DOCUMENT RESUME

ED 388 689 TM 023 920

AUTHOR	Stuck, Ivan
TITLE	Heresies of the New Unified Notion of Test Validity.
PUB DATE	Apr 95
NOTE	31p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 19-21, 1995).
PUB TYPE	Viewpoints (Opinion/Position Papers, Essays, etc.) (120) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Evaluation Methods; *Measurement Techniques; Measures (Individuals); *Predictive Validity; Research Problems; Selection; *Statistical Bias; Test Items; *Test Validity

ABSTRACT

By focusing on "appropriateness" and "adequacy" of inference and action, unified validity may be misused in rejecting valid test outcomes. The notion of levels of validity is challenged, the necessity of assumption is argued, and experience is proposed as the basis of validity. "Consequential validity" is interpreted as an optional predictive validity, a tangential validity that depends on organizational or political prerogative. Measurement validity is distinguished from test validity, which usually has more importance. Test validities such as content and predictive validities are perceived as demonstrable in contrast to construct measurement, which can never be proven. The claim that all validity is construct validity is challenged. The same claim can be made for any type of validity, and not all valid tests require constructs. Tests with valid constructs may not provide adequate predictive and content validity, whereas selection tests may be adequate. It is argued that test purpose is a more important validity issue than credibility of construct. "Structural item validity" is suggested as an alternative descriptor for items free of random and systematic bias; systematic bias does not constitute measurement of an additional construct. Noting the futility of attempting to demonstrate "construct validity," it is suggested that the term be renamed "construct feasibility." (Contains 11 references.) (Author/SLD)

Heresies of the New Unified Notion of Test Validity

Ivan Stuck, American College Testing
INTERNET Stuck@ACT-ACT4-PO.act.org

ABSTRACT

By focusing on "appropriateness" and "adequacy" of inference and action, unified validity may be misused in rejecting valid test outcomes. The notion of "levels of validity" is challenged, the necessity of assumption is argued, and experience is proposed as the basis of validity. "Consequential validity" is interpreted as an optional predictive validity; a "tangential validity" that depends on organizational or political prerogative. Measurement validity is distinguished from test validity which usually has more importance. Test validities such as content and predictive validities are perceived as demonstrable in contrast to construct measurement which can never be proven. The claim is challenged that all validity is construct validity-- the same claim can be made for any type of validity, and not all valid tests require constructs. Tests with valid constructs may not provide adequate predictive and content validity whereas selection tests may be adequate. It is argued that test purpose is a more important validity issue than credibility of construct. "Stuctural item validity" is suggested as an alternative descriptor for items free of random and systematic bias; systematic bias does not constitute measurement of an additional construct. Noting the futility of attempting to demonstrate "construct validity", it is suggested that the term be renamed "construct feasibility".

NOTE: The ideas contained in this paper are solely those of the author, and do not reflect the institutional views of ACT.

Running head: HERESIES OF UNIFIED VALIDITY

Paper presented at the annual meeting of the National Council of Measurement in Education, 1995, San Fransciso

DATE: 10/11/95

ERIC/TME Clearinghouse Report

Document Cover Sheet

TM #: 023920

Title: HERESIES OF THE NEW UNIFIED NOTION OF TEST VALIDIT

Number of Pages: 28

Publication Date: 04/21/95

Document Level: 1

Notes:

Suggested Descriptors:

Heresies of the New Unified Notion of Test Validity

Ivan Stuck, American College Testing

The most authoritative rendering of the notion of a 'unitary' test validity is found in Messick's (1989) chapter in the third edition of **Educational Measurement.** Other discussions of the topic include Angoff (1988), Cronbach (1988), and Messick (1988) in **Test Validity**; Geisinger (1992); Moss (1992); and Shepard (1993) in **Review of Research in Education, 19.** Having initially accepted Messick's views as unobjectionable, the author re-evaluated his thinking about validity in the course of his work on a domain-referenced testing program. He concluded that the unitary validation thinking was problematic on a number of issues including the following: (1) that validation could be an evaluation, (2) that validities of inference and action could be drawn from test scores, (3) that social consequences are a necessary component of validity, (4) that construct validity is the whole of validity, and (5) that validation is a perpetual process. The following discussion addresses these issues by recasting some assumptions about test validity and by challenging some contemporary views.

CAN EVALUATORS DETERMINE WHAT IS APPROPRIATE AND ADEQUATE?

