DOCUMENT RESUME

ED 395 031 TM 025 049

AUTHOR Messick, Samuel
TITLE Validity of Test Interpretation and Use.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-90-11
PUB DATE Aug 90
NOTE 33p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Concurrent Validity; *Construct Validity; *Content Validity; Criteria; Educational Assessment; *Predictive Validity; *Scores; *Test Interpretation; Test Use
IDENTIFIERS *Social Consequences

ABSTRACT

Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment. 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. The key issues of test validity are the meaning, relevance, and utility of scores; the import or value implications of scores as a basis for action; and the functional worth of scores in terms of the social consequences of their use. For some time, test validity has been broken into content validity, predictive validity and concurrent criterion-related validity, and construct validity. The only form of validity neglected or bypassed in these traditional formulations is that bearing on the social consequences of test interpretation and use. Validity becomes a unified concept when it is recognized, or assured, that construct validation subsumes considerations of content, criteria, and consequences. Speaking of validity as a unified concept does not mean that it cannot be differentiated into facets to underscore particular issues. The construct validity of score meaning is the integrating force that unifies validity issues into a unitary concept. (Contains 1 table and 25 references.) (SLD)

RESEARCH REPORT

VALIDITY OF TEST INTERPRETATION AND USE

Samuel Messick

Educational Testing Service
Princeton, New Jersey
August 1990

VALIDITY OF TEST INTERPRETATION AND USE

Samuel Messick
Educational Testing Service

VALIDITY OF TEST INTERPRETATION AND USE

Samuel Messick[1]
Educational Testing Service

Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *interpretations* and *actions* based on test scores or other modes of assessment. 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 here 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 device (such as work samples, portfolios, or realistic problem simulations). This general usage subsumes qualitative as well as quantitative summaries. It applies, for example, to protocols, to clinical interpretations, to behavioral or performance judgments or ratings, and to computerized verbal score reports. Nor are scores in this general sense limited to behavioral consistencies and attributes of persons, such as persistence and verbal ability. Scores may refer as well to functional consistencies and attributes of groups, situations or environments, and objects or institutions, as in measures of group solidarity, situational

---

[1]This article appears in M. C. Alkin, (Ed.), *Encyclopedia of Educational Research* (6th ed.), New York: Macmillan, 1991. Grateful acknowledgements are due Walter Emmerich, Robert Linn, and Lawrence Stricker for their helpful comments.

stress, quality of artistic products, and such social indicators as school drop-out rate.

Broadly speaking, validity is an inductive summary of both the existing evidence for and the actual as well as potential consequences of score interpretation and use. Hence, what is to be validated is not the test or observation device as such but the inferences derived from test scores or other indicators (Cronbach, 1971) -- inferences about score meaning or interpretation and about the implications for action that the interpretation entails. In essence, then, test validation is empirical evaluation of the meaning and consequences of measurement.

It is important to note that validity is a matter of degree, not all or none. Furthermore, over time, the existing validity evidence becomes enhanced (or contravened) by new findings. Moreover, projections of potential social consequences of testing become transformed by evidence of actual consequences and by changing social conditions. In principle, then, validity is an evolving property and validation is a continuing process -- except, of course, for tests that are demonstrably inadequate or inappropriate for the proposed interpretation or use. In practice, because validity evidence is always incomplete, validation is essentially a matter of making the most reasonable case, on the basis of the balance of evidence available, both to justify current use of the test and to guide current research needed to advance understanding of what the test scores mean and of how they function in the applied context. This validation research to extend the evidence in hand then serves either to corroborate or to revise prior validity judgments.

To validate an interpretive inference is to ascertain the extent to which multiple lines of evidence are consonant with the inference, while

establishing that alternative inferences are less well supported. Consonant research findings supportive of a purported score interpretation or a proposed test use are called *convergent* evidence. For example, convergent evidence for an arithmetic word-problem test interpreted as a measure of quantitative reasoning might indicate that the scores correlate substantially with performance on logic problems, discriminate mathematics majors from English majors, and predict success in science courses. Research findings that discount alternative inferences, and thereby give greater credence to the preferred interpretation, are called *discriminant* evidence. For example, to counter the possibility that the word-problem test is in actuality a reading test in disguise, one might demonstrate that correlations with reading scores are not unduly high, that loadings on a verbal comprehension factor are negligible, and that the reading level required by the items is not taxing for the population group in question. *Both* convergent and discriminant evidence are fundamental in test validation (Campbell & Fiske, 1959).

