Validity of Psychological Assessment

Validation of Inferences From Persons' Responses and Performances as Scientific Inquiry Into Score Meaning

Samuel Messick
Educational Testing Service

The traditional conception of validity divides it into three separate and substitutable types—namely, content, criterion, and construct validities. This view is fragmented and incomplete, especially because it fails to take into account both evidence of the value implications of score meaning as a basis for action and the social consequences of score use. The new unified concept of validity interrelates these issues as fundamental aspects of a more comprehensive theory of construct validity that addresses both score meaning and social values in test interpretation and test use. That is, unified validity integrates considerations of content, criteria, and consequences into a construct framework for the empirical testing of rational hypotheses about score meaning and theoretically relevant relationships, including those of an applied and a scientific nature. Six distinguishable aspects of construct validity are highlighted as a means of addressing central issues implicit in the notion of validity as a unified concept. These are content, substantive, structural, generalizability, external, *and* consequential *aspects of construct validity. In effect, these six aspects function as general validity criteria or standards for all educational and psychological measurement, including performance assessments, which are discussed in some detail because of their increasing emphasis in educational and employment settings.*

Validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment (Messick, 1989b). Validity is not a property of the test or assessment as such, but rather of the meaning of the test scores. These scores are a function not only of the items or stimulus conditions, but also of the persons responding as well as the context of the assessment. In particluar, what needs to be valid is the meaning or interpretation of the score; as well as any implications for action that this meaning entails (Cronbach, 1971). The extent to which score meaning and action implications hold across persons or population groups and across settings or contexts is a persistent and perennial empirical question. This is the main reason that validity is an evolving property and validation a continuing process.

The Value of Validity

The principles of validity apply not just to interpretive and action inferences derived from test scores as ordinarily conceived, but also to inferences based on any means of observing or documenting consistent behaviors or attributes. Thus, the term *score* is used generically in its broadest sense to mean any coding or summarization of observed consistencies or performance regularities on a test, questionnaire, observation procedure, or other assessment devices such as work samples, portfolios, and realistic problem simulations.

This general usage subsumes qualitative as well as quantitative summaries. It applies, for example, to behavior protocols, to clinical appraisals, to computerized verbal score reports, and to behavioral or performance judgments or ratings. Scores in this sense are not limited to behavioral consistencies and attributes of persons (e.g., persistence and verbal ability). Scores may also refer to functional consistencies and attributes of groups, situations or environments, and objects or institutions, as in measures of group solidarity, situational stress, quality of artistic products, and such social indicators as school dropout rate.

Hence, the principles of validity apply to all assessments, including performance assessments. For example, student portfolios are often the source of inferences—not just about the quality of the included products but also about the knowledge, skills, or other attributes of the student—and such inferences about quality and constructs need to meet standards of validity. This is important because performance assessments, although long a staple of industrial and military applications, are now touted as purported instruments of standards-based education reform because they promise positive consequences for teaching and learning. Indeed, it is precisely because of

Editor's note. Samuel M. Turner served as action editor for this article.

This article was presented as a keynote address at the Conference on Contemporary Psychological Assessment, Arbetspsykologiska Utvecklingsinstitutet, June 7–8, 1994, Stockholm, Sweden.

Author's note. Acknowledgments are gratefully extended to Isaac Bejar, Randy Bennett, Drew Gitomer, Ann Jungeblut, and Michael Zieky for their reviews of various versions of the manuscript.

Correspondence concerning this article should be addressed to Samuel Messick, Educational Testing Service, Princeton, NJ 08541.

Vol. 50, No. 9, 741–749

Samuel Messick
Photo by William Monachan, Educational Testing Service, Princeton, NJ.

such politically salient potential consequences that the validity of performance assessment needs to be systematically addressed, as do other basic measurement issues such as reliability, comparability, and fairness. The latter reference to fairness broaches a broader set of equity issues in testing that includes fairness of test use, freedom from bias in scoring and interpretation, and the appropriateness of the test-based constructs or rules underlying decision making or resource allocation, that is, distributive justice (Messick, 1989b).

These issues are critical for performance assessment—as they are for all educational and psychological assessment—because validity, reliability, comparability, and fairness are not just measurement principles, they are social values that have meaning and force outside of measurement whenever evaluative judgments and decisions are made. As a salient social value, validity assumes both a scientific and a political role that can by no means be fulfilled by a simple correlation coefficient between test scores and a purported criterion (i.e., classical criterion-related validity) or by expert judgments that test content is relevant to the proposed test use (i.e., traditional content validity).

Indeed, validity is broadly defined as nothing less than an evaluative summary of both the evidence for and the actual—as well as potential—consequences of score interpretation and use (i.e., construct validity conceived comprehensively). This comprehensive view of validity integrates considerations of content, criteria, and consequences into a construct framework for empirically testing rational hypotheses about score meaning and utility. Therefore, it is fundamental that score validation is an empirical evaluation of the meaning and consequences of measurement. As such, validation combines scientific inquiry with rational argument to justify (or nullify) score interpretation and use.