Messick (1989) calls validity "an...evaluative judgment" of support for the "adequacy and appropriateness of inferences and actions" when those are based on assessment (p. 1). This definition is in contrast with the classic definition paraphrased from Garrett (1947, p. 394). "The validity of a test is the extent to which it measures what it purports to measure." Messick's validity is a validity of inference and policy and not of the validity of the test or measurement scale itself. Whereas conventional test validity might err by either selecting or failing to select marginally meritorious candidates, Messick's policy validity is apt to sacrifice the most meritorious candidates in order to select the mediocre; by a slight modification in the intended inference or policy, the conventional inference or action can be ruled invalid allowing the targeted candidate to be selected. Determining what is "appropriate" or "adequate" is perhaps suitable for personal or religious decisions, but for secular and socially responsible policy, these terms are hopelessly vague and subject to abuse.

Consider two examples of how Messick's validity might be misused:

> An otherwise valid test might indicate that a particular candidate was the best qualified for a particular opportunity, but if the candidate's selection was considered inappropriate due to other unmet selection goals, the basis would exist to find the test an inadequate basis for the action originally intended.

> A "favorite son" candidate is favored for a position which is supposed to be open to both internal and external candidates. A job description is revised to recast a new vision for the performance of the position. Coincidentally, the "favorite son" candidate fits the revised job description perfectly and is hired.

The selection of policy or inferences that are "adequate and appropriate" reaches far beyond the test to thwart any commitment to the most meritorious candidate. The best qualified candidates can be positioned for rejection in order to accommodate political or personal interests. No doubt Messick's concern was for the dangers of exceeding the valid use of a test. As these examples demonstrate, however, damage can also be done by rejecting the validity of a fair test. Test validity is easily sabotaged by simply modifying the criterion. Messicks' evaluative approach makes this injustice more possible.

Allowing validity to be determined by "evaluative judgment" rather than by a rule, allows any conceivable judgment by any convenient argument. Either (1) allowing the judgment to be made by non-experts or (2) making the judgment unduly complicated encourages incompetent policy. Responsible professionals will avoid either (a) attempting to determine what inferences are "appropriate" for others or (b) attempting to determine what action is "adequate" for someone else's organization. Ideally, a test should be selected or developed with reference to appropriate inference and adequate action, before the test is administered.

Because Messick's validity is subjective, it will be found and not found across evaluators even when the test is truly valid. It will likewise be found and not found across situations and sites. Because his evaluation of validity is inherently unreliable at finding validity when it exists, it is therefore inherently invalid. Messick's notion might well be identified as a model of test invalidity.

IS VALIDITY SOMETHING TO BE MEASURED?

Messick (1989) states that "Validity is a(n)... judgment of **the degree** to which..." inferences and actions are supported by evidence and theory (p. 1). Elsewhere he proclaims "...validity is itself clearly a construct and subject to construct validation" (Messick, 1988, p. 43). This emphasis suggests that he believes that validity has a literal existence, is quantifiable on some kind of scale, and should be measured. This contrasts with a less restricted view that sees validity as a nominal level judgment; validity either maintaining or not dependent on whether (or not) a specified purpose has been served. By restricting validity to only evaluations that reflect an underlying score continuum the cost of validity increases, the likelihood of establishing validity is decreased substantially, and the incentive for trivializing validation study is markedly increased. Messick (1989) further states: "...validity is a matter of degree, not all or none" (p. 13). Whether intended or not, Messick is "talking turkey" with us; Messick poses a false dichotomy stating that validity falls in the middle of a continuum and not at the extreme ends of the scales. This begs the question of whether validity needs to be measured at all.

We might agree that a hammer is a valid tool for driving a nail--is it meaningful to quantify the degree to which the hammer is valid? If we administer a spelling test of 100 words that students have studied in one week, do we need to quantify how valid the spelling test was? The proposition can be strengthened further. Suppose that we don't know whether or not the validity for a particular test is quantifiable, but we do know from observation that persons who do well on the test are correspondingly competent at the skills that we intended to measure. Is the test invalid for lack of a stated **level** of validity? On the one hand it is easy to think of ways to make a valid tool more or less valid, it is not so easy to propose a unidimensional scale for measuring validity that remains consistent with the purpose of the test. I submit that there is no compelling reason to require validity to be stated in quantitative terms; in many situations the attempt to quantify validity will be frivolous. For instance, can a valid sample (of a fixed size) be more or less valid than a parallel sample? What is essential is to decide in a straightforward way if the test meets its intended purpose.

NEED EDUCATION BE VALID?