To validate an action inference requires validation not only of score meaning but also of value implications and action outcomes, especially appraisals of the relevance and utility of the test scores for particular applied purposes and of the intended as well as unintended social consequences of using the scores for applied decision making. For example, let us assume that the previously considered word-problem scores, on the basis of convergent and discriminant evidence, are indeed interpretable in terms of the *construct* of quantitative reasoning. The term "construct" has come to be used generally in the validity literature to refer to score meaning -- typically, but not necessarily, by attributing consistency in test responses and score correlates to some quality, attribute, or trait of persons or other objects of

measurement. This usage signals that score interpretations are (or should be) constructed to explain and predict (or less ambitiously, to summarize or at least be compatible with) score properties and relationships.

Given this quantitative reasoning interpretation, the use of these scores in college admissions (action implications) would be supported by judgmental and statistical evidence that such reasoning skills are implicated in or facilitative of college learning (relevance); that the scores usefully predict success in the freshman year (utility); and, that any adverse impact against females or minority groups, for instance, is not due to male- or majority-oriented item content or to other sources of construct-irrelevant test variance but, rather, reflects authentic group differences in construct-relevant quantitative performance (appraisal of consequences or side effects). Thus, the key issues of test validity are the meaning, relevance, and utility of scores, the import or value implications of scores as a basis for action, and the functional worth of scores in terms of the social consequences of their use.

MULTIPLE LINES OF EVIDENCE FOR UNIFIED VALIDITY

Although there are different sources and mixes of evidence for supporting score-based inferences, validity is a unitary concept. Validity always refers to the degree to which evidence and theory support the adequacy and appropriateness of interpretations and actions based on test scores. Furthermore, although there are many ways of accumulating evidence to support a particular inference, these ways are essentially the methods of science. Inferences are hypotheses, and the validation of inferences is hypothesis testing. However, it is not hypothesis testing in isolation but, rather,

theory testing more generally because the source, meaning, and import of score-based hypotheses derive from the interpretive theories of score meaning in which these hypotheses are rooted. As a consequence, test validation is basically both theory-driven and data-driven. Hence, test validation embraces all of the experimental, statistical, and philosophical means by which hypotheses and scientific theories are evaluated. What follows amplifies these two basic points -- namely, that validity is a unified though faceted concept and that validation is scientific inquiry into score meaning.

Sources of validity evidence. The basic sources of validity evidence are by no means unlimited. Indeed, if asked where to turn for such evidence, one finds that there are only a half dozen or so main research strategies and associated forms of evidence. The number of forms is arbitrary, to be sure, because instances can be sorted in various ways and categories set up at different levels of generality. But a half dozen or so categories of the following sort provide a workable level for highlighting similarities and differences among validation approaches:

1. Appraise the relevance and representativeness of the test content in relation to the content of the behavioral or performance domain about which inferences are to be drawn or predictions made.

2. Examine relationships among responses to the tasks, items, or parts of the test -- that is, delineate the internal structure of test responses.

3. Survey relationships of the test scores with other measures and background variables -- that is, elaborate the test's external structure.

4. Directly probe the ways in which individuals cope with the items or tasks, in an effort to illuminate the *processes* underlying item response and task performance.

5. Investigate uniformities and differences in these test processes and structures over time or across groups and settings -- that

is, ascertain that the generalizability (and limits) of test interpretation and use are appropriate to the construct and contexts at issue.

6. Evaluate the degree to which test scores display appropriate or theoretically expected variations as a function of instructional and other interventions or as a result of experimental manipulation of content and conditions.

7. Appraise the value implications and social consequences of interpreting and using the test scores in the proposed ways, scrutinizing not only the intended outcomes but also unintended side effects -- in particular, evaluate the extent to which (or, preferably, discount the possibility that) any adverse consequences of testing derive from sources of score invalidity such as irrelevant test variance.