Comprehensiveness of Construct Validity

In principle as well as in practice, construct validity is based on an integration of any evidence that bears on the interpretation or meaning of the test scores—including content- and criterion-related evidence—which are thus subsumed as part of construct validity. In construct validation the test score is not equated with the construct it attempts to tap, nor is it considered to define the construct, as in strict operationism (Cronbach & Meehl, 1955). Rather, the measure is viewed as just one of an extensible set of indicators of the construct. Convergent empirical relationships reflecting communality among such indicators are taken to imply the operation of the construct to the degree that discriminant evidence discounts the intrusion of alternative constructs as plausible rival hypotheses.

A fundamental feature of construct validity is construct representation, whereby one attempts to identify through cognitive-process analysis or research on personality and motivation the theoretical mechanisms underlying task performance, primarily by decomposing the task into requisite component processes and assembling them into a functional model or process theory (Embretson, 1983). Relying heavily on the cognitive psychology of information processing, construct representation refers to the relative dependence of task responses on the processes, strategies, and knowledge (including metacognitive or self-knowledge) that are implicated in task performance.

Sources of Invalidity

There are two major threats to construct validity: In the one known as *construct underrepresentation,* the assessment is too narrow and fails to include important dimensions or facets of the construct. In the threat to validity known as *construct-irrelevant variance,* the assessment is too broad, containing excess reliable variance associated with other distinct constructs as well as method variance such as response sets or guessing propensities that affects responses in a manner irrelevant to the interpreted construct. Both threats are operative in all assessments. Hence a primary validation concern is the extent to which the same assessment might underrepresent the focal construct while simultaneously contaminating the scores with construct-irrelevant variance.

There are two basic kinds of construct-irrelevant variance. In the language of ability and achievement testing, these might be called *construct-irrelevant difficulty* and *construct-irrelevant easiness.* In the former, aspects of the task that are extraneous to the focal construct make the task irrelevantly difficult for some individuals or groups. An example is the intrusion of undue reading comprehension requirements in a test of subject matter knowledge. In general, construct-irrelevant difficulty leads to construct scores that are invalidly low for those individuals adversely affected (e.g., knowledge scores of poor readers or examinees with limited English proficiency). Of course, if concern is solely

with criterion prediction and the criterion performance requires reading skill as well as subject matter knowledge, then both sources of variance would be considered criterion-relevant and valid. However, for score interpretations in terms of subject matter knowledge and for any score uses based thereon, undue reading requirements would constitute construct-irrelevant difficulty.

Indeed, construct-irrelevant difficulty for individuals and groups is a major source of bias in test scoring and interpretation and of unfairness in test use. Differences in construct-irrelevant difficulty for groups, as distinct from construct-relevant group differences, is the major culprit sought in analyses of differential item functioning (Holland & Wainer, 1993).

In contrast, construct-irrelevant easiness occurs when extraneous clues in item or task formats permit some individuals to respond correctly or appropriately in ways irrelevant to the construct being assessed. Another instance occurs when the specific test material, either deliberately or inadvertently, is highly familiar to some respondents, as when the text of a reading comprehension passage is well-known to some readers or the musical score for a sight reading exercise invokes a well-drilled rendition for some performers. Construct-irrelevant easiness leads to scores that are invalidly high for the affected individuals as reflections of the construct under scrutiny.

The concept of construct-irrelevant variance is important in all educational and psychological measurement, including performance assessments. This is especially true of richly contextualized assessments and so-called "authentic" simulations of real-world tasks. This is the case because "paradoxically, the complexity of context is made manageable by contextual clues" (Wiggins, 1993, p. 208). And it matters whether the contextual clues that people respond to are construct-relevant or represent construct-irrelevant difficulty or easiness.

However, what constitutes construct-irrelevant variance is a tricky and contentious issue (Messick, 1994). This is especially true of performance assessments, which typically invoke constructs that are higher order and complex in the sense of subsuming or organizing multiple processes. For example, skill in communicating mathematical ideas might well be considered irrelevant variance in the assessment of mathematical knowledge (although not necessarily vice versa). But both communication skill and mathematical knowledge are considered relevant parts of the higher-order construct of mathematical power, according to the content standards delineated by the National Council of Teachers of Mathematics (1989). It all depends on how compelling the evidence and arguments are that the particular source of variance is a relevant part of the focal construct, as opposed to affording a plausible rival hypothesis to account for the observed performance regularities and relationships with other variables.

