including observable attributes, theory-based construct interpretations, and causal interpretations, where feasible. In research contexts, where the goal is to develop explanations of phenomena, theory-based interpretations are highly valued. In many applied contexts, observable attributes (defined in terms of a target domain of possible performances which covers the content of a course, skill in some activity, or a performance domain associated with some activity) may be entirely adequate. Even in these latter cases it would be desirable to have explanatory models for the observed performances (Zumbo, 2007, 2009), but it is not necessary to have such models to validate IUAs that do not make causal claims. In many applied contexts, we need an indication of how well individuals can perform in certain contexts or domains (e.g., in reading college texts, in debugging computer programs, in preparing a legal documents, or taking a medical history), but we do not need a detailed explanation of the processes they employ in accomplishing such tasks.

Recap of Section "Theory-Based Interpretations"

This discussion of theory-based interpretations has made two main points about argument-based validation: that more-ambitious interpretations (e.g., theory-based interpretations) tend to be more useful than less-ambitious interpretations but also need more evidence for their validation, and that interpretations and uses can change over time in response to new developments. We get to design and develop tests and to specify a proposed interpretation of the test scores. Of course, the task at hand, the context, and the target population shape and limit our options, but we always have choices in test design and score interpretation.

Theory-based interpretations can be particularly powerful and useful, because once a theoretical construct is measured, it can be used (in conjunction with the theory) to draw theory-based inferences about the performances covered by the theory. If the theory is a very general theory (e.g., Newtonian mechanics), we may be able to make predictions about a wide range of phenomena based on a single measurement (e.g., the mass of an object), and if the theory is well supported and the observed scores have been validated as measures of a construct in the theory, we can make these predictions with confidence. But as Humpty Dumpty suggested, there is no free lunch! Validating test scores as indicators of a theoretical construct requires enough evidence to support the theory and to support the appropriateness of the indicator as a measure of the construct.

Under the argument-based approach to validation, interpretations can change over time as uses change and as new evidence becomes available. For example, if we decided to use a validated measure of competence in algebra (defined as an observable attribute) to place students into different science courses, we would need additional evidence to support this extended IUA (e.g., evidence that the test scores predict performance in the science courses). If we decide to extend the interpretation in another direction by providing an explanation for the scores in terms of a process theory and interpreting the scores as indicators of a construct in the theory, we need to support this more-ambitious interpretation.

The proposed interpretation should be consistent with the purpose at hand. It should not understate the inferences and assumptions inherent in the claims that are being made, thereby begging various questions. It also should not overstate the claims being made (e.g., by adding causal assumptions in cases where such assumptions are not needed). Overstating the claims puts unnecessary burdens on test development and validation and makes it more likely that good tests with defensible interpretations will be rejected because they do not meet the inflated requirements. Adding unnecessary inferences and assumptions to the IUA can set up a straw man that may be too easy to knock down. A Goldilocks criterion is applicable here: the proposed IUA should support the proposed interpretation and use in the target population and the anticipated contexts of use. It should be neither too limited nor too ambitious.

Score Uses

In this section, I discuss the role of consequences in evaluating test score uses, and in doing so I emphasize three additional points about test-score uses:

6. The evaluation of test score uses requires an evaluation of the consequences of the proposed uses, and negative consequences can render a score use unacceptable.
7. The rejection of a score use does not necessarily invalidate a prior, underlying score interpretation.
8. The validation of the score interpretation on which a score use is based does not, in itself, validate the score use.

In developing these points, I am arguing for including evaluations of specific kinds of consequences (but not all consequences) under the heading of validity.

Test scores typically are used to make decisions (e.g., selection, instructional planning, accountability) about individuals (or groups). The general notion is that the test scores are interpreted as indicating something about the test takers (e.g., their level of achievement in some domain), and a decision is made in accordance with a *decision rule* which stipulates that certain actions be taken given certain test scores. The decision inference takes an interpreted score as its datum and yields a decision as the claim. The IUA for score-based decisions generally will involve a chain of inferences leading to claims about attributes of test takers and then to decisions based on the estimated values of the attribute.

The decision rule is the warrant for the decisions, and the backing for this warrant typically consists of analyses indicating the appropriateness of the decision rule given the purpose of the program. Decision rules are evaluated in terms of their expected success in achieving their goals at a reasonable cost and with acceptable consequences. The IUA for a score-based decision includes all of the inferences leading up to the decision, plus the decision rule itself. The decision rule tends to be the capstone of the IUA for score-based decisions, and in many cases the rest of the IUA is developed to support the decision rule.

The backing for the decision warrant would need to make the case that the decision rule works well for the target population as a whole in the sense that the positive consequences outweigh the negative consequences in some sense (e.g., expected utility). The testing program is evaluated in terms of how well it achieves the goals of the program and in terms of any unintended consequences (e.g., adverse impact and negative systemic effects) that are perceived to be potentially serious for the population of interest or some substantial subset of the population.

Idiosyncratic consequences that are unique to a particular test taker (e.g., getting nervous about the test or unhappy with the results) generally would not be considered in evaluating the decision rule, as such, unless their occurrence indicates a larger problem of some kind. Within Toulmin's model, such unique events could be considered under the heading of exceptions to the warrant. The test user could address these issues on a case-by-case basis, but the existence of such exceptions generally would not have much impact on the evaluation of the decision rule as a warrant for decisions. In the evaluation of decision rules, the focus would be on consequences that apply across the population or a major part of the population. That is, the focus would be on "... the consequences of the repeated and pervasive practices of testing" (Moss, 1998, p. 11).

Like all presumptive inferences, the decision inference is always qualified to some extent; we are never sure that we are making the right decision for the individual

under consideration. For example, if the decision rule is being used to place students into one of a sequence of courses, the qualifier might be stated in terms of the probability that a student with a particular score will succeed in the course. The qualifiers for binary decisions (e.g., pass/fail) often are specified in terms of false-positive and false-negative error rates.

Decision Rules and Consequences

Decision rules are evaluated in terms of their overall consequences (or utilities, or values) for some population (Cascio, 1993; Cronbach & Gleser, 1965; Peterson, 2009). In formal applications of decision theory, the value or utility of every possible outcome of the decision rule is specified, the probabilities of the various outcomes of the decision rule are estimated for the population, and the decision rule is evaluated in terms of some summary statistic, like the expected utility of the outcomes over the relevant population (or, more conservatively, in terms of minimizing the maximum possible loss).

In practice, the analysis of outcomes usually is not straightforward because generally it is not feasible to specify the required utility functions in any detail; the principle of weighing positive consequences (or gains) against negative consequences (or losses) in evaluating decision procedures is, however, generally accepted (Cronbach, 1971: Heubert & Hauser, 1999; Messick, 1989; Shepard, 1993, 1997).

> To be given due consideration in the first place, a policy or hypothesis must have some explicit or implicit warranting case in its support … And there are important questions that need to be raised early on about this "live option:" Would this policy be likely to solve the policy problem that it is directed at? Is it just? What are the likely social and political and economic consequences, and are these morally permissible? How are the unintended consequences of this policy to be weighed against the putative benefits? (Phillips, 2007, pp. 394–395)

A decision rule that achieves its goals at an acceptable cost and with acceptable consequences is considered a success. A decision rule that does not achieve its goals or has unacceptable consequences is considered a failure. The backing for a decision warrant is derived from analyses of its consequences.

Evidence for the accuracy of the information on which decisions are based is clearly relevant to the evaluation of a decision rule, mainly because the use of inaccurate information as the basis for a decision would be unacceptable ethically, legally, and socially, and because more accurate information generally should lead to better outcomes. If the information is shown to be erroneous, the rational basis for the decision rule is undermined. However, even indisputable evidence for the validity of a proposed test-score interpretation does not, in itself, justify a score-based decision rule (Cascio, 1993; Cronbach, 1971, 1988; Cronbach & Gleser, 1965). A highly accurate diagnostic test for an untreatable disease would not be used for screening for that disease, especially if the test involved serious side effects or risks (Hershey & Asch, 2001). A decision procedure that does not achieve its goals or has serious negative consequences is not likely to be adopted, even if it is based on perfectly accurate information.

It can be difficult to get agreement on the values/utilities to be assigned to different outcomes and for claims to be convincing to diverse audiences (even data-based

claims); the assumptions on which the claims depend must be credible to those audiences (Cronbach, 1982, 1988; Ryan, 2002). This concern is especially salient for high-stakes testing programs, because perceptions about the seriousness of various kinds of consequences may be highly variable across stakeholders and across groups of stakeholders. Those who have to suffer negative consequences have a tendency to see them as more serious than those who do not suffer them.