The guiding principle of test validation is that the test content, the internal and external test structures, the operative response processes, the degree of generalizability (or lack thereof), the score variations as a function of interventions and manipulations, and the social consequences of the testing should all make theoretical sense in terms of the attribute or trait (or, more generally, the construct) that the test scores are interpreted to assess. Research evidence that does not make theoretical sense calls into question either the validity of the measure or the validity of the construct, or both, granted that the validity of the research itself is not also questionable.

One or another of these forms of validity evidence, or combinations thereof, have in the past been accorded special status as a so-called *type* of validity. But because all of these forms of evidence bear fundamentally on the valid interpretation and use of scores, it is not a type of validity but the relation between the evidence and the inference to be drawn that should determine the validation focus. That is, one should seek evidence to support (or undercut) the proposed score interpretation and test use as well as to

discount plausible rival interpretations. In this enterprise, the varieties of evidence are not alternatives but rather complements to one another. This is the main reason that validity is now recognized as a unitary concept (APA, 1985) and why each of the historic types of validity is limiting in some way.

TRADITIONAL TYPES OF VALIDITY AND THEIR LIMITATIONS

At least since the early 1950s, test validity has been broken into three or four distinct types -- or, more specifically, into three types, one of which comprises two subtypes. These are content validity, predictive and concurrent criterion-related validity, and construct validity. These three traditional validity types have been described, with slight paraphrasing, as follows (APA, 1954, 1966):

> Content validity is evaluated by showing how well the content of the test samples the class of situations or subject matter about which conclusions are to be drawn.
>
> Criterion-related validity is evaluated by comparing the test scores with one or more external variables (called criteria) considered to provide a direct measure of the characteristic or behavior in question.
>
> Predictive validity indicates the extent to which an individual's future level on the criterion is predicted from prior test performance.
>
> Concurrent validity indicates the extent to which the test scores estimate an individual's present standing on the criterion.
>
> Construct validity is evaluated by investigating what qualities a test measures, that is, by determining the degree to which certain explanatory concepts or constructs account for performance on the test.

With some important shifts in emphasis, these validity conceptions are found in current testing standards and guidelines. They are given here in their classic or traditional version to provide a benchmark against which to appraise the import of subsequent changes, such as a shift in the focus of content validity from the sampling of situations or subject matter to the

sampling of domain behaviors or processes and a shift in construct validity from being in contradistinction to content and criterion validities to subsuming the other validity types.

Historically, distinctions were not only drawn among three types of validity, but each was related to particular testing aims (APA, 1954, 1966). This proved to be especially insidious because it implied that there were testing purposes for which one or another type of validity was sufficient. For example, content validity was deemed appropriate to support claims about an individual's present performance level in a universe of tasks or situations, criterion-related validity for claims about a person's present or future standing on some significant variable different from the test, and construct validity for claims about the extent to which an individual possesses some trait or quality reflected in test performance.

However, for reasons expounded in detail shortly (see also Messick, 1989a, 1989b), neither content nor criterion-related validity alone is sufficient to sustain any testing purpose while the generality of construct validity needs to be attuned to the relevance, utility, and consequences of score interpretation and use in particular applied settings. By comparing these so-called validity types with the half dozen or so forms of evidence outlined earlier, one can quickly discern what evidence each validity type relies on as well as what each leaves out. The remainder of this section underscores salient properties and critical limitations of the traditional "types" of validity.

Content validity. In its perennial form, content validity is based on expert judgments about the relevance of the test content to the content of a particular behavioral domain of interest and about the representativeness with

which item or task content covers that domain. For example, the relevance and representativeness of the items in a chemistry achievement test might be appraised relative to material typically covered in curriculum and text book surveys, the items in a clerical job selection test relative to job properties and functions revealed through a job analysis, and the items in a personality test relative to the behaviors and applicable situations implicated in a particular trait theory. Thus, the heart of the notion of so-called content validity is that the test items are samples of a behavioral domain or item universe about which inferences are to be drawn or predictions made.