If testing is inappropriate except where a degree of validity can be specified, then won't the same follow for the stipulation of "valid" requirements in educational programs? Must we provide a validity coefficient for the value of each required course in high school? How many basketball players have been denied careers in basketball due to academic failure in subjects that have no direct relation to the practice of professional basketball? Is it really sufficiently appropriate to require a basketball player to know history? Why should a student be compelled to take a course of study where there is no quantified proof that the study will improve the student's future practice? Where is the quantified evidence for the validity of interview procedures, personal reference, and ratings of performance frequently inplemented in educational practice?

If the censorship of testing can be justified merely for lack of quantified validity, the same arguments can be used to censor education and (ultimately) free speech in public institutions.

THE BASIS OF VALIDITY: MEASUREMENT OR ASSUMPTION?

An important dichotomy that does need to be addressed is whether validity is ultimately measurable or whether it must be assumed. The latter position was argued convincingly by Ebel (1961) and his logic remains unassailable. If you use a correlation to argue validity, you must validate the criterion. To validate the criterion, you must use some other basis that requires additional validation. A cyclic referral of the validity question to an endless chain of additional measurements, each requiring its own validation, never resolves the validity question; a basic **assumption** of measurement validity must be made at some point if we are to ever conclude that a scale is valid. Reason suggests that the basic validity assumption be addressed by experienced observers and that it be as defensible as possible. The validity issue should not be "all or none" nor should it be "a matter of degree", these are unduly difficult to support--it should be yes or no; this tool is helpful or it is not.

THE BASIS OF VALIDITY: TEST SCORES OR EXPERIENCE?

Messick and others take the position that validity is a judgment based on inferences made from test scores. Cronbach's (1971) classic line is that "One validates, not a test, but an interpretation of data arising from a specified procedure" (p. 447). I am not impressed by these assertions. I am convinced that test validity is based on the experience of various subject matter experts whose inferences about item performance are often well established before the test is built. Without the basis of experience there would be no test blueprint, test item, or even scale construct. Additional psychometric guidance in item and test structure can be helpful in further polishing and refining the test. Again, it is the experience of experts coupled with the observations of psychometricians that can reduce erosions of reliability caused by measurement error and test bias. It must be admitted, however, that in subject areas where expertise is lacking, the impact of measurement error and test bias may suppress the test's validity. The solution is not, however, to emphasize the discovery of invalidity from test outcomes after the test is built and administered, but to evaluate the expertise of the subject matter experts prior to the construction and administration of the test.

CONSEQUENTIAL VALIDITY: PRUDENCE OR POLITICAL CORRECTNESS?

Messick (1989, p.20) is concerned with the impact of names and labels on measurement scales as well as the validity of inferences and actions that result from the testing program. The names and labels may be the creation of the testing community, however, the inferences and actions will mostly occur outside the testing community. One thing that Messick dismisses too quickly is that the same persons who are wont to make misguided inferences and actions based on well developed test instruments are quite willing to make the same misguided inferences and actions without the benefit of well developed test instruments; when push comes to shove managers will do something that seems plausible--like testing--or they could use a worse alternative. To Messick's credit, he does note that Ebel (1980, pp.34-35) proposed weighing the potential consequences of test use against the potential consequences of not using the test. Indeed, in most selection contexts the question is not whether a test can be used, but rather which test will be the most valid. Most selection can't be postponed for lack of validity.

Political retaliation and the threat of litigation would seem to cultivate caution on the part of scale developers and labelers on a more immediate basis than the spectacle of actual harm. Furthermore, political and legal repercussions need not represent any authentic danger to succeed in derailing a testing program. Hence, the suggestion that test validity requires a correct language invites political censorship and politicalization of the testing enterprise. Although the basis of concern today may be social consequences, the political consequence may well be politically influenced test outcomes. Using inoffensive labels is acceptable if they are not misleading; allowing the testing enterprise to accommodate political pressure is unacceptable. The critical social consequence of a valid test should be a constantly improving level of quality products or services to the consumer. Beyond that essential and ambitious goal, the other social consequences of testing remain virtually independent of the intentions of the testing community. It is not the test, after all, that treats examinees unjustly; injustice requires a social context. A well developed test represents greater fairness and opportunity than most other sectors of the social context will provide. We cannot afford to surrender as a political hostage the label "test validity."