A further complication arises when construct-irrelevant variance is deliberately capitalized upon to produce desired social consequences, as in score adjustments for minority groups, within-group norming, or sliding band procedures (Cascio, Outtz, Zedeck, & Goldstein, 1991; Hartigan & Wigdor, 1989; Schmidt, 1991). However, recognizing that these adjustments distort the meaning of the construct as originally assessed, psychologists should distinguish such controversial procedures in applied testing practice (Gottfredson, 1994; Sackett & Wilk, 1994) from the valid assessment of focal constructs and from any score uses based on that construct meaning. Construct-irrelevant variance is always a source of invalidity in the assessment of construct meaning and its action implications. These issues portend the substantive and consequential aspects of construct validity, which are discussed in more detail later.

Sources of Evidence in Construct Validity

In essence, construct validity comprises the evidence and rationales supporting the trustworthiness of score interpretation in terms of explanatory concepts that account for both test performance and score relationships with other variables. In its simplest terms, construct validity is the evidential basis for score interpretation. As an integration of evidence for score meaning, it applies to any score interpretation—not just those involving so-called "theoretical constructs." Almost any kind of information about a test can contribute to an understanding of score meaning, but the contribution becomes stronger if the degree of fit of the information with the theoretical rationale underlying score interpretation is explicitly evaluated (Cronbach, 1988; Kane, 1992; Messick, 1989b). Historically, primary emphasis in construct validation has been placed on internal and external test structures—that is, on the appraisal of theoretically expected patterns of relationships among item scores or between test scores and other measures.

Probably even more illuminating in regard to score meaning are studies of expected performance differences over time, across groups and settings, and in response to experimental treatments and manipulations. For example, over time one might demonstrate the increased scores from childhood to young adulthood expected for measures of impulse control. Across groups and settings, one might contrast the solution strategies of novices versus experts for measures of domain problem-solving or, for measures of creativity, contrast the creative productions of individuals in self-determined as opposed to directive work environments. With respect to experimental treatments and manipulations, one might seek increased knowledge scores as a function of domain instruction or increased achievement motivation scores as a function of greater benefits and risks. Possibly most illuminating of all, however, are direct probes and modeling of the processes underlying test responses, which are becoming both more accessible and more powerful with continuing developments in cognitive psychology (Frederiksen, Mislevy, & Bejar, 1993; Snow & Lohman, 1989). At the simplest level, this might involve querying respondents about their solution processes or asking them to think aloud while responding to exercises during field trials.

In addition to reliance on these forms of evidence, construct validity, as previously indicated, also subsumes content relevance and representativeness as well as cri-

terion-relatedness. This is the case because such information about the range and limits of content coverage and about specific criterion behaviors predicted by the test scores clearly contributes to score interpretation. In the latter instance, correlations between test scores and criterion measures—viewed within the broader context of other evidence supportive of score meaning—contribute to the joint construct validity of both predictor and criterion. In other words, empirical relationships between predictor scores and criterion measures should make theoretical sense in terms of what the predictor test is interpreted to measure and what the criterion is presumed to embody (Gulliksen, 1950).

An important form of validity evidence still remaining bears on the social consequences of test interpretation and use. It is ironic that validity theory has paid so little attention over the years to the consequential basis of test validity, because validation practice has long invoked such notions as the functional worth of the testing—that is, a concern over how well the test does the job for which it is used (Cureton, 1951; Rulon, 1946). And to appraise how well a test does its job, one must inquire whether the potential and actual social consequences of test interpretation and use are not only supportive of the intended testing purposes, but also at the same time consistent with other social values.

With some trepidation due to the difficulties inherent in forecasting, both potential and actual consequences are included in this formulation for two main reasons: First, anticipation of likely outcomes may guide one where to look for side effects and toward what kinds of evidence are needed to monitor consequences; second, such anticipation may alert one to take timely steps to capitalize on positive effects and to ameliorate or forestall negative effects.

However, this form of evidence should not be viewed in isolation as a separate type of validity, say, of "consequential validity." Rather, because the values served in the intended and unintended outcomes of test interpretation and use both derive from and contribute to the meaning of the test scores, appraisal of the social consequences of the testing is also seen to be subsumed as an aspect of construct validity (Messick, 1964, 1975, 1980). In the language of the Cronbach and Meehl (1955) seminal manifesto on construct validity, the intended consequences of the testing are strands in the construct's nomological network representing presumed action implications of score meaning. The central point is that unintended consequences, when they occur, are also strands in the construct's nomological network that need to be taken into account in construct theory, score interpretation, and test use. At issue is evidence for not only negative but also positive consequences of testing, such as the promised benefits of educational performance assessment for teaching and learning.

A major concern in practice is to distinguish adverse consequences that stem from valid descriptions of individual and group differences from adverse consequences that derive from sources of test invalidity such as construct underrepresentation and construct-irrelevant variance. The latter adverse consequences of test invalidity present measurement problems that need to be investigated in the validation process, whereas the former consequences of valid assessment represent problems of social policy. But more about this later.