Although there often are disagreements about the values (or utilities) associated with various outcomes and about the criteria for evaluating overall utility for a decision rule (e.g., maximizing expected utility vs. minimizing maximum loss), there is broad consensus on the central role of outcomes/consequences/utilities in evaluating decision procedures (Cascio, 1993; Peterson, 2009). Claims about the efficacy of score-based decision procedures require evidence that the intended outcomes are being achieved at a reasonable cost and with limited negative consequences.

Evaluating the Consequences of Score-Based Decision Procedures

Test-based decision procedures have a range of potential consequences, and the kinds of consequences that are considered have tended to expand over time as our understanding of how testing programs function in the world has grown (Bachman & Palmer, 2010; Camilli, 2006; Crooks, 1988; Messick, 1989). In this section, I review three major categories of outcomes/consequences that have played and continue to play a major role in evaluating score-based decision rules: (1) the extent to which the intended outcomes are achieved, (2) differential impact on groups (particularly adverse impact on legally protected groups), and (3) positive and negative systemic effects (particularly in education):

> Testing programs should be evaluated to see if they are achieving their stated purpose. As part of the evaluation, consequences—both positive or intended and negative or unintended—should be carefully monitored and weighed. Newly developed tests, a ubiquitous feature of educational policy today, need to be studied for their impact on particular groups of test takers and their effects on curriculum and instruction. (National Research Council, 2007, p. 13)

The consequences of testing programs are not necessarily easy to identify or evaluate, but this difficulty does not diminish their importance.

We evaluate decision procedures in terms of the intended outcomes and unintended consequences that are considered serious (in the sense that they can have substantial effects on a significant number of individuals). The three kinds of consequences (intended effects, adverse impact, and unintended systemic effects) that have been identified by the profession and society (i.e., legislatures, the courts, and public opinion) as being particularly relevant (Pyburn, Ployhart, & Kravitz, 2008; Sireci & Green, 2000; Sireci & Parker, 2006) tend to be of particular interest to different groups of stakeholders (Ryan, 2002). As sponsoring agencies develop and implement testing programs to serve particular purposes, the extent to which the programs achieve the intended outcomes at an acceptable price tends to be the main focus along with concerns that the testing process be fair and be perceived as being fair. Second, with the development of greater sensitivity to questions of civil rights, differential impact of a testing program on various groups has become a major concern. Third,

as it has become evident that testing programs (particularly high-stakes programs) can have substantial, unintended effects on how institutions function (e.g., on what is included in school curricula), such systemic effects have become major concerns, especially in education.

The central issue is not whether consequences should be considered in evaluating decision rules. Rather, the main concerns are: the kinds of consequences that should be considered, who should take responsibility for evaluating these consequences, and how the positive and negative consequences should be weighed against each other. The measurement community can have input on the consequences to be considered and on their weighting (Cronbach, 1988; Linn, 1998), but we do not make the policy decisions and we have to work in an evolving social, political, and legal context.

Intended outcomes. Score-based decision procedures are intended to achieve some desired outcomes (e.g., placing students in courses in which they learn the most or hiring the most productive workers) and are evaluated in terms of how well they achieve these goals. Institutions have a very long tradition of using tests to predict outcomes (e.g., in college admissions, course placement, and employment), of using the predictions to promote the desired outcomes, and of evaluating the decision programs in terms of the extent to which they achieve their intended outcomes as measured by some criterion. Standard 1.22 of the 1999 *Standards* states that:

> When it is clearly stated or implied that a recommended test use will result in a specific outcome, the basis for expecting that outcome should be presented, together with relevant evidence. (American Educational Research Association et al., 1999, p. 23)

A testing program that does not achieve its goals might fail because of poor design or implementation, for technical reasons (e.g., poor generalizability, weak equating), or for contextual reasons. In any case, failure to achieve the goals counts against the program.

Assuming that the intended outcomes are achieved reasonably well, the costs of the program and any immediate negative consequences (e.g., any lack of procedural fairness) would be evaluated. If the perceived costs and/or any immediate negative consequences exceed the perceived benefits, the program will get a negative evaluation. This kind of cost-benefit analysis of testing programs has been around for a long time and generally has been discussed under the heading of validity (Cascio, 1993; Cronbach & Gleser, 1965). The Taylor-Russell tables, which yield the expected proportion of selected candidates who will be successful based on predictive validity coefficients, selection ratios, and base rates, and were used to evaluate the effectiveness of selection procedures, go back over 70 years (Taylor & Russell, 1939). Interest in the intended consequences of score-based decision rules goes back at least to the 1920s.

A more recent development is the use of testing programs as the engine of reform in test-based accountability programs in education (Koretz & Hamilton, 2006; Lane & Stone, 2002; Linn, 2009). The idea is to set high standards of student achievement and then to hold schools, districts, and teachers accountable for having their students achieve these standards. In these new programs, the test scores are not used

simply to select promising candidates or to monitor progress. The testing programs are expected to improve the educational programs by clearly articulating goals in terms of test performance and by incorporating incentives for good performance and penalizing poor performance. In evaluating the effectiveness of these accountability systems, it would be appropriate to examine how well they achieve their goals in terms of the motivation and practices of school administrators and teachers and in terms of improved student performance on the state standards (Lane & Stone, 2002). The extent to which these goals are being achieved can be investigated by studying what is happening in the schools and by examining student performance on the state tests and other tests (Linn, 2009).

Adverse impact. Before 1960, fairness issues did not get a lot of attention in the testing literature, but the civil rights movement, legislation, and litigation made fairness a major issue in evaluating testing programs (Cascio, Jacobs, & Silva, 2010; Cole & Moss, 1989; Ebel, 1966). By 1970, legislation and judicial decisions put fairness and equity issues for minorities and women firmly on the agenda (Wigdor & Garner, 1982), and, more recently, the impact of test-based decisions on individuals with disabilities has become a major issue.

The Supreme Court's decision in Griggs v. Duke Power Co. (1971) established the framework for subsequent analyses. Before 1965, all of the black employees at the Duke Power Company were in the labor department (the lowest salaried of its five departments). In 1965, as the Civil Rights Act of 1964 was about to take effect, the Duke Power Company introduced two aptitude tests (on which black applicants scored lower than white applicants) for anyone seeking employment in the upper level jobs. In a unanimous decision, the Supreme Court (Griggs v. Duke Power Co., 1971) found that the new requirements were "adopted 'without meaningful study' relating them to performance" (Guion, 1998, p. 190). The testing procedure limited the opportunities of groups protected under the Civil Rights Act, and no effort had been made to demonstrate any countervailing benefits:

> The message from the Court . . . was that Title VII covers "the consequences of employment practices, not simply the motivation of employers," and that a "manifest relationship" between the challenged practice and the "employment in question" must be proven if adverse impact is shown. (Gutman, 2005, p. 28)

That is, if the decision rule has adverse impact on a protected group, it has to be counterbalanced by positive consequences (e.g., enhanced performance on the job) if the program is to be considered legally acceptable.

In 1978, the federal agencies responsible for enforcing civil rights legislation published the *Uniform Guidelines on Employee Selection Procedures* (EEOC, Civil Service Commission, Department of Labor, & Department of Justice, 1979), which reinforced this principle by requiring that, if a selection procedure produced adverse impact for a protected group (e.g., racial minorities), the procedure could not be used for employment decisions unless the test scores were shown to be related to performance on the job. Adverse impact was said to exist if the selection rate for the protected group was less than four-fifths of that for the majority group. The *Uniform Guidelines* struck a balance between the employer's interest in hiring the best

candidates for each job and society's interest in group equity. At about the same time, the courts decided that a state cannot deprive a student of a diploma on the basis of test scores without providing the student with an opportunity to learn the content being assessed (Debra P. v. Turlington, 1981); again, the focus was on the balancing of positive and negative consequences, including adverse impact (Jaeger, 1989). Evaluation of differential impact has become a routine part of most high-stakes testing programs.

More recent court decisions have expanded on and modified the requirement that adverse impact (a negative consequence) needs to be offset by some legitimate institutional interest (positive consequences), but the basic principle is alive and well (Guion, 1998; Pyburn et al., 2008) and has spawned an extensive literature on how to balance concerns about adverse impact against the benefits associated with score-based selection (Cascio et al., 2010; De Corte, Lievens, & Sackett, 2007; Sackett, De Corte, & Lievens, 2010).

Note that adverse impact does not, in itself, indicate that the decision rule is not appropriate in a particular context:

> The idea that fairness requires overall passing rates to be comparable across groups is not generally accepted in the professional literature. Most testing professionals would probably agree that while group differences in testing outcomes should in many cases trigger heightened scrutiny for possible sources of test bias, outcome differences across groups do not in themselves indicate that a testing application is biased or unfair. (American Educational Research Association et al., 1985, p. 75)

If the skills and abilities being evaluated can "be shown to be necessary to safe and efficient job performance" (Campion, 1983, p. 529) and there are no indications that the testing procedures are unduly affected by irrelevant variance, the negative consequences do not rule out the use of the scores.