A well developed test will be biased towards persons who are more proficient. No test can be perfectly free of irrelevant biasing factors and no test can perfectly reflect the examinee's level of proficiency. Thus it must be expected that groups differing on biological factors will exhibit at least some small systematic differences in test outcomes. Inequivalence of group mean scores does not mean that the test is unfair even though the difference is bias in a technical sense. Any other means of selection will likewise exhibit random and systematic bias; there will not be a fairer alternative means of selection than an appropriate and well developed formal test.

Tests often find value as competitive contests (i.e. quiz shows, spelling bees) in which competitors rely not only on proficiency but on luck, and on the ability to avoid careless mistakes. Contests are accepted and well understood by every level of society. Contestants are always expected to absorb some minor injustices as a part of the contest; they play their best and are rewarded when they can outplay their opponents. A test is fair if the examinee has had the opportunity to learn or attain the proficiency needed to properly meet the challenges posed in the test. It is a given that certain ancillary proficiencies such as mastery of language, verbal reasoning, mathematical reasoning, short and long term memory, motivation, and style of anxiety reduction will enhance or restrict the examinees' ability to perform optimally. In most testing situations, the candidate has the option of retesting, further diminishing the potential for inadequate assessment.

A great deal of the measurement research which purports to find ethnic and gender bias is ultimately political in nature. Generally, so-called racial groups include groupings which have their origins in politics: they confound both race and ethnicity; hence, their study can indicate neither cultural nor organic effects. Beyond the cultural confounding of the groupings is an incredible research error: the use of political variables to predict educational outcomes. Research on ACT assessment outcomes by Noble, Crouse, Sawyer, and Gillespie (1992) has demonstrated that when the appropriate educational variables are entered into the prediction of test outcomes, gender and ethnicity have only a marginal impact on test outcome. In a multiple regression study ACT scores were associated with predictors as follows:

40% - 64%	coursework and grade variables
05% - 08%	background and aspiration variables
05% - 07%	high school variables
01%	**ethnic and gender variables**

For the most part, the social consequences of testing can be understood as simply the continuing consequences of the family's pursuit of educational opportunities. In short, these tests are indicating essentially what they are supposed to--educational proficiency. When the proper educational variables have been replaced by atheoretical political variables, bogus findings of test bias are predictable.

FACETS OF CONSTRUCT VALIDITY OR PREDICTIVE VALIDITY?

One irony of the unified position is that it both advocates (1) the identifying of constructs, and (2) the avoidance of negative social consequences. By tagging individuals and groups with construct scores, inter-group comparisons become inevitable. Because many biological traits are linked to family ancestry, group trait differences are likely to suggest stereotypes that will sometimes seem offensive. While this obvious social consequence is not a sufficient reason to curtail personality research altogether,

MODIFIED FACETS OF VALIDITY TABLE

	TEST INTERPRETATION (construct interpretaticn)	TEST USE (score interpretation)
EVIDENTIAL BASIS	construct validity	utility
CONSEQUENTIAL BASIS	labeling	associated outcomes

it would seem prudent to avoid the scaling of constructs where such measures are not essentia˙ to the purpose of the test. In particular, tests having the purposes of knowledge retrieval or prediction would not require well defined personality constructs.

Messick (1989) creates a two by two matrix which he labels "Facets of Validity" (p.20). Test Interpretation and Test Use are each matched with Evidential Basis and Consequential Basis. The table is modified to reflect my interpretation of Messick's point. I would label the first column "Construct Interpretation" and the second column "Score Interpretation". The language of the construct theory might suggest value implications and the score interpretation (passing standard) might have social implications. Presumably, the importance of the value implications depends on sensitivity towards the putative construct in question. A "moral maturity" score might evoke more angst than a "love of flowers" score. Conversely, overly ambitious labels such as "Fascist scale" may simply evoke scorn and ridicule.

The "Test Use" column refers to social policy outcomes triggered by the interpretation of test scores. One such outcome might be an uneven representation of political groupings. I have split the matrix into two separated columns to make the assertion that the two columns are largely independent; where a pass standard is set depends on predictable outcomes and is virtually independent of the meaning of the construct at the pass standard. Indeed, when Messick calls for positive social consequences from testing, he is requiring a predictive validity of positive social outcomes stemming from use of the test. Presuming that the same number of people will be selected regardless of the instrument used, Messick is discounting the interests of persons who would otherwise be successful under formal test selection; by obtaining an alternative selection, the same number of (otherwise successful) candidates will be disenfranchised--is this a less onerous social consequence? I perceive the appropriate aspects of social consequence to be (1) whether the product or service user realizes a benefit from more competent service and better quality products as a result of appropriate selection policy, and (2) whether competition among the candidates to meet selection standards encourages better education and preparation for service.