Thus, the process of construct validation evolves from these multiple sources of evidence a mosaic of convergent and discriminant findings supportive of score meaning. However, in anticipated applied test use, this mosaic of general evidence may or may not include pertinent specific evidence of (a) the relevance of the test to the particular applied purpose and (b) the utility of the test in the applied setting. Hence, the general construct validity evidence may need to be buttressed in applied instances by specific evidence of relevance and utility.

In summary, the construct validity of score interpretation comes to undergird all score-based inferences—not just those related to interpretive meaningfulness but also the content- and criterion-related inferences specific to applied decisions and actions based on test scores. From the discussion thus far, it should also be clear that test validity cannot ***rely*** on any one of the supplementary forms of evidence just discussed. However, neither does validity ***require*** any one form, granted that there is defensible convergent and discriminant evidence supporting score meaning. To the extent that some form of evidence cannot be developed—as when criterion-related studies must be forgone because of small sample sizes, unreliable or contaminated criteria, and highly restricted score ranges—heightened emphasis can be placed on other evidence, especially on the construct validity of the predictor tests and on the relevance of the construct to the criterion domain (Guion, 1976; Messick, 1989b). What ***is*** required is a compelling argument that the available evidence justifies the test interpretation and use, even though some pertinent evidence had to be forgone. Hence, validity becomes a unified concept, and the unifying force is the meaningfulness or trustworthy interpretability of the test scores and their action implications, namely, construct validity.

Aspects of Construct Validity

However, to speak of validity as a unified concept does not imply that validity cannot be usefully differentiated into distinct aspects to underscore issues and nuances that might otherwise be downplayed or overlooked, such as the social consequences of performance assessments or the role of score meaning in applied use. The intent of these distinctions is to provide a means of addressing functional aspects of validity that help disentangle some of the complexities inherent in appraising the appropriateness, meaningfulness, and usefulness of score inferences.

In particular, six distinguishable aspects of construct validity are highlighted as a means of addressing central issues implicit in the notion of validity as a unified concept. These are content, substantive, structural, generalizability, external, and consequential aspects of construct validity. In effect, these six aspects function as general validity criteria or standards for all educational and psychological measurement (Messick, 1989b). Following a

capsule description of these six aspects, some of the validity issues and sources of evidence bearing on each are highlighted:

- The content aspect of construct validity includes evidence of content relevance, representativeness, and technical quality (Lennon, 1956; Messick, 1989b);
- The substantive aspect refers to theoretical rationales for the observed consistencies in test responses, including process models of task performance (Embretson, 1983), along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks;
- The structural aspect appraises the fidelity of the scoring structure to the structure of the construct domain at issue (Loevinger, 1957; Messick 1989b);
- The generalizability aspect examines the extent to which score properties and interpretations generalize to and across population groups, settings, and tasks (Cook & Campbell, 1979; Shulman, 1970), including validity generalization of test criterion relationships (Hunter, Schmidt, & Jackson, 1982);
- The external aspect includes convergent and discriminant evidence from multitrait–multimethod comparisons (Campbell & Fiske, 1959), as well as evidence of criterion relevance and applied utility (Cronbach & Gleser, 1965);
- The consequential aspect appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness, and distributive justice (Messick, 1980, 1989b).

Content Relevance and Representativeness

A key issue for the content aspect of construct validity is the specification of the boundaries of the construct domain to be assessed—that is, determining the knowledge, skills, attitudes, motives, and other attributes to be revealed by the assessment tasks. The boundaries and structure of the construct domain can be addressed by means of job analysis, task analysis, curriculum analysis, and especially domain theory, in other words, scientific inquiry into the nature of the domain processes and the ways in which they combine to produce effects or outcomes. A major goal of domain theory is to understand the construct-relevant sources of task difficulty, which then serves as a guide to the rational development and scoring of performance tasks and other assessment formats. At whatever stage of its development, then, domain theory is a primary basis for specifying the boundaries and structure of the construct to be assessed.

However, it is not sufficient merely to select tasks that are relevant to the construct domain. In addition, the assessment should assemble tasks that are representative of the domain in some sense. The intent is to insure that all important parts of the construct domain are covered, which is usually described as selecting tasks that sample domain processes in terms of their functional importance, or what Brunswik (1956) called *ecological sampling*. Functional importance can be considered in terms of what people actually do in the performance domain, as in job analyses, but also in terms of what characterizes and differentiates expertise in the domain, which would usually emphasize different tasks and processes. Both the content relevance and representativeness of assessment tasks are traditionally appraised by expert professional judgment, documentation of which serves to address the content aspect of construct validity.