According to Cronbach (1980), "Logically, . . . content validation is established only in test construction, by specifying a domain of tasks and sampling rigorously. The inference back to the domain can then be purely deductive" (p. 105). But this inference is not from the sample of test items to the domain of knowledge or skill or whatever construct is germane, but to the "domain" of *tasks* deemed relevant to that construct. In this regard, it is useful to distinguish the *domain* of knowledge or other construct from the *universe* of relevant tasks (Messick, 1989b). Judgments of relevance are critical in specifying the universe of tasks, and judgments of relevance and representativeness help support inferences from the test sample to the task universe. However, these inferences must be tempered by recognizing that the test not only samples the task universe but casts the sampled tasks in a test format, thereby raising the spectre of context effects or irrelevant method variance possibly distorting test performance vis-à-vis domain performance. Such effects will be discussed shortly. In any event, inferences about the extent to which either the test sample or the task universe taps the construct

domain of knowledge, skill, or other attribute require not content judgment but, rather, construct evidence.

Inconsistency or confusion with respect to this distinction between construct domain and task universe is apparent historically, especially in relation to the form of evidence offered to support relevance and representativeness. Content validity has been conceptualized over the years in three closely related but distinct ways: in terms of how well the *content* of the test samples the content of the domain of interest (APA, 1954, 1966), the degree to which the *behaviors* exhibited in test performance constitute a representative sample of behaviors displayed in the desired domain performance (APA, 1974), and the extent to which the *processes* employed by the examinee in arriving at test responses are typical of the processes underlying domain responses (Lennon, 1956). Yet, in practice, content-related evidence usually takes the form of consensual professional judgments about the content relevance of (presumably construct-valid) items to the specified domain and about the representativeness with which test content covers the domain content. But inferences regarding behaviors require evidence of response or performance consistency and not just judgments of content, whereas inferences regarding processes require construct-related evidence (Loevinger, 1957).

To be more precise about the variety of validity evidence that is ignored or left out, content validity per se is not concerned with response processes, internal and external test structures, performance differences across groups and settings, responsiveness of scores to experimental intervention, or with social consequences. Thus, content validity provides judgmental evidence in support of the domain relevance and representativeness of the content of the test instrument, rather than evidence in support of inferences to be made from

test scores. Response consistencies and test scores are not even addressed in typical accounts of content validity. Some test specifications, to be sure, do refer to desired cognitive levels or response processes. But validity in these instances, being inferred not from test content but from consistencies in test responses and their correlates, is clearly construct-related.

At a fundamental level, then, so-called content validity does not qualify as validity at all, although such considerations of content relevance and representativeness clearly do and should influence the nature of score inferences supported by other evidence. That is, content relevance and representativeness of the test should be consistent with the range or generality of the construct interpretation advanced. Contrariwise, the generality of the construct interpretation should be limited by the content relevance and representativeness of the test, unless sustained by other evidence of generalizability such as external correlations or factor patterns with broader construct measures.

In addition, the ubiquitous problem of irrelevant test variance, especially method variance, is simply not confronted in the content validity framework, even though irrelevant variance serves to subvert judgments of content relevance. Method variance refers to all systematic effects associated with a particular measurement procedure that are extraneous to the focal construct being measured (Campbell & Fiske, 1959). Included are all of the context effects or situational factors (such as an evaluative atmosphere) that influence test performance differently from domain performance (Loevinger, 1957). For example, experts may judge items ostensibly tapping knowledge or reasoning as highly relevant to domain problem solving, but the items might instead (or in addition) measure reading comprehension. Or,

objective multiple-choice items aimed at knowledge or skill might contain such transparent distractors that they primarily reflect merely testwiseness or common sense. As another instance, subjective scores for the persuasiveness of writing might primarily reflect prowess in punctuation and grammar or be influenced by the length of the writing sample produced.

Indeed, irrelevant test variance contributes, along with other factors, to the ultimate frailty of traditional content validation, namely, that expert judgment is fallible and may imperfectly apprehend domain structure or inadequately represent test structure, or both. Thus, as previously indicated, content validity alone is insufficient to sustain any testing purpose, with the possible exception of test samples that are truly domain samples observed under naturalistic domain conditions. Even here, however, the legitimacy of the test sample as an exemplar of the construct domain must ultimately rest on construct-related evidence. The way out of this impasse is to evaluate (and inform) expert judgment on the basis of other evidence about the structure of the behavioral domain under consideration as well as about the structure of the test responses -- namely, through construct-related evidence.