In the past, assessments that were correlated with performance but not based directly on the performance of interest often were used for selection. For example, physical strength and endurance are needed in many occupations (e.g., police, firefighters), and these attributes were assessed indirectly through height and weight requirements which were easily and accurately measured and were positively related to strength (Jackson, 1994). The use of these convenient, but indirect, indicators was rejected by the courts because their use had adverse impact (particularly on women) and had no demonstrated relationship to job performance (Campion, 1983). The height and weight requirements were perceived as only loosely connected to job performance (Jackson, 1994), and they were replaced by more direct measures of job-related strength and simulations of job performance (Campion, 1983; De Corte et al., 2007; Jackson, 1994).

Legal precedents, the Equal Employment Opportunity Commission (EEOC) *Guidelines*, and the *Standards* have assigned unintended consequences (particularly adverse impact) an important but complicated role in the evaluation of decision rules. All else being equal, adverse impact counts against the acceptability of a decision rule but it has to be weighed against the positive consequences associated with the testing program. How the trade-off between positive and negative consequences

should be resolved in practice is not always so clear, but it is clear that there is a trade-off.

Systemic effects. In educational and clinical contexts, testing programs are expected to support the primary goal of helping clients (e.g., students, patients) in some way, and any effects that a testing program has on the achievement of this primary goal would be relevant to the evaluation of the program. A testing program will be valued as part of an educational system to the extent that it promotes the goals of the program at an acceptable cost and without substantial negative consequences. The fact that the test scores provide accurate predictions of some criterion is not, in itself, a sufficient justification for the use of the tests.

Testing programs have long been known to have strong effects on how schools function, on how and what teachers teach, and on what students study and learn (Crooks, 1988; Frederiksen, 1984; Frederiksen & Collins, 1989; Herman & Golan, 1993; Heubert & Hauser, 1999; Lane & Stone, 2002). Some of the effects (e.g., motivating students to study certain content) were among the intended outcomes of many of the testing programs, but other effects were not intended. The systemic effects of testing programs on educational institutions can be positive or negative, but in either case, they have a major role in evaluating decision rules.

Cronbach and Snow's (1977) analyses of aptitude-treatment interactions indicate conditions under which a decision rule can enhance the effectiveness of an educational program and conditions under which the decision rule can interfere with the effectiveness of the program. For example, suppose that a university uses scores on an aptitude test to assign freshman engineering students to a regular mathematics sequence or to an accelerated sequence. Assume that the test scores accurately predict performance in the two sequences but that the score-based regression lines for predicting performance on a common criterion assessment in the two sequences are parallel, with students at every test score predicted to achieve more in the accelerated sequence than they would in the regular sequence. In this case, the test scores are not useful for assigning students to the two sequences, because the optimal policy is to assign everyone to the accelerated sequences (and presumably, to scrap the regular sequence). The validity of the interpretation in terms of predicted performance in the two course sequences does not provide justification for the use of the test scores in assigning each student to one of the two sequences.

Shepard (1993) provided a good example of the need for critical evaluation of the decision rule. Suppose that a "readiness" test provides a "valid" measure of certain basic skills (e.g., letter recognition) and scores on the test are positively correlated with performance in kindergarten. Does this justify its use in deciding whether to admit children to kindergarten this year or hold them back until next year? The use of the test might be predicated on an assumption that a low "readiness" score indicates a developmental lag that will be resolved by waiting a year. However, if the low score indicates a home environment that does not promote learning of the skills, keeping the child out of school for another year would be counterproductive. The program would not serve its intended purpose and would have negative consequences for those children who have not had an opportunity to learn the skills at home. The validity of the proposed interpretation in terms of mastery of certain skills does not justify the

use of the test scores for what is essentially a placement decision (kindergarten or home).

The Committee on Appropriate Test Use of the National Research Council (Heubert & Hauser, 1999) echoed Shepard's point:

> Determining whether the use of a test for promotion, tracking, and graduation produces better overall educational outcomes requires that the intended benefits of the policy be weighed against unintended negative consequences. These costs and benefits must also be balanced with those of making high-stakes decisions about students in other ways, without tests. (p. 23)

The Committee suggested that the use be evaluated in terms of three criteria: measurement validity, attribution of cause, and effectiveness of treatment. Measurement validity corresponds roughly to the validity of the underlying interpretation of the scores, while causal attribution and concerns about the effectiveness of treatment introduce the issue of consequences. The Committee argued that "neither test scores nor other information" should be used to place students in classes where they are "worse off than they would be in other placements" (Heubert & Hauser, 1999, p. 282).

Frederiksen (1984) provided a classic example of the positive and negative consequences that testing programs can have on teaching and learning. He recalled working on validity studies during World War Two and being surprised that verbal and reading comprehension tests were the best predictors of grades in gunner's mate school. Later he worked at the school and found that the instruction was delivered via lecture-demonstration and the study of manuals, and examinations were based on the lectures and manuals. Frederiksen's group introduced performance tests that required students to actually service real guns, and grades declined sharply. As the tests were given at the end of each unit, students got the word and began practicing assembling and disassembling guns. The instructors also got the point and "moved out the classroom chairs and lecture podium and brought in more guns and gunmounts" (Frederiksen, 1984, p. 201), and the pass rates on the new performance tests increased. The validity coefficients also changed, with mechanical aptitude and knowledge becoming the best predictors of grades in gunner's mate school.

> Note that no attempt was made to change the curriculum or teacher behavior. The dramatic changes in achievement came about solely through a change in the tests. The moral is clear: It is possible to influence teaching and learning by changing the tests of achievement. (Frederiksen, 1984, p. 201)

Testing programs can have dramatic systemic consequences, positive or negative.

As noted earlier, current educational accountability programs are designed to have effects on educational systems (e.g., improvements in school effectiveness and student achievement). In some current accountability programs, the design of the test gets little attention and the focus is on the decision rule as a policy tool. For example, the NCLB program uses student test scores to hold schools accountable for student progress by penalizing schools that don't meet certain benchmarks, but each state gets to use its own test and its own performance standards; the NCLB legislation

specifies the decision rules in some detail but leaves test design and development to the states (NCLB, 2002).

Test-based accountability programs have a range of potential benefits and costs (Hamilton, 2003; Kane, 2002b; Koretz & Hamilton, 2006; Lane, Parke, & Stone, 1998; Linn, 2009; McNeil, 2005; Mehrens, 1997). The potential benefits include increases in student achievement on the content areas covered by the tests (Linn, 2005) and possibly improvements in public confidence in the schools. The potential costs include the time and resources spent on testing, possible narrowing of the curriculum and of student options (e.g., fewer advanced placement courses), and increased dropout rates. These positive and negative consequences are likely to have different impacts on different groups and in different schools (Lane & Stone, 2002).

Accountability programs implement policies, and policies are evaluated in terms of their consequences (Cronbach, 1982; Phillips, 2007). However, any large-scale educational intervention is likely to have a variety of systemic effects—some of which may be positive and some negative—in addition to the intended effects (Crooks, 1988; Frederiksen, 1984; Madaus, 1988; Moss, 1998). The accountability program is an educational intervention, and a serious evaluation of an accountability program would require an evaluation of both intended and unintended outcomes.

Balancing different kinds of consequences. The evaluation of the decision rule, or decision warrant, requires evidence that goes beyond that typically required for other inferences in the IUA (Cronbach & Gleser, 1965). Consequences are the bottom line in evaluating decision rules, but in any decision context the range of consequences to be considered is limited to certain recognized categories of potentially serious consequences. For score-based decision rules, the focus has been on intended consequences, adverse impact, and negative systemic effects. As is true for most warrants, decision makers have latitude in the decision rules that they choose to implement, but once the decision rule is specified, the argument-based approach requires backing for this warrant and the backing involves evidence about consequences.

Negative consequences count against a decision rule, but they can be offset by positive consequences. A program can have substantial negative consequences and still be acceptable if the benefits outweigh any negative consequences. Negative consequences that are not offset by positive consequences tend to render a decision rule unacceptable (at least for stakeholders who are concerned about these consequences).