Substantive Theories, Process Models, and Process Engagement

The substantive aspect of construct validity emphasizes the role of substantive theories and process modeling in identifying the domain processes to be revealed in assessment tasks (Embretson, 1983; Messick, 1989b). Two important points are involved: One is the need for tasks providing appropriate sampling of domain processes in addition to traditional coverage of domain content; the other is the need to move beyond traditional professional judgment of content to accrue empirical evidence that the ostensibly sampled processes are actually engaged by respondents in task performance.

Thus, the substantive aspect adds to the content aspect of construct validity the need for empirical evidence of response consistencies or performance regularities reflective of domain processes (Loevinger, 1957). Such evidence may derive from a variety of sources, for example, from "think aloud" protocols or eye movement records during task performance; from correlation patterns among part scores; from consistencies in response times for task segments; or from mathematical or computer modeling of task processes (Messick, 1989b, pp. 53–55; Snow & Lohman, 1989). In summary, the issue of domain coverage refers not just to the content representativeness of the construct measure but also to the process representation of the construct and the degree to which these processes are reflected in construct measurement.

The core concept bridging the content and substantive aspects of construct validity is representativeness. This becomes clear once one recognizes that the term *representative* has two distinct meanings, both of which are applicable to performance assessment. One is in the cognitive psychologist's sense of representation or modeling (Suppes, Pavel, & Falmagne, 1994); the other is in the Brunswikian sense of ecological sampling (Brunswik, 1956; Snow, 1974). The choice of tasks or contexts in assessment is a representative sampling issue. The comprehensiveness and fidelity of simulating the construct's realistic engagement in performance is a representation issue. Both issues are important in educational and psychological measurement and especially in performance assessment.

Scoring Models As Reflective of Task and Domain Structure

According to the structural aspect of construct validity, scoring models should be rationally consistent with what is

known about the structural relations inherent in behavioral manifestations of the construct in question (Loevinger, 1957; Peak, 1953). That is, the theory of the construct domain should guide not only the selection or construction of relevant assessment tasks but also the rational development of construct-based scoring criteria and rubrics.

Ideally, the manner in which behavioral instances are combined to produce a score should rest on knowledge of how the processes underlying those behaviors combine dynamically to produce effects. Thus, the internal structure of the assessment (i.e., interrelations among the scored aspects of task and subtask performance) should be consistent with what is known about the internal structure of the construct domain (Messick, 1989b). This property of construct-based rational scoring models is called *structural fidelity* (Loevinger, 1957).

Generalizability and the Boundaries of Score Meaning

The concern that a performance assessment should provide representative coverage of the content and processes of the construct domain is meant to insure that the score interpretation not be limited to the sample of assessed tasks but be broadly generalizable to the construct domain. Evidence of such generalizability depends on the degree of correlation of the assessed tasks with other tasks representing the construct or aspects of the construct. This issue of generalizability of score inferences across tasks and contexts goes to the very heart of score meaning. Indeed, setting the boundaries of score meaning is precisely what generalizability evidence is meant to address.

However, because of the extensive time required for the typical performance task, there is a conflict in performance assessment between time-intensive depth of examination and the breadth of domain coverage needed for generalizability of construct interpretation. This conflict between depth and breadth of coverage is often viewed as entailing a trade-off between validity and reliability (or generalizability). It might better be depicted as a trade-off between the valid description of the specifics of a complex task and the power of construct interpretation. In any event, such a conflict signals a design problem that needs to be carefully negotiated in performance assessment (Wiggins, 1993).

In addition to generalizability across tasks, the limits of score meaning are also affected by the degree of generalizability across time or occasions and across observers or raters of the task performance. Such sources of measurement error associated with the sampling of tasks, occasions, and scorers underlie traditional reliability concerns (Feldt & Brennan, 1989).

Convergent and Discriminant Correlations With External Variables

The external aspect of construct validity refers to the extent to which the assessment scores' relationships with other measures and nonassessment behaviors reflect the expected high, low, and interactive relations implicit in the theory of the construct being assessed. Thus, the meaning of the scores is substantiated externally by appraising the degree to which empirical relationships with other measures—or the lack thereof—are consistent with that meaning. That is, the constructs represented in the assessment should rationally account for the external pattern of correlations. Both convergent and discriminant correlation patterns are important, the convergent pattern indicating a correspondence between measures of the same construct and the discriminant pattern indicating a distinctness from measures of other constructs (Campbell & Fiske, 1959). Discriminant evidence is particularly critical for discounting plausible rival alternatives to the focal construct interpretation. Both convergent and discriminant evidence are basic to construct validation.

Of special importance among these external relationships are those between the assessment scores and criterion measures pertinent to selection, placement, licensure, program evaluation, or other accountability purposes in applied settings. Once again, the construct theory points to the relevance of potential relationships between the assessment scores and criterion measures, and empirical evidence of such links attests to the utility of the scores for the applied purpose.