Criterion-related validity. As contrasted with content validity, criterion-related validity is based on the degree of empirical correlation between the test scores and criterion scores. This correlation then serves as a basis for using the test scores to predict an individual's standing on a criterion measure of interest such as grade-point average in college or success on a job. As such, criterion-related validity only emphasizes selected parts of the test's external structure. The interest is not in the pattern of relationships of the test scores with other measures generally, but

instead is more narrowly focussed to spotlight selected relationships with measures held to be criterial for a particular applied purpose in a specific applied setting. Thus, there are as many criterion-related validities for the test scores as there are criterion measures and settings, and the extent to which a criterion correlation can be generalized across settings and times has become an important and contentious empirical question (Schmidt, Hunter, Pearlman, & Hirsh, with commentary by Sackett, Schmitt, Tenopyr, Keho, & Zedeck, 1985).

Essentially, then, criterion-related validity is not concerned with any other sorts of evidence except specific test-criterion correlations or, more generally, the regression system linking the criterion to the predictor scores. However, criterion scores are *measures* to be evaluated like all measures. They too may be deficient in capturing the criterion domain of interest and may be contaminated by irrelevant variance -- as in supervisors' ratings, for example, which are typically distorted by selective perception and by halo effects or other biases. Consequently, potentially deficient and contaminated criterion measures cannot serve as the unequivocal standards for validating tests, as is intrinsic in the criterion-oriented approach to validation.

Thus, as indicated previously, criterion-related validity per se is insufficient to sustain any testing purpose, with the possible (though extremely unlikely) exception of predictor tests having high correlations with uncontaminated complete criteria. Even here, however, the legitimacy of the criterion measure as an exemplar of the criterion domain -- that is, the extent to which it captures the criterion construct -- ultimately needs to rest on construct-related evidence and rational arguments (Thorndike, 1949).

The way out of this paradox -- that criteria, being measures that need to be evaluated in the same manner as tests, cannot serve as the standard for evaluating themselves -- is to evaluate both the criterion measures and the tests in relation to construct theories of the criterion domain.

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 aspects 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.

There are two major threats to construct validity: One is *construct underrepresentation* -- that is, the test is too narrow and fails to include important dimensions or facets of the construct; the other is *construct irrelevant variance* -- that is, the test is too broad and contains excess reliable variance associated with other distinct constructs as well as method variance making items or tasks easier or harder for some respondents in a manner irrelevant to the interpreted construct. 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." Hence, one should not belabor whether or not construct evidence is needed because the score in question might not refer to a theoretical construct -- as, for example, in arguing that teacher competence (referring to the repertoire of specific things that teachers know, do, or believe) "does not seem like a theoretical construct" (Mehrens, 1987, p. 215). It does not matter whether one contends that competence, knowledge, skill, or belief are constructs. If test scores are interpreted in these terms, then convergent and discriminant evidence should be provided that high scorers exhibit domain competence (that is, enabling knowledge and skill) in task performance -- as opposed to answering the test items on some other basis such as rote memory, testwiseness, or common sense. More importantly, one must be cautious about interpreting low scores as lack of competence without first discounting a number of plausible rival hypotheses for poor test performance such as anxiety, fatigue, low motivation, limited English proficiency, or handicapping conditions. And, the discounting of plausible rival hypotheses is the hallmark of construct validation (Messick, 1989b).

Rather than questioning the construct basis of a particular score interpretation, it is more prudent to simply admit the ubiquity of constructs, recognizing that what is often in dispute is the degree to which they are explicitly theoretical, that is, based on an elaborated theory or embedded in a nomological network of expected relationships. However, to the extent that the score interpretation has little or only vague theoretical underpinnings,

construct validation becomes even more important because it then serves to clarify as well as to support score meaning.

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. 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 of score meaning, however, are studies of expected performance differences over time (such as increased impulse-control scores from childhood to young adulthood); across groups and settings (as in contrasting, for measures of domain problem-solving, the solution strategies of novices versus experts or, for measures of creativity, the creative productions of individuals in self-determined versus directive work environments); and, in response to experimental treatments and manipulations (such as 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 are direct probes and modeling of the processes underlying test responses (for example, via "thinking aloud" protocols during task performance), an approach becoming both more accessible and more powerful with continuing developments in cognitive psychology (Snow & Lohman, 1989).