Messick's table does touch upon the dimensionality of validity. One's attribution of validity will hinge on a variety of other tangential predictive validities that are especially important in a particular testing context. For Messick, science and social consequence are paramount tangential validities--even if they seem to work at cross-purposes. From other vantages important tangential outcomes might be the test's potential for fiscal solvency, academic motivation, political appeal, cultural appeal, professional bonding, career opportunity, increasing product and service quality for the consumer, and reward for attainment--to only name a few. All tangential validities have some legitimacy in most contexts, but may be particularly important in special contexts. Parallel to Messick's evaluation of validity according to his particular vantage, other test critics will attend to other tangential predictive validities, either finding validity or not depending on a particular view of test validity. It is important to note that these tangential predictive validities are voluntary burdens that are not universally essential and that may often be unfair. Tangential validities may be critical in a particular context, or to particular authority figures. For many test contexts, potential social consequence and the pursuit of science are simply not realistic nor relevant concerns. To encumber all testing efforts with demands for additional compulsory tangential predictive validities is presumptuous and unrealistic.

TEST VALIDITY AND MEASUREMENT SCALE VALIDITY: ARE THEY THE SAME?

There are no intrinsic limits to the number of purposes that can exist for testing. There is also much testing that occurs external to formal testing, such as the interview, the evaluation of resumes, or the evaluation of letters of recommendation.

As we are well aware, testing is a tradition not only in the area of education and psychology, but throughout all the sciences and fine art disciplines. I submit that a formal test requires several essential elements: (1) an examinee, (2) a test administrator, (3) a set of challenges, (4) a system of observation (5) a rule for determining test outcome, and (6) anticipated consequences of test outcome. Furthermore, I distinguish between "test" and "measurement"; observing that one can take measurements external to a testing context and that one can test without taking measurements. In the testing literature, unfortunately, the terms are often used synonymously.

Stevens (1946) defines measurement as "...the assignment of numerals to objects or events according to rules." He also states that "...measurement aspires to create interval level scales, and it sometimes succeeds. The problem usually is to devise operations for equalizing the units of the scales..." Measurement is an elaborate observation that requires theoretical and technological justification beyond what is needed for mere empirical observation. Measurement and test are qualitatively different operations which may be combined or remain independent. As a test, you might ask each candidate to face a doorway and to attempt to pass through it, those who can pass through the doorway without lowering their heads would pass the test and the others would be rejected. This test would not use a measurement scale thus it would not be "measurement". If on the other hand you measure each examinee with a ruler and reject those who are taller than a specified height, then your test is also "measurement". If you measure the height of someone merely out of curiosity, this would be measurement, but it would not be a test. To establish the validity of a test which also uses measurement, it is only essential to establish the validity of the test when the specified pass standard is used. The validity of the measurement itself is incidental when the test demonstrates acceptable predictive validity. When Cronbach (1971) makes the statement that "One does not validate the test but rather an interpretation of data arising from a specified procedure" (p. 447), I suggest that he is confounding measurement and testing. In his quotation, the term "test" represents a measurement scale, while his "interpretation of data" is, in fact, the test pass standard that is applied using the measurement as a means of observation. What he seems to be saying is that the measurement scale used in making the observations is not being validated but rather it is the rule for determinining the test outcome that must be validated.

DOES IT MATTER HOW VALIDITY EVIDENCE IS ACCUMULATED?

A direct means of validation is preferred by most people who must assess validity. Whether it is for identifying a driver with a pictured drivers license, or whether it is finding the fine print on a hundred dollar bill, the simplest, most direct validation is ideal. A simple term for direct validation is the word "proof". For tests, proof of the test's validity is a reasonable and desirable expectation when a test is to be used for an important purpose. If the purpose of the test is to predict future performance, we want to see evidence that the test has been predictive in the past. If the purpose of the test is to assess achievement in a course of study, we want to know how well the content of the test represents the content of the course. Other information about the test may be helpful also, but we want that additional information to inform us regarding how well the purpose of the test was met.

Evidence about the predictiveness of a final exam in chemistry might be interesting but it is irrelevant to the question of content validity. Likewise, it would not be important to know how well a licensure test represents the content of a particular course taken by a candidate. The licensure test is supposed to be broadly predictive in nature, hence, the evidence desired for this test is evidence of greater competence for those who pass the test.