Consequences As Validity Evidence

The consequential aspect of construct validity includes evidence and rationales for evaluating the intended and unintended consequences of score interpretation and use in both the short- and long-term. Social consequences of testing may be either positive, such as improved educational policies based on international comparisons of student performance, or negative, especially when associated with bias in scoring and interpretation or with unfairness in test use. For example, because performance assessments in education promise potential benefits for teaching and learning, it is important to accrue evidence of such positive consequences as well as evidence that adverse consequences are minimal.

The primary measurement concern with respect to adverse consequences is that any negative impact on individuals or groups should not derive from any source of test invalidity, such as construct underrepresentation or construct-irrelevant variance (Messick, 1989b). In other words, low scores should not occur because the assessment is missing something relevant to the focal construct that, if present, would have permitted the affected persons to display their competence. Moreover, low scores should not occur because the measurement contains something irrelevant that interferes with the affected persons' demonstration of competence.

Validity as Integrative Summary

These six aspects of construct validity apply to all educational and psychological measurement, including performance assessments. Taken together, they provide a way of addressing the multiple and interrelated validity questions that need to be answered to justify score interpretation and use. In previous writings, I maintained that it is "the relation between the evidence and the inferences

drawn that should determine the validation focus" (Messick, 1989b, p. 16). This relation is embodied in theoretical rationales or persuasive arguments that the obtained evidence both supports the preferred inferences and undercuts plausible rival inferences. From this perspective, as Cronbach (1988) concluded, validation is evaluation argument. That is, as stipulated earlier, validation is empirical evaluation of the meaning and consequences of measurement. The term *empirical evaluation* is meant to convey that the validation process is scientific as well as rhetorical and requires both evidence and argument.

By focusing on the argument or rationale used to support the assumptions and inferences invoked in the score-based interpretations and actions of a particular test use, one can prioritize the forms of validity evidence needed according to the points in the argument requiring justification or support (Kane, 1992; Shepard, 1993). Helpful as this may be, there still remain problems in setting priorities for needed evidence because the argument may be incomplete or off target, not all the assumptions may be addressed, and the need to discount alternative arguments evokes multiple priorities. This is one reason that Cronbach (1989) stressed cross-argument criteria for assigning priority to a line of inquiry, such as the degree of prior uncertainty, information yield, cost, and leverage in achieving consensus.

Kane (1992) illustrated the argument-based approach by prioritizing the evidence needed to validate a placement test for assigning students to a course in either remedial algebra or calculus. He addressed seven assumptions that, from the present perspective, bear on the content, substantive, generalizability, external, and consequential aspects of construct validity. Yet the structural aspect is not explicitly addressed. Hence, the compensatory property of the usual cumulative total score, which permits good performance on some algebra skills to compensate for poor performance on others, remains unevaluated in contrast, for example, to scoring models with multiple cut scores or with minimal requirements across the profile of prerequisite skills. The question is whether such profile scoring models might yield not only useful information for diagnosis and remediation but also better student placement.

The structural aspect of construct validity also received little attention in Shepard's (1993) argument-based analysis of the validity of special education placement decisions. This was despite the fact that the assessment referral system under consideration involved a profile of cognitive, biomedical, behavioral, and academic skills that required some kind of structural model linking test results to placement decisions. However, in her analysis of selection uses of the General Aptitude Test Battery (GATB), Shepard (1993) did underscore the structural aspect because the GATB within-group scoring model is both salient and controversial.

The six aspects of construct validity afford a means of checking that the theoretical rationale or persuasive argument linking the evidence to the inferences drawn touches the important bases; if the bases are not covered, an argument that such omissions are defensible must be provided. These six aspects are highlighted because most score-based interpretations and action inferences, as well as the elaborated rationales or arguments that attempt to legitimize them (Kane, 1992), either invoke these properties or assume them, explicitly or tacitly.

In other words, most score interpretations refer to relevant content and operative processes, presumed to be reflected in scores that concatenate responses in domain-appropriate ways and are generalizable across a range of tasks, settings, and occasions. Furthermore, score-based interpretations and actions are typically extrapolated beyond the test context on the basis of presumed relationships with nontest behaviors and anticipated outcomes or consequences. The challenge in test validation is to link these inferences to convergent evidence supporting them and to discriminant evidence discounting plausible rival inferences. Evidence pertinent to all of these aspects needs to be integrated into an overall validity judgment to sustain score inferences and their action implications, or else provide compelling reasons why there is not a link, which is what is meant by validity as a unified concept.