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

In one way or another, then, these three traditional types of validity, taken together, make explicit reference to all but one of the forms of validity evidence mentioned earlier. This occurs in spite of the ad hoc singularity of reference of both content- and criterion-related validity, but because of the comprehensiveness of reference of construct validity. The only form of validity evidence bypassed or neglected in these traditional formulations is that bearing on the social consequences of test interpretation and use.

It is ironic that little attention has been paid over the years to the consequential basis of test validity, because validity has been cogently conceptualized in the past in terms of the functional worth of the testing -- that is, in terms of how well the test does the job it is employed to do (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 at the same time are consistent with other social values. However, this form of evidence should not be viewed in isolation as a fourth

validity type, 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 social consequences of the testing is also seen to be subsumed as an aspect of construct validity (Messick, 1964, 1975, 1980).

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 uses of tests, this mosaic of *general* evidence may or may not include pertinent *specific* evidence of the relevance of the test to the particular applied purpose and 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. Relevance of test use entails evidence, including content- and criterion-related evidence, that the test validly reflects processes or constructs deemed important in the applied domain. For example, evidence of relevance might involve the judgmental linking of item content -- or the empirical linking of item or test scores -- to domain performance dimensions derived from job- or task-analyses. Utility of test use captures the benefits relative to the costs of the testing, usually as a function of the degree of test-criterion correlation (Cronbach & Gleser, 1965). In a sense, then, the very generality of construct validation may bring about some limitations or potential limitations in applied test usage -- unless the mosaic of general construct findings includes evidence of relevance and utility in the applied setting or until such evidence can be developed. That is, the general evidence for construct validity of score

meaning may not be precise or specific enough to warrant use of the test for the particular purpose in particular applied settings.

Offsetting limitations of validity types. Although the three traditional types of validity are now viewed as aspects of a unitary concept of validity, it is still useful to underscore the salient strengths and especially the weaknesses of these erstwhile validity types in order to illuminate both the need for and the power of the unified validity view. In content validation the touchstone is expert judgment specifying the relevance and representativeness of the test content vis-à-vis domain content. In criterion-oriented validity the touchstone is the criterion measure, which serves as the standard for evaluating the relevance and utility of the test scores. However, the basic problem is that these touchstones are not only fallible or subject to error, but possibly bogus because expert judgment may not only be unreliable but biased while criteria may not only be contaminated but incomplete. In contrast, in construct validation the touchstone is convergent and discriminant evidence corroborating score meaning and discounting plausible rival interpretations. Although any piece of evidence may be fallible and some may be spurious, the continuing construct validation process attempts to appraise, and take into account, the nature and extent of such distortion in the evolving validity judgment.

Indeed, such convergent and discriminant evidence provides a rational basis for evaluating the other two suspect touchstones of content and criterion validity, both generally in test interpretation as well as in specific instances of applied test use. In other words, construct-related evidence is critical in the very delineation of content domains and in the conceptualization and measurement of applied criteria -- that is, in precisely

those aspects of domain coverage and criterion prediction that are at the heart of traditional content- and criterion-oriented validities. From this standpoint, the construct validity of score interpretation comes to undergird *all* score-based inferences -- not just those related to interpretive meaningfulness but including the content- and criterion-related inferences specific to applied decisions and actions based on test scores.

On the other hand, as previously indicated, the mosaic of general construct validity evidence supportive of score interpretation may still be limited with respect to particular proposed uses of tests. In such instances, the mosaic needs to be extended to include evidence of the relevance of the test to the applied purpose and the utility of the test in the applied setting. Thus, considerations of specific content and selected criteria resurface as part of the general construct validity of score meaning whenever the test is used for a particular applied purpose. Granted, neither content- nor criterion-validity can stand alone to support the specific testing application because, ultimately, score meaning for both tests and criteria is needed to justify test use (Loevinger, 1957; Thorndike, 1949). But in the context of broader construct validity evidence for score interpretation, they can support (or counter) the action implications of score meaning that provide the rational basis for the proposed use.