In reviewing a National Academy of Science report on ability testing (Wigdor & Garner, 1982), Messick suggested that the report was dispassionate and wise but that it "evinces a pervasive institutional bias" (1982, p. 9) by focusing on common analytic models for selection and classification, which emphasize the intended outcomes of the decision rule:

> Consider that, for the most part, the utility of a test for selection is appraised statistically in terms of the correlation coefficient between the test and the criterion . . . but this correlation is directly proportional to the obtained gains over random selection in the criterion performance of the selected group (Brogden, 1946; Cronbach & Gleser, 1965). . . . Our traditional statistics tend to focus on the accepted group and on minimizing the number of poor performers who are accepted,

> with little or no attention to the rejected group or those rejected individuals who would have performed adequately if given the chance. (Messick, 1982, p. 10)

Messick went on to suggest that

> By giving primacy to productivity and efficiency, the Committee simultaneously downplays the significance of other important goals in education and the workplace. (p. 11)

It certainly is appropriate to evaluate a decision rule in terms of the extent to which it achieves the goals of the program, but it is also important to attend to unintended effects that have potentially serious consequences.

Bad Decisions Based on Sound Score Interpretations

A finding that a decision rule has unacceptable consequences does not necessarily imply that there is anything wrong with the test being used or with the score interpretations. As Messick (1989) argued, negative consequences suggest a closer scrutiny of the testing program, but they do not, in themselves, invalidate the score interpretation. Popham (1997, p. 11) made this point by postulating a school board that adopts some bizarre decision rules based on test scores, "from which valid inferences can be drawn regarding the levels of a middle-school student's mathematics achievement":

> For example, the board requires that any girl who scores below the overall median will be . . . prohibited from engaging in any extracurricular activities, . . . and . . . boys who scored below the median would be expelled. (Popham, 1997, pp. 11–12)

Popham is assuming that the interpretation of the scores in terms of mathematics achievement is unproblematic, and goes on to argue that:

> Such absurd decisions, while deplorable, do not alter one whit the validity of the test-based inferences about students' mathematics achievement. Those inferences are just as accurate as they were before the board made its ludicrous test-use decisions." (Popham, 1997, p. 11–12)

That is, inappropriate test use does not imply that a prior underlying score interpretation of the test scores is invalid.

Popham's hypothetical example also illustrates a second, equally important point—that evidence for the validity of a prior, underlying interpretation does not, in itself, justify any particular uses of the scores (Shepard, 1997). In Popham's example, "absurd" and "deplorable" decisions are based on scores "from which valid inferences can be drawn" (Popham, 1997, p. 11). That is, even if the underlying interpretation is accepted as completely valid, the consequences of a proposed decision rule have to be evaluated to evaluate the score-based decisions. If the proposed use were more sensible than Popham's examples, the additional evidence needed to justify the decision procedure might be readily available; for example, if the math test scores were used by teachers to plan future instruction, the expected consequences probably would be mostly positive and the program might be justified without a lot of additional study, but even in such reasonable cases, some evaluation of the consequences is needed to justify the decision procedure.

To make his point forcefully, Popham (1997) posited decision rules that were especially "absurd" and "deplorable," and it is the expected consequences that make them absurd and deplorable. There does not seem to be any sensible goal to be achieved by the decision rules, and yet the rules almost certainly would have serious negative consequences for many students (e.g., being expelled). The lack of any likely positive consequences and the high likelihood of serious negative consequences make the decision rule indefensible, even though the underlying interpretation of the test scores is assumed to be valid.

Negative consequences generally do not count against the validity of a prior interpretation of test scores, but they do count against the decision rule. Even a very strong validity argument for some score interpretation (like that assumed by Popham) does not, in itself, justify the decision rule. To take a validation of a proposed interpretation of test scores in terms of some attribute as justification for a decision rule is to beg the question of whether the rule makes sense and is justified in a particular context for a particular purpose. In some cases, the extension of an IUA from a claim about an attribute of a test taker to a decision about the test taker may seem like a minor extension (e.g., if the test and the proposed interpretation were designed to support the decision procedure and there is no indication of any serious negative consequences), but even in these cases consequences merit enough attention to determine whether the intended consequences are likely to be achieved and the potential for negative consequences is indeed small. More generally, the evaluation of consequences becomes a major issue if the decisions being made are high stakes, the potential for negative consequences is substantial, and/or success in achieving the intended outcomes is doubtful. In these cases, it is important not to leap to the conclusion that the proposed use makes sense just because a supporting interpretation has been validated.

Score Interpretations and Consequences

Although negative consequences do not necessarily count against the validity of a prior underlying interpretation, they typically do trigger investigations into the causes of the negative consequences. If it is shown that the negative consequences result from defects in the test or testing procedures, the negative consequences count against the validity of the underlying score interpretation (Kane, 2010; Messick, 1989; Sackett, 1998). The negative consequences can draw attention to defects in the testing program which may not have been recognized in advance. For example, the fact that the test did not sample the full target domain of interest might not be given much attention in some contexts, but this "limitation" would become a serious defect if it led to substantial adverse impact for a particular group. In addition to undermining claims about the legitimacy of the decision rule, any defects in the testing procedures would count against the validity of the underlying interpretation of the scores. Test developers try to minimize any potential negative consequences in testing programs (e.g., differential impact due to flaws in the test) by developing content that is neutral, representative, or balanced and by employing statistical quality control based on analyses of differential item functioning (Camilli, 2006).

In general, any findings (including consequences) that are inconsistent with the proposed interpretation of test scores count against that interpretation. For example,

Guion (1998) suggests that if a test is used to hire mechanics because it purports to be a measure of flexibility in thinking about problems,

> then one consequence of its use is that high scorers are likely to think of and try alternative explanations for mechanical problems—hence, to solve more of them. If they actually do solve more problems than low scorers, it is evidence of valid measurement of the construct as well as evidence supporting the predictive hypothesis. (Guion, 1998, p. 248)

Finding that high scorers are less effective at solving novel problems would count against the predictive hypothesis, and in this case it also would count against the construct interpretation, which is closely tied to the prediction (Guion, 1998). Outcomes that are inconsistent with a proposed interpretation cast doubt on the interpretation.

Information about specific negative consequences also can help to pinpoint problems in the testing program (Kane, 2010). For example, finding that a placement system is assigning many students to courses that are too difficult for them might indicate that the cutscores need to be adjusted (a defect in the decision rule) or that the test does not give enough attention to some skills required in the courses.

In the absence of indications that they result from a defect in the testing procedure, negative consequences do not count against the validity of a prior interpretation of the test scores. However, even if they are not associated with any flaw in the test, they do count against the decision rule. If these negative consequences are not outweighed by positive consequences, they can lead to a rejection of the decision rule.

Responsibility for Evaluating Decision Rules

The two main candidates for evaluating the consequences of test use are the user (who decides how the scores are to be used) and the test developer. In some cases, the test developer and user are identical. Assuming that they are different, the test users tend to be in the best position to evaluate the consequences of their own decision rules and have a responsibility to do so (Cronbach, 1980; Taleporos, 1998).

> Though the developer of a test should help the user in any practicable way, validation is the interpeter's responsibility. A medical school, for example, has to take responsibility for the way *it* integrates the Medical College Admissions Test into its decisions. (Cronbach, 1980, p. 100)

Test users presumably know how they are using the tests: the intended outcomes, the population being tested, and the context in which the decisions will be made. The user also is in a position to spot unanticipated consequences. Test users may choose to rely on evidence and analyses provided by a test developer in evaluating potential consequences, and it would be reasonable to expect that test developers would provide support for their proposed interpretation of the scores and for any test uses that they recommend, explicitly or implicitly (Shepard, 1997), but test users always have responsibility for their choices, especially if they choose to use the test scores in ways that go beyond the uses recommended by the developer.

If a test developer markets a test as a measure of some attribute, it would seem that the developer would be responsible for justifying the claim that the test scores can be validly interpreted in terms of the attribute. In addition, if a developer claims

or suggests (with a label like "readiness") that their test can be used to achieve certain goals (e.g., effective accountability), the developer would be expected to justify these claims by evaluating the consequences of the expected score use. It also seems reasonable to expect test publishers to anticipate common uses and consequences of their tests (Green, 1998; Moss, 1998), even if these uses are not explicitly advocated by the publisher. Some consequences of high-stakes testing programs (e.g., teaching to the test) occur regularly enough to be anticipated as possibilities:

> In most instances, the test publisher will have the greatest technical capability and the most information about the likely impact of different uses of his or her test based on experience in other states or districts. Experience has shown that, if there are high stakes associated with test results, teachers will teach to the test. (Linn, 1998, p. 29)

Teaching to the test may be considered good or bad depending on the circumstances, but in high-stakes contexts its occurrence should not be considered a surprise. Some potential consequences of test use (e.g., adverse impact) are sufficiently familiar that they are routinely monitored in most high-stakes assessment programs.