Meaning and Values in Test Validation

The essence of unified validity is that the appropriateness, meaningfulness, and usefulness of score-based inferences are inseparable and that the integrating power derives from empirically grounded score interpretation. As seen in this article, both meaning and values are integral to the concept of validity, and psychologists need a way of addressing both concerns in validation practice. In particular, what is needed is a way of configuring validity evidence that forestalls undue reliance on selected forms of evidence as opposed to a pattern of supplementary evidence, that highlights the important yet subsidiary role of specific content- and criterion-related evidence in support of construct validity in testing applications. This means should formally bring consideration of value implications and social consequences into the validity framework.

A unified validity framework meeting these requirements distinguishes two interconnected facets of validity as a unitary concept (Messick, 1989a, 1989b). One facet is the source of justification of the testing based on appraisal of either evidence supportive of score meaning or consequences contributing to score valuation. The other facet is the function or outcome of the testing—either interpretation or applied use. If the facet for justification (i.e., either an evidential basis for meaning implications or a consequential basis for value implications of scores) is crossed with the facet for function or outcome (i.e., either test interpretation or test use), a four-fold classification is obtained, highlighting both meaning and values in both test interpretation and test use, as represented by the row and column headings of Figure 1.

These distinctions may seem fuzzy because they are not only interlinked but overlapping. For example, social consequences of testing are a form of evidence, and other forms of evidence have consequences. Furthermore, to

Figure 1
Facets of Validity as a Progressive Matrix

	Test Interpretation	Test Use
Evidential Basis	Construct Validity (CV)	CV + Relevance/Utility (R/U)
Consequential Basis	CV + Value Implications (VI)	CV + R/U + VI + Social Consequences

interpret a test is to use it, and all other test uses involve interpretation either explicitly or tacitly. Moreover, utility is both validity evidence and a value consequence. This conceptual messiness derives from cutting through what indeed is a unitary concept to provide a means of discussing its functional aspects.

Each of the cells in this four-fold crosscutting of unified validity are briefly considered in turn, beginning with the evidential basis of test interpretation. Because the evidence and rationales supporting the trustworthiness of score meaning are what is meant by construct validity, the evidential basis of test interpretation is clearly construct validity. The evidential basis of test use is also construct validity, but with the important proviso that the general evidence supportive of score meaning either already includes or becomes enhanced by specific evidence for the relevance of the scores to the applied purpose and for the utility of the scores in the applied setting, where utility is broadly conceived to reflect the benefits of testing relative to its costs (Cronbach & Gleser, 1965).

The consequential basis of test interpretation is the appraisal of value implications of score meaning, including the often tacit value implications of the construct label itself, of the broader theory conceptualizing construct properties and relationships that undergirds construct meaning, and of the still broader ideologies that give theories their perspective and purpose—for example, ideologies about the functions of science or about the nature of the human being as a learner or as an adaptive or fully functioning person. The value implications of score interpretation are not only part of score meaning, but a socially relevant part that often triggers score-based actions and serves to link the construct measured to questions of applied practice and social policy. One way to protect against the tyranny of unexposed and unexamined values in score interpretation is to explicitly adopt multiple value perspectives to formulate and empirically appraise plausible rival hypotheses (Churchman, 1971; Messick, 1989b).

Many constructs such as competence, creativity, intelligence, or extraversion have manifold and arguable value implications that may or may not be sustainable in terms of properties of their associated measures. A central issue is whether the theoretical or trait implications and the value implications of the test interpretation are commensurate, because value implications are not ancillary but, rather, integral to score meaning. Therefore, to make clear that score interpretation is needed to appraise value implications and vice versa, this cell for the consequential basis of test interpretation needs to comprehend both the construct validity as well as the value ramifications of score meaning.

Finally, the consequential basis of test use is the appraisal of both potential and actual social consequences of the applied testing. One approach to appraising potential side effects is to pit the benefits and risks of the proposed test use against the pros and cons of alternatives or counterproposals. By taking multiple perspectives on proposed test use, the various (and sometimes conflicting) value commitments of each proposal are often exposed to open examination and debate (Churchman, 1971; Messick, 1989b). Counterproposals to a proposed test use might involve quite different assessment techniques, such as observations or portfolios when educational performance standards are at issue. Counterproposals might attempt to serve the intended purpose in a different way, such as through training rather than selection when productivity levels are at issue (granted that testing may also be used to reduce training costs, and that failure in training yields a form of selection).

What matters is not only whether the social consequences of test interpretation and use are positive or negative, but how the consequences came about and what determined them. In particular, it is not that adverse social consequences of test use render the use invalid but, rather, that adverse social consequences should not be attributable to any source of test invalidity, such as construct underrepresentation or construct-irrelevant variance. And once again, in recognition of the fact that the weighing of social consequences both presumes and contributes to evidence of score meaning, of relevance, of utility, and of values, this cell needs to include construct validity, relevance, and utility, as well as social and value consequences.