Although in practice each of the three traditional validation approaches thus has real or potential problems, these are offset by treating the three conjointly -- or, rather, by recognizing and assuring that construct validation subsumes considerations of content, criteria, and consequences in test interpretation and use. Thus, test validity cannot *rely* on any one of the six or seven forms of evidence discussed earlier, but neither does

it *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 the relevance of the construct to the criterion domain (Guion, 1976; Messick, 1989b). 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.

FACETS OF UNIFIED VALIDITY

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. However, to speak of validity as a unified concept is not to imply that validity cannot be usefully differentiated into facets to underscore issues and nuances that might otherwise be downplayed or overlooked, such as the social consequences of testing or the role of score meaning in applied test use. The intent of these distinctions or facets 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, what is needed is a way of configuring validity evidence that forestalls undue reliance on selected forms of evidence, that highlights the important though subsidiary role of specific content- and criterion-related

evidence in support of construct validity in testing applications, and that formally brings consideration of value implications and social consequences into the validity framework.

Meaning and values as ways of configuring validity evidence. 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, being based on appraisal of either evidence supportive of score meaning or of consequences contributing to score valuation. The other facet is the function or outcome of the testing, being either interpretation or applied use. If the facet for justification (that is, 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 (that is, 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 Table 1.

Table 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

The four cells thereby generated correspond to four interrelated aspects of the basic validity question, which might be phrased as follows: To what degree if at all, on the basis of evidence and rationales, should the test scores be interpreted and used in the manner proposed? Four distinguishable but interrelated aspects of this question ask what balance of evidence sustains the interpretation or meaning of the scores; what evidence supports not only score meaning, but also the relevance of the scores to the particular applied purpose and the utility of the scores in the applied setting; what makes credible the value implications of the score interpretation and any associated implications for action; and, what signifies the functional worth of the testing in terms of its intended and unintended consequences?

Let us briefly consider in turn each of the cells in this four-fold crosscutting of unified validity, beginning with the evidential basis of test interpretation. Because the evidence and rationales supporting the trustworthiness of score meaning is 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.

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 (for example, about the functions of science or the

nature of the human being as a learner) that give theories their perspective and purpose (Messick, 1989b). For example, a contrast between behavior that is variable as opposed to repetitive might be given a construct label of "flexibility versus rigidity," but the value implications of score interpretation would be very different if the label were instead "confusion versus control." Similarly, a construct and its associated measures interpreted as "inhibited versus impulsive" would have different value implications if it were instead labeled "self-controlled versus self-expressive."

Many constructs such as competence, creativity, intelligence, or extraversion have manifold and arguable value implications which may or may not be sustainable in terms of properties of their associated measures. A central issue is whether or not 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 thus 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 performance standards are at issue. Or 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.

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 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.

Thus, 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 aspects of construct validity need to be emphasized, in addition to the general mosaic of evidence, 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 added to 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 Table 1. 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.

References[2]

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, **51**(2, Part 2).

American Psychological Association. (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

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

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

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

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

---

[2]For an extensive bibliography on this topic see S. Messick (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

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 -- Proceedings of the 1979 ETS Invitational Conference* (pp. 99-108). San Francisco: Jossey-Bass.

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

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.

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. *Psychological Reports*, 3, 635-694 (Monograph Supplement 9).

Mehrens, W. A. (1987). Validity issues in teacher licensure tests. *Journal of Personnel Evaluation in Education*, 1, 195-229.

Messick, S. (1964). Personality measurement and college performance. *Proceedings of the 1963 Invitational Conference on Testing Problems* (pp. 110-129). Princeton, NJ: Educational Testing Service. [Reprinted in A. Anastasi (Ed.). (1966). *Testing problems in perspective* (pp. 557-572). Washington, DC: American Council on Education.]

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.

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

Schmidt, F. L., Hunter, J. E., Pearlman, K., & Hirsh, H. R., with commentary by Sackett, P. R., Schmitt, N., Tenopyr, M. L., Keho, J., & Zedeck, S. (1985). Forty questions about validity generalization and meta-analysis. *Personnel Psychology*, 38, 697-798.

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.

Thorndike, R. L. (1949). *Personnel selection*. New York: Wiley.