Although some consequences may not be identified in advance (Green, 1998; Reckase, 1998), it seems reasonable to expect that the users and test developers will monitor the actual consequences associated with test use. The evaluation of the consequences can be a complex and contentious issue, with a range of potential consequences to be evaluated (Lane et al., 1998) and many individuals, groups, and organizations involved in the evaluation (Linn, 1998). As noted above, test users generally are in the best position to identify unintended consequences, but test publishers also have responsibility for the consequences of uses they explicitly or implicitly advocate (Green, 1998; Linn, 1998; Moss, 1998).

The Evolving Role of Intended and Unintended Consequences in Validation

Traditional definitions of validity in terms of how well a testing program achieves its goals (Cureton, 1951) necessarily raise questions about consequences, both positive and negative (Cronbach & Gleser, 1965; Linn, 1997; Moss, 1992; Shepard, 1997). As noted earlier, Kelley (1927) suggested that validity had its roots in educational consequences because score use was "intimately connected with the weal of . . . pupils" (p. 13). The focus was on both the plausibility of proposed interpretations and the appropriateness of uses, but the two issues were not clearly distinguished. Gradually, the distinction between validating an interpretation and validating a use was clarified (Guion, 1974; Linn, 1997; Messick, 1975, 1989, 1994; Moss, 1992; Shepard, 1993, 1997).

In the first edition of *Educational Measurement*, Cureton defined validity in terms of how well testing programs achieved their intended goals:

> The essential question of test validity is how well a test does the job it is employed to do. The same test may be used for several different purposes, and its validity may be high for one, moderate for another and low for a third. (Cureton, 1951, p. 621)

For the first half of the 20th century, validity was goal-driven, aimed at predicting (and to the extent possible, maximizing) some desired outcome, or criterion. It is not clear that writers like Kelley and Cureton and other testing professionals in the

first half of the 20th century entertained the notion that testing programs could have serious negative consequences; rather, the focus was on optimizing the benefits of testing.

Starting with the civil rights movement in the 1960s, concerns developed about potential social consequences of testing programs; of particular concern was differential impact on racial and ethnic minorities, women, and individuals with disabilities. Later, in evaluating the impact of large-scale standardized testing programs on educational outcomes and on educational practices, a range of potentially negative consequences was identified; these consequences generally were considered under the heading of validity (Moss, 1992). The language-testing community has given a lot of attention to the myriad ways in which testing programs can have an impact (or "washback") on schools, teaching, and learning (Alderson & Wall, 1993; Chapelle, 1999; Fulcher & Davidson, 2007).

Although they differed somewhat in emphasis, both Cronbach (1971, 1980, 1988) and Messick (1975, 1980, 1989, 1994) included both interpretive accuracy and the consequences of score-based decisions under the heading of validity. The discussion of validity in the *Standards* (1985, 1999) also has covered the goals and consequences of the testing program, as well as interpretations of test scores *per se*:

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

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

Note, however, that a finding that the test is inappropriate for a particular use generally would not count against other more appropriate uses or against a proposed interpretation of the scores *per se*. The relationship between the decision rule and the rest of the IUA follows the typical pattern; the rejection of any inference in the argument invalidates any interpretation or use that depends on that inference, but it does not invalidate more limited interpretations or uses that do not depend on the inference.

Our evolving conception of validation has included concerns about consequences since its inception (Kelley, 1927), but the range of consequences to be considered has expanded over time. Early on, the focus was on intended outcomes, particularly the outcomes of selection systems (Brogden, 1946, 1949; Cascio, 1993; Cronbach & Gleser, 1965; Taylor & Russell, 1939). More recently, adverse impact and unintended systemic effects have been recognized as serious issues that need to be addressed in evaluating high-stakes testing programs.

Including the Evaluation of Consequences in Validation

A role for consequences in the evaluation of testing programs has been recognized since the 1920s (Brogden, 1946, 1949; Cronbach, 1971, 1988; Cronbach & Gleser, 1965; Cureton, 1951; Gulliksen, 1950; Kelley, 1927; Taylor & Russell, 1939), and it is not that Messick (1981, 1989) or Cronbach (1971, 1988) added the consideration of consequences to validity (Moss, 1998; Shepard, 1997). Cronbach and Messick changed things in two major ways. First, they (mainly Messick) made the role of social consequences in validity more explicit; second, they (mainly Cronbach) gave consequences a major role in evaluating decision rules.

There seems to be broad agreement on the following propositions:

First, an evaluation of a decision rule requires an evaluation of its overall consequences for the intended population, with the positive consequences weighed against the negative consequences. Although there are differences of opinion about the details and about whether the evaluation of the consequences of decision rules should be included under the heading of validity, nobody suggests that consequences should be ignored.

Second, an evaluation of the decision rule depends on the consequences of the decision rule for the population of interest and does not focus on specific individual outcomes which generally can be addressed by test users on an individual basis (e.g., possibly under the heading of "exceptions").

Third, an argument for or against the use of a decision rule based on analyses of consequences typically depends on values (e.g., that increases in some criterion are positive or that certain changes in curricula are negative), and for the argument to be persuasive, these value assumptions have to be accepted by the stakeholders.

Fourth, even very strong evidence for an underlying score interpretation does not, in itself, justify a decision rule that is based on the scores. The justification of a decision rule requires that its positive consequences outweigh its negative consequences.

Fifth, negative consequences count against a prior underlying interpretation of scores if and only if the consequences indicate that some part of the IUA for the interpretation is not plausible.

The issue that is most in dispute can be specified fairly sharply. The question is whether the evaluation of unintended consequences (particularly social consequences, like adverse impact) should be included under the heading of "validity" or under some other heading, such as: "utility" (Lissitz & Samuelsen, 2007; Scriven, 2002), "evaluation" (Shadish, Cook, & Campbell, 2002), "informed social debate" (Maguire, Hattie, & Brian, 1994), "justification" (Cizek, 2012), "overall quality" (Borsboom et al., 2004), or as an important consideration "orthogonal to validity" (Popham, 1997). Even if one accepts the claim that the evaluation of consequences is essential to the evaluation of decision rules and the desirability of giving serious attention to consequences, it does not automatically follow that the evaluation of decision rules in terms of their consequences should be included under the heading of validity.

One option is to allow for the validation of test-score uses by including the evaluation of at least some major categories of unintended consequences in the validation of the score use. A second option focuses validation on score interpretation (and perhaps some intended consequences) and excludes unintended consequences from

validity. Under the second option, most of the consequences of score uses would be evaluated under some heading other than validity (e.g., "utility").

I am clearly in favor of the first option. I think that the evaluation of consequences of score-based decision rules should be included under the heading of validity, but in taking this position I have indicated the kinds of consequences that I think should be included (those with a potential for substantial impact in the population of interest, particularly adverse impact and systemic consequences). Negative consequences play a direct and immediate role in the evaluation of score uses, and they have a more limited and less direct role in the validation of underlying interpretations associated with score uses.

Evaluating consequences is difficult, especially in cases where there are substantial positive and negative consequences which have to be weighed against each other. But evaluating consequences is a necessary component in evaluating score-based decision procedures. The fact that an important question cannot be answered easily and unambiguously is not a good reason for not asking it. It is good to know what we do not know or are not sure of.

Decision rules are evaluated in terms of their overall consequences (e.g., in terms of their expected utility over possible outcomes), and our inability to specify all of the potential consequences in precise quantitative terms does not remove the need to evaluate consequences. The range of consequences to be considered has evolved over time, but it is still limited to a few major concerns, particularly intended outcomes, adverse impact, and systemic effects, or washback. The consequences that are relevant to validity share two characteristics: they are evaluated over some population or some subpopulation, and they are widely accepted as legitimate public concerns.

Some authors have argued for a much more limited definition of validity that primarily involves score interpretations and excludes the evaluation of most consequences, especially social consequences (Borsboom et al., 2004; Cizek, 2012; Green, 1998; Mehrens, 1997; Popham, 1997; Reckase, 1998; Sackett, 1998; Scriven, 2002; Wiley, 1991). I do not have the space to analyze these different positions, but I do want to make a few general points about the implications of not addressing consequences as part of validation.

The most commonly made case for separating consequences from validity is that this will simplify validity or keep it from getting more complicated; this case is dubious. If we do not include the evaluation of consequences under validity, we would seem to have two options. First, the measurement community could accept responsibility for evaluating intended consequences and major sources of unintended consequences (e.g., adverse impact, systemic effects) but address these questions under some other heading (Cizek, 2012). It does not seem to me that we gain much by doing this, and we may exacerbate some existing problems. Assuming that we are going to take responsibility for both the validity of a proposed interpretation and for the evaluation of the consequences of score use (under some other heading), it is not clear that we have simplified the overall task (especially if the subtasks are going to be carried out by different groups that will have to coordinate their efforts). Most of the work on the intended consequences of testing programs has been lodged squarely under the heading of validity, particularly under the heading of criterion-related validity evidence, and most research on unintended consequences also has

been associated with validity (Camilli, 2006; Cole & Moss, 1989; Willingham and Cole, 1997; Xi, 2010). Given this history, if we separate the evaluation of consequences from validity we run the risk of encouraging a tendency to beg the question of consequences in which a validated interpretation (e.g., in terms of an observable attribute or trait) will be taken to justify (or validate) the proposed use of the scores.