Some measurement specialists argue that adding value implications and social consequences to the validity framework unduly burdens the concept. However, it is simply not the case that values are being ***added*** to validity in this unified view. Rather, values are intrinsic to the meaning and outcomes of the testing and have always been. As opposed to adding values to validity as an adjunct or supplement, the unified view instead exposes the inherent value aspects of score meaning and outcome to open examination and debate as an integral part of the validation process (Messick, 1989a). This makes explicit what has been latent all along, namely, that validity judgments ***are*** value judgments.

A salient feature of Figure 1 is that construct validity appears in every cell, which is fitting because the construct validity of score meaning is the integrating force that unifies validity issues into a unitary concept. At the same time, by distinguishing facets reflecting the justification and function of the testing, it becomes clear that distinct features of construct validity need to be emphasized, in addition to the general mosaic of evidence, as one moves

from the focal issue of one cell to that of the others. In particular, the forms of evidence change and compound as one moves from appraisal of evidence for the construct interpretation per se, to appraisal of evidence supportive of a rational basis for test use, to appraisal of the value consequences of score interpretation as a basis for action, and finally, to appraisal of the social consequences—or, more generally, of the functional worth—of test use.

As different foci of emphasis are highlighted in addressing the basic construct validity appearing in each cell, this movement makes what at first glance was a simple four-fold classification appear more like a progressive matrix, as portrayed in the cells of Figure 1. From one perspective, each cell represents construct validity, with different features highlighted on the basis of the justification and function of the testing. From another perspective, the entire progressive matrix represents construct validity, which is another way of saying that validity is a unified concept. One implication of this progressive-matrix formulation is that both meaning and values, as well as both test interpretation and test use, are intertwined in the validation process. Thus, validity and values are one imperative, not two, and test validation implicates both the science and the ethics of assessment, which is why validity has force as a social value.

REFERENCES

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley: University of California Press.

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

Cascio, W. F., Outtz, J., Zedeck, S., & Goldstein, I. L. (1991). Statistical implications of six methods of test score use in personnel selection. *Human Performance, 4,* 233–264.

Churchman, C. W. (1971). *The design of inquiring systems: Basic concepts of systems and organization.* New York: Basic Books.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

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. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 34–35). Hillsdale, NJ: Erlbaum.

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

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana: University of Illinois Press.

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

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (1st ed., pp. 621–694). Washington, DC: American Council on Education.

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

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

Frederiksen, N., Mislevy, R. J., & Bejar, I. (Eds.). (1993). *Test theory for a new generation of tests.* Hillsdale, NJ: Erlbaum.

Gottfredson, L. S. (1994). The science and politics of race-norming. *American Psychologist, 49,* 955–963.

Guion, R. M. (1976). Recruiting, selection, and job placement. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 777–828). Chicago: Rand McNally.

Gulliksen, H. (1950). Intrinsic validity. *American Psychologist, 5,* 511–517.

Hartigan, J. A., & Wigdor, A. K. (Eds.). (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery.* Washington, DC: National Academy Press.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning.* Hillsdale, NJ: Erlbaum.

Hunter, J. E., Schmidt, F. L., & Jackson, C. B. (1982). *Advanced meta-analysis: Quantitative methods of cumulating research findings across studies.* San Francisco: Sage.

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

Lennon, R. T. (1956). Assumptions underlying the use of content validity. *Educational and Psychological Measurement, 16,* 294–304.

Loevinger, J. (1957). Objective tests as instruments of psychological theory [Monograph]. *Psychological Reports, 3,* 635–694 (Pt. 9).

Messick, S. (1964). Personality measurement and college performance. In *Proceedings of the 1963 Invitational Conference on Testing Problems* (pp. 110–129). Princeton, NJ: Educational Testing Service.

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. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18(2),* 5–11.

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

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23(2),* 13–23.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

Peak, H. (1953). Problems of observation. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences* (pp. 243–299). Hinsdale, IL: Dryden Press.

Rulon, P. J. (1946). On the validity of educational tests. *Harvard Educational Review, 16,* 290–296.

Sackett, P. R., & Wilk, S. L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist, 49,* 929–954.

Schmidt, F. L. (1991). Why all banding procedures are logically flawed. *Human Performance, 4,* 265–278.

Shepard, L. A. (1993). Evaluating test validity. *Review of research in education, 19,* 405–450.

Shulman, L. S. (1970). Reconstruction of educational research. *Review of Educational Research, 40,* 371–396.

Snow, R. E. (1974). Representative and quasi-representative designs for research on teaching. *Review of Educational Research, 44,* 265–291.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263–331). New York: Macmillan.

Suppes, P., Pavel, M., & Falmagne, J.-C. (1994). Representations and models in psychology. *Annual Review of Psychology, 45,* 517–544.

Wiggins, G. (1993). Assessment: Authenticity, context, and validity. *Phi Delta Kappan, 75,* 200–214.