Second, we can restrict our attention to score interpretations *per se* and assume that some other group of stakeholders will take responsibility for evaluating the consequences of score use, particularly the unintended consequences of score use (e.g., adverse impact, systemic effects). However, the measurement community has a long tradition of evaluating intended and unintended consequences (Camilli, 2006; Crooks, 1988; Frederiksen, 1984; Moss, 1992). Given that it is the measurement community that has the most experience with the operation of testing programs and with the trade-offs required in such programs, leaving the evaluation of most consequences to others seems problematic.

Barring a seismic shift in the legal and social context, testing programs are going to be evaluated in terms of their overall consequences, including some social consequences (e.g., fairness and systemic effects).

> The courts have taken a more practical, less theoretical view on validity and tend to emphasize evidence based on test content and testing consequences. (Sireci and Parker, 2006, p. 27)

If the measurement community were to give less attention to the role of consequences in evaluating score-based decision rules, other groups (e.g., the courts, legislatures, the media) would take up the burden. All test users are subject to legal and ethical restraints. They cannot ignore consequences, even if they would prefer to do so.

On a more fundamental level, we all are expected to take responsibility for what we do. Many tests are specifically designed to serve a particular function (e.g., licensure, placement, program evaluation) and are evaluated in terms of their suitability for this function. If validity is to be considered "the most fundamental consideration in developing and evaluating tests" (American Educational Research Association et al., 1999, p. 9), it needs to address the suitability of the test for its intended function. Those who make decisions about test use necessarily bear much of the responsibility for evaluating the suitability of their intended score use and the suitability of the test for the use, but a test developer who provides a test to the user (e.g., under contract) shares this responsibility. In particular, test developers who suggest that a test can be used in a particular way have an obligation to support the claims that they make.

Recap of Section "Score Uses"

I have argued for a somewhat limited but important role for consequences in validation. The role is limited in several ways. First, the consequences to be considered under the heading of validity are limited to those that either apply to the population of test takers as a whole or to a substantial subset of the population. Serious negative consequences for individuals would need to be addressed, but they would be

addressed as special cases (or as exceptions, in Toulmin's terminology). The decision rule is evaluated in terms of how well it works in general over the population, just as we evaluate the effectiveness of regression-based predictions and the overall dependability of generalizations in terms of standard errors defined over the population.

Second, negative consequences do not generally count against an underlying interpretation, unless the nature of the consequences indicates some defect in the test or in the IUA supporting the interpretation. A finding of serious negative consequences generally would trigger a careful review of the evidence for the validity of the interpretation on which the use of the test scores is based, but it would not in itself count against the validity of the interpretation (Messick, 1989).

Third, support for an interpretation on which a score use is based does not validate the score use. As illustrated by Popham (1997), a well-designed test with a clear interpretation can be used in inappropriate (even ridiculous) ways. So even strong support for an interpretation does not justify any score use. In addition, it is necessary to make the case that the proposed use is likely to achieve its goals and is not likely to have serious negative consequences.

A role for consequences in validation is particularly important because we are in the habit of running the interpretations and uses of scores together in our validity analyses. For example, criterion-related validity evidence is relevant to score interpretations (as predictors of future performance), to their use (e.g., in selection or placement), and to the evaluation of the intended outcomes (or utilities) of these uses (Brogden, 1946; Cronbach & Gleser, 1965).

Excluding consequences from validity is likely to encourage score users to conclude that a score use is automatically justified by evidence for the validity of an underlying score interpretation, and perhaps by some analyses of intended consequences. Assuming that validity is the bottom line in evaluating testing programs, the question of consequences (particularly unintended consequences) should not be ignored (or begged) in evaluating score-based decisions.

Concluding Remarks

Our conception of validity has expanded gradually over the last 120 years, starting from a very general content-based model and adding the criterion model, trait and factor models, the construct model, and concerns about fairness and other consequences. Along the way, a host of specific "validities" were developed as efforts were made to evaluate particular interpretations and uses of test scores and as new kinds of interpretations and uses of test scores were proposed. These developments greatly enriched our conception of validity, but they also tended to fragment the validity framework as more and more types of validity (or types of validity evidence) were introduced.

The tendency toward expansion and fragmentation led to ongoing attempts to unify the field through general definitions of validity and general approaches to validation that are designed to serve as umbrellas under which specific methods and models could reside (sometimes uncomfortably). By 1950, the criterion model had emerged as the gold standard, but this model did not provide a good home for attributes for which criterion measures were not available; Cronbach and Meehl (1955)

introduced the construct validity model to fill the gap. Loevinger (1957) suggested that the construct model could provide a general framework for validity, and this idea gradually gained traction, but the original "strong form" of construct validity was not feasible in most cases and the "weak form" was too loose to be satisfactory. Messick (1989) strengthened the philosophical underpinnings of the field by proposing a unified model based on general principles of construct validity, but this unitary formulation of validity was quite abstract and expansive and, as a result, hard to implement.

The argument-based approach to validation (Cronbach, 1988; House, 1980; Kane, 1992; Shepard, 1993) takes a more pragmatic approach to unification. At its core is the basic scientific and social principle that public claims should be supported by appropriate evidence. The argument-based approach suggests that "to validate an interpretation or use of measurements is to evaluate the rationale, or argument, for the proposed conclusions and decisions" (Kane, 2006, p. 17). It calls for "a clear statement of the proposed interpretations and uses and a critical evaluation of these interpretations and uses" (Kane, 2006, p. 17).

The inferences and assumptions on which the proposed interpretations and uses depend are to be laid out as an IUA, and the coherence of this argument and the plausibility of its inferences and assumptions are to be evaluated in a validity argument. The IUA provides a framework for validation that delimits the claims that need to be checked, and as a result, validation does not need to be considered an open-ended, never-ending process. The validation of ambitious interpretations and uses that make strong claims can require an extensive set of empirical studies and analyses, but even in these cases, the validation would be limited to the evidence needed to evaluate the IUA and nothing more. The validation of simple interpretations (e.g., in terms of observable attributes) would require a very limited range of evidence (e.g., mainly evidence for the appropriateness of scoring, the generalizability of the scores, and extrapolation to the target domain). It is the claims and decisions based on the test scores that are to be validated.

A strongly supported interpretation does not in itself justify a decision rule. It is certainly possible to conceive of an accurate measure of some attribute being used badly. It also is easy to conceive of a well-designed decision process that fails because of an inadequate test, one that does not "support the interpretation . . . entailed by proposed uses of tests" (American Educational Research Association et al., 1999, p. 9). A chain of reasoning is only as strong as its weakest link (Crooks et al., 1996), and strong evidence for part of an argument does not compensate for weaknesses in other parts of the argument.

Although the argument-based approach is systematic, it is not automatic or algorithmic. The IUA has to be developed, and its inferences and assumptions have to be evaluated. Specifying an IUA requires care and insight, and developing the evidence to support the claims being made typically requires technical skill and ingenuity, but we do not need to reinvent the wheel for each validation effort. Some inferences and assumptions (e.g., scoring, generalization, scaling, extrapolation, and decisions) are commonly found in IUAs, and we have a lot of experience in evaluating these inferences and assumptions. Validation is not easy, but generally it is possible to do a reasonably good job of validation with a manageable level of effort (with the

requirements determined by the extent and complexity of the proposed interpretation and use and the stakes associated with the testing program).

The argument-based approach provides explicit, albeit contingent, guidance for validation; the evidence required for validation is the evidence needed to evaluate the proposed IUA. We have some flexibility in what we include in the IUA, but once it is specified, the requirements for validation are well defined. The IUA provides a template for validation and a basis for evaluating the adequacy of the validity argument. If the IUA is coherent and complete and if all of its inferences and assumptions are plausible given the evidence, the proposed interpretations and uses can be considered valid. If the IUA is incomplete or if some of its inferences or assumptions are shaky, the validity argument is inadequate. A failure to specify the proposed interpretations and uses clearly and in some detail makes a fully adequate validation difficult if not impossible because implicit inferences and assumptions would not be evaluated. Most fallacies in presumptive reasoning involve the tacit acceptance of doubtful assumptions, and an important function of external critics and alternative interpretations is to make these assumptions explicit.

The argument-based approach to validation is contingent. The validity argument is based on the IUA. Score interpretations that make very modest claims (e.g., narrowly defined observable attributes) do not require much evidence for validation. Ambitious interpretations and uses (e.g., those involving causal claims or high stakes decisions) can require an extended research program for their validation. The argument-based approach does not provide a single, simple recipe or checklist for validation, but it does provide a generally applicable methodology for validation. The proposed score-based interpretations and uses are to be developed as an IUA, the coherence and completeness of this argument is to be evaluated, and its inferences and assumptions are to be checked.

Note

[1]In this paper, I have used the expression "interpretation/use argument" (or IUA) in place of my earlier terminology, "interpretive argument," in order to recognize the importance of score uses in determining score interpretations and to acknowledge the importance of score uses (as well as contexts and test-taker populations) in validation.

References

Alderson, J., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, *14*(2), 115–129.

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

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (1974). *Standards for educational and psychological tests and manuals*. Washington, DC: American Educational Research Association.

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

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

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

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 9–13). Hillsdale, NJ: Lawrence Erlbaum.

Bachman, L. (2002). Alternative interpretations of alternative assessments: Some validity issues in educational performance assessments. *Educational Measurement: Issues and Practice*, *21*(3), 5–18.

Bachman, L. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, *2*, 1–34.

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

Blair, J. (1995). Informal logic and reasoning in evaluation. In D. Fournier (Ed.), *Reasoning in evaluation: Inferential links and leaps* (pp. 71–80). San Francisco, CA: Jossey-Bass.

Bonner, S. (2005, April). *Investigating the cognitive processes in responding to MBE questions*. Paper presented at the meeting of the National Council on Measurement in Education, Montreal, Canada.

Boorstin, D. (1983). *The discoverers*. New York, NY: Random House.

Borsboom, D., Cramer, A., Kievit, R., Zand Scholten, A., & Franic, S. (2009). The end of construct validity. In R. Lissitz (Ed.), *The concept of validity* (pp. 135–170). Charlotte, NC: Information Age Publishers.

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

Brennan, R. (2001a). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, *36*, 295–317.

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

Bridgeman, P. (1927). *The logic of modern physics*. New York, NY: Macmillan.

Brogden, H. E. (1946). On the interpretation of the correlation coefficient as a measure of predictive efficiency. *Journal of Educational Psychology*, *37*, 65–76.

Brogden, H. E. (1949). When testing pays off. *Personnel Psychology*, *2*, 171–183.

Camilli, G. (2006). Test fairness. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221–256). Westport, CT: American Council on Education and Praeger.

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

Campion, M. (1983). Personnel selection for physically demanding jobs: Review and recommendations. *Personnel Psychology*, *36*, 527–550.

Carroll, L. (2010). Through the looking glass. (Original work published 1871.) Retrieved March 26, 2010, from www.literature.org/authors/carroll-lewis/through-the-looking-glass

Cascio, W. (1993). Assessing the utility of selection decisions: Theoretical and practical considerations. In N. Schmitt & W. Borman (Eds.), *Personnel selection in organizations* (pp. 310–340). San Francisco, CA: Jossey-Bass.

Cascio, W., Jacobs, R., & Silva, J. (2010). Validity, utility, and adverse impact: Practical implications from 30 years of data. In J. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 271–288). New York, NY: Routledge.

Chapelle, C. A. (1999). Validation in language assessment. *Annual Review of Applied Linguistics*, *19*, 254–272.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2008). *Building a validity argument for the test of English as a foreign language*. New York, NY: Routledge.

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

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

Clauser, B. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological Measurement*, *24*, 310–324.

Clauser, B., Margolis, M., & Case, S. (2006). Testing for licensure and certification in the professions. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 701–731). Westport, CT: American Council on Education and Praeger.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 201–219). New York, NY: American Council on Education and Macmillan.

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

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

Cronbach, L. J. (1980). Validity on parole: How can we go straight? *New Directions for Testing and Measurement: Measuring Achievement Over a Decade*, *5*, 99–108.

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

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: University of Illinois Press.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions*. Urbana: University of Illinois Press.

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

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

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York, NY: Irvington.

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

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

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–694). Washington, DC: American Council on Education.

Debra P. v. Turlington (1981). 644 F. 2d 397 (5th Cir.) 564 F. Supp. 177 (M. D. Fla. 1983).

De Corte, W., Lievens, F., & Sackett, P. R. (2007). Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. *Journal of Applied Psychology*, *92*, 1380–1393.

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

Ebel, R. (1966). The social consequences of educational testing. In A. Anastasi (Ed.), *Testing problems in perspective: Twenty-fifth anniversary volume of topical readings from the invitational conference in testing problems* (pp. 18–28). Washington, DC: American Council on Education.

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

Embretson, S. (1984). A general multicomponent latent trait model for measuring learning and change. *Psychometrika*, *49*, 175–186.

Embretson, S. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*, 380–396.

Embretson, S., & McCollam, K. (2000). Psychometric approaches to understanding and measuring intelligence. In R. Sternberg (Ed.), *Handbook of intelligence* (pp. 423–444). Cambridge, UK: Cambridge University Press.

Ennis, R. H. (1973). Operational definitions. In H. Brody, R. Ennis, & L. Krimerman (Eds.), *Philosophy of educational research* (pp. 650–669). New York, NY: Wiley.

Equal Employment Opportunity Commission (EEOC), Civil Service Commission, Department of Labor, & Department of Justice (1979). Adoption by four agencies of Uniform Guidelines on employee selection procedures. *Federal Register*, *43*, 38290–38315.

Feest, U. (2005). Operationism in psychology: What the debate is about, what the debate should be about. *Journal of the History of the Behavioral Sciences*, *49*(2), 131–149.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York, NY: American Council on Education and Macmillan.

Flockton, L., & Crooks, T. (2002). *Social studies assessment results 2001*. Dunedin, New Zealand: Educational Assessment Research Unit, University of Otago.

Frederiksen, N. (1984). The real test bias. *American Psychologist*, *39*, 193–202.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, *18*(9), 27–32

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. New York, NY: Routledge.

Galison, P. (1987). *How do experiments end?* Chicago, IL: University of Chicago Press.

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

Griggs v. Duke Power Company (1971). 401 U.S. 424.

Guion, R. (1974). Open a window: Validities and values in psychological measurement. *American Psychologist*, *29*, 287–296.

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

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

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

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley.

Gutman, A. (2005). Adverse impact: Judicial, regulatory, and statutory authority. In F. L. Landy (Ed.) *Employment discrimination litigation* (pp. 20–46). San Francisco, CA: Jossey-Bass.

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

Haertel, E. H. (2006). Reliability. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: American Council on Education and Praeger.

Hamilton, L. (2003). Assessment as a policy tool. *Review of Research in Education*, *27*, 25–68.

Hansen, H., & Pinto, R. (1995). *Fallacies, classical and contemporary readings*. University Park: Pennsylvania State University Press.

Herman, J. L., & Golan, S. (1993). The effects of standardized testing on teaching and schools. *Educational Measurement: Issues and Practice*, *12*(4), 20–25.

Hershey, J., & Asch, D. (2001). A change of heart: Unexpected responses to medical testing. In S. Hoch, H. Kumreuthen, & R. Gunther (Eds.), *Wharton on making decisions* (pp. 225–242). New York, NY: Wiley.

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

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

Jackson, A. S. (1994). Pre-employment physical evaluation. *Exercise and Sport Science Review*, *22*, 53–90.

Jaeger, R. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–534). New York, NY: American Council on Education and Macmillan.

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

Kane, M. (1986). The future of testing for licensure and certification examinations. In B. Plake & J. Will (Eds.), *The future of testing* (pp. 145–181). Hillsdale, NJ: Lawrence Erlbaum.

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

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

Kane, M. (2002a). Inferences about variance components and reliability-generalizability coefficients in the absence of random sampling. *Journal of Educational Measurement*, *32*, 165–181.

Kane, M. (2002b). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, *21*(1), 31–41.

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

Kane, M. (2010). *Errors of measurement, theory, and public policy*. Princeton, NJ: Educational Testing Service.

Kane, M. (2011). The errors of our ways. *Journal of Educational Measurement*, *48*, 12–30.

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

Kelley, T. (1927). *Interpretation of educational measurements*. Yonkers, NY: World Book Co.

Kolen, M. (2006). Scaling and norming. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155–220). Westport, CT: American Council on Education and Praeger.

Koretz, D. M., & Hamilton, L. S. (2006). Testing for accountability in K-12. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 531–578). Westport, CT: American Council on Education and Praeger.

Lakatos, I. (1970). Falsification and the methodology of scientific research programs. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91–196). London, UK: Cambridge University Press.

Lane, S., Parke, C., & Stone, C. (1998). A framework for evaluating the consequences of assessment programs. *Educational Measurement: Issues and Practice*, *17*(2), 24–28.

Lane, S., & Stone, C. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and Practice*, *21*(1), 23–30.

Lane, S., & Stone, C. (2006). Performance assessment. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387–431). Westport, CT: American Council on Education and Praeger.

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

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

Linn, R. L. (2005). Conflicting demands of No Child Left Behind and state systems: Mixed messages about school performance. *Education Policy Analysis Archives*, *13*(33), (http://epaa.asu.edu/epaa/v13n33/, accessed June 28, 2005).

Linn, R. L. (2009). The concept of validity in the context of NCLB. In R. Lissitz (Ed.), *The concept of validity* (pp. 195–212). Charlotte, NC: Information Age Publishers.

Lissitz, R., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*, 437–448.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Madaus, G. F. (1988). The influences of testing on the curriculum. In L. N. Tanner (Ed.), *Critical issues in curriculum* (pp. 83–121). Chicago, IL: University of Chicago Press.

Maguire, T., Hattie, J., & Brian, H. (1994). Construct validity and achievement assessment. *The Alberta Journal of Educational Research*, *40*(2), 109–126.

Marion, S., & Perie, M. (2009). In introduction to validity arguments in alternate assessments. In W. D. Schaeffer & R. W. Lissitz (Eds), *Alternate assessment: Proceedings from the 8th Annual MARCES Conference* (pp. 113–126). Baltimore, MD: Brookes Publishing.

McNeil, L. (2005). Faking equity: High-stakes testing and the education of Latino youth. In A. Valenzuela (Ed.), *Leaving children behind: How "Texas-style" accountability fails Latino youth* (pp. 57–111). Albany: State University of New York Press.

Meehl, P. (1950). On the circularity of the law of effect. *Psychological Bulletin*, *47*, 52–75.

Meehl, P. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis: University of Minnesota.

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

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, *30*, 955–966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*, 1012–1027.

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

Messick, S. (1982). The values of ability testing: Implications of multiple perspectives about criteria and standards. *Educational Measurement: Issues and Practice*, *1*(3), 9–12.

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

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

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*, 13–23.

Mill, J. S. (2002 [1843]). *A system of logic*. Honolulu, HI: University Press of the Pacific.

Mislevy, R. (1996). Test theory reconceived. *Journal of Educational Measurement*, *33*, 379–416.

Mislevy, R. (2004). Can there be reliability without "reliability?" *Journal of Educational and Behavioral Statistics*, *29*, 241–244.

Mislevy, R. (2006). Cognitive psychology and educational assessment. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257–305). Westport, CT: American Council on Education and Praeger.

Mislevy, R. (2009). Validity from the perspective of model-based reasoning. In R. Lissitz (Ed.), *The concept of validity* (pp. 83–108). Charlotte, NC: Information Age Publishers.

Mislevy, R., Steinberg, L., & Almond, R. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3–62.

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

Moss, P. (1994). Can there be validity without reliability? *Educational Researcher*, *23*, 5–12.

Moss, P. (1995). Themes and variations in validity theory. *Educational Measurement: Issues and Practice*, *4*(2), 5–13.

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

Moss, P. (2007). Reconstructing validity. *Educational Researcher*, *36*, 470–476.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment.* J. Pellegrino, N. Chudowski, & R. Glaser (Eds.), Committee on the Foundations of Assessment. Board on Testing and Assessment, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Research Council (2007). *Lessons learned about testing: Ten years of work at the National Research Council*. [Booklet.] Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press. Retrieved January 28, 2013, from http://www7.nationalacademies.org/dbasse/Lessons_Learned_Brochure_pdf.pdf

No Child Left Behind (NCLB) Act (2002). Pub. L. No. 107–110, 115 Stat. 1435.

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

Petersen, N. (2007). Equating: Best practices and challenges to best practices. In N. Dorans, M. Pommerich, & P. Holland (Eds.), *Linking and aligning scores and scales* (pp. 59–72). New York, NY: Springer-Verlag.

Peterson, M. (2009). *An introduction to decision theory*. Cambridge, UK: Cambridge University Press.

Phillips, D. (2007). Adding complexity: Philosophical perspectives on the relationship between evidence and policy. In P. Moss (Ed.), *Evidence and decision making* (pp. 376–402). Malden, MA: Blackwell.

Pinto, R. (2001). *Argument, inference and dialectic*. Dordrect, The Netherlands: Kluwer Academic Publishers.

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

Popper, K. R. (1962). *Conjecture and refutation: The growth of scientific knowledge*. New York, NY: Basic Books.

Pyburn, K., Ployhart, R., & Kravitz, D. (2008). The diversity-validity dilemma: Overview and legal context. *Personnel Psychology*, *61*, 143–151.

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

Ryan, K. (2002). Assessment validation in the context of high-stakes testing assessment. *Educational Measurement: Issues and Practice*, *21*(1), 7–15.

Sackett, P. R. (1998). Performance assessment in education and professional certification: Lessons for personnel selection? In M. D. Hakel (Ed.), *Beyond multiple choice: Evaluating alternatives for traditional testing for selection* (pp. 113–129). Mahwah, NJ: Lawrence Erlbaum.

Sackett, P., De Corte, W., & Lievens, F. (2010). Decision aids for addressing the validity-adverse impact trade-off. In J. Outtz (Ed.), *Adverse impact: Implications for organizational staffing and high stakes selection* (pp. 453–472). New York, NY: Routledge.

Scriven, M. (1987). Validity in personnel evaluation. *Journal of Personnel Evaluation in Education*, *1*, 9–23.

Scriven, M. (2002). Assessing six assumptions in assessment. In H. Braun, D. Jackson, & D. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 255–275). Mahwah, NJ: Lawrence Erlbaum.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Shepard, L. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of research in education* (pp. 405–450). Washington, DC: American Educational Research Association.

Shepard, L. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, *16*(2), 5–24.

Sireci, S. (1998). The construct of content validity. *Social Indicators Research*, *45*, 83–117.

Sireci, S. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. Lissitz (Ed.), *The concept of validity* (pp. 19–38). Charlotte, NC: Information Age Publishers.

Sireci, S. G., & Green, P. C. (2000). Legal and psychometric criteria for evaluating teacher certification tests. *Educational Measurement: Issues and Practice*, *19*(1), 22–31.

Sireci, S. G., & Parker, P. (2006). Validity on trial: Psychometric and legal conceptualizations of validity. *Educational Measurement: Issues and Practice*, *2*(3), 27–34.

Sireci, S., Scarpati, S., & Li, S. (2005). Test accommodations for students with disabilities: An analysis of the interaction hypothesis. *Review of Educational Research*, *75*, 457–490.

Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology*, *15*, 201–292.

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

Taylor, H., & Russell, J. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology*, *23*, 565–578.

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

Thorndike, E. L. (1918). Individual differences. *Psychological Bulletin*, *15*, 148–159.

Toulmin, S. (1958). *The uses of argument*. Cambridge, UK: Cambridge University Press.

Toulmin, S. (2001). *Return to reason*. Cambridge, MA: Harvard University Press.

von Mayrhauser, R. T. (1992). The mental testing community and validity: A prehistory. *American Psychologist*, *47*, 244–253.

Walton, D. (1989). *Informal logic: A handbook for critical argumentation*. Cambridge, UK: Cambridge University Press.

Wigdor, A., & Garner, W. (1982). *Ability testing: Uses, consequences, and controversies*. Washington, DC: National Academies Press.

Wiley, D. (1991). Test validity and invalidity reconsidered. In R. Snow & D. Wiley (Eds.), *Improving inquiry in social science* (pp. 75–107). Hillsdale, NJ: Lawrence Erlbaum.

Willingham, W., & Cole, N. (1997). *Gender and fair assessment*. Mahwah, NJ: Lawrence Erlbaum.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*(2), 147–170.

Yen, W., & Fitzpatrick, A. (2006). Item response theory. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport, CT: American Council on Education and Praeger.

Zumbo, B. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Vol. 26: Psychometrics* (pp. 45–79). Amsterdam, The Netherlands: Elsevier Science.

Zumbo, B. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. Lissitz (Ed.), *The concept of validity* (pp. 65–82). Charlotte, NC: Information Age Publishers.

Author

MICHAEL T. KANE serves as the Samuel J. Messick Chair in Validity at Educational Testing Service, Rosedale Road, Princeton, NJ 08541; mkane@ets.org. His primary research interests include validity theory, generalizability theory, licensure and certification testing, and standard setting.