Measurement scales are also instruments which are frequently used in testing, classification, and psychological research. Measurement scales are also commonly referred to as "tests" even when they aren't being used to "test" anyone. The primary purpose of the measurement scale is not to "test" but rather to measure: to assign the proper numerical value to some level of ability (or trait) that the subject is expected to possess. The gradients on the ability scales are expected to have equal and meaningful intervals. When observable behavior is being measured, the validity of the measurement scales can be verified directly. For unobservable hypothesized abilities and traits, however, it is impossible to prove the validity of a scale. In this case, an inductive, or argumentative approach must be used to show that the underlying theory corresponds to the response data. An argumentative approach is less than "proof" because without direct observation of the hypothesized construct, too many plausible alternative explanations will remain to account for the outcomes. Instead of proof, for defending opinions about constructs, there is only argument and incomplete evidence. The term construct validity was originally devised to address the need for a means of establishing the credibility of a measurement scale for something that was speculative and difficult to verify. According to Cronbach and Meehl (1955) "A construct is some postulated attribute of people assumed to be reflected in test performance....The constructs in which tests are to be interpreted are certainly not likely to be physiological. Most often they will be traits..." It is important to note that a hypothetical mental ability will often be associated with a predictable outcome that can be observed. Part of the argument for the hypothetical ability might be the predictive value of the measurement scale. The finding of an actual predictive relation between construct and outcome may bolster support for the feasibility of the construct, but the converse does not hold. Construct feasibility does not contribute to predictive validity in a practical way.

In order for the construct-outcome relations to be usable for a predictive test, predictive validity must be demonstrated explicitly for the specific pass standard. In addition, it would often be desirable to expand the predictor base or otherwise modify or confound the theoretical source of prediction. The construct-outcome relationship that supports a construct validity argument is likely to lack the right predictive range and strength to establish predictive validity for a particular outcome criterion.

For test validities such as predictive validity and content validity, direct and elegant proof of validity can be assembled. Extraneous information will be largely unappreciated. For construct validation, an elaborate argument will be required to show the correspondence between the construct theory and the structure of the data from the scale. Merely accumulating important information about the test or measurement scale is hardly appropriate for supporting either test or scale validity.

HOW DOES VALIDITY UNITE?

Messick argues for a unitary notion of validity. In this unitary validity notion, distinctions between measurement validity vs test validity and distinctions between the purposes of tests would be simply ignored. Messick (1989) christens the new notion of validity with a familiar label, "construct validity". He states, "Construct validity is based on an integration of any evidence that bears on the interpretation or meaning of the test scores" (p. 17). The words "any evidence" make the frame of inquiry almost infinite. He also states that "...constructs are not explicitly defined, but rather, are more like 'open concepts'." He further remarks that "Almost any kind of information about a test can contribute to an understanding of its construct validity...". Messick borrows heavily from the original idea of construct validity but he is clearly defining a novel concept of validation and not merely extending the old. Like the old version, the "new" construct validation continues to offer weak evidence, it continues to be used to justify a measurement scale (as opposed to a test), and it continues to evaluate the consistency between measurement scores and a construct theory.

The proposed unity of test validity suggests that one validity is suitable for all purposes. Does this mean that there can only be one purpose for testing?

IS CONSTRUCT VALIDITY SUFFICIENT FOR ALL TEST PURPOSES?

To support the premise that construct validity is "the whole" of validity, Messick (1989) observes that a variety of evidence can be used to support construct validity (content-related evidence, predictive evidence, and etc.), hence all validity evidence is thought to be "subsumed" within construct validation (p. 17). According to the same logic, if a measurement textbook includes a section of text devoted to the statistical roots of test measurement, we might conclude that statistics are subsumed by the discipline of educational measurement. We might also conclude that all validity is face validity because any kind of validity will enhance the appearance of an instrument's validity.

It is equally true that all types of validity are subsumed in content validity and that all types of validity are subsumed in predictive validity (or in any other type of validity that you wish to specify). For any type of validity one can demonstrate that that single type of validity can be a general model under which all other validity types fit as special cases. What cannot be done, however, is to use the wrong type of validity to fully justify the practical purposes of a test. For instance, a test that is determined to have construct validity as measuring fourth grade spelling will not satisfy the teacher who really needs to know how well the students learned the particular words that were taught in a specific fourth grade classroom. Likewise, this same construct-valid test will not predict fifth-grade spelling as well as a test that has been constructed for the purpose of predicting fifth grade spelling (from fourth grade words). The fact that a particular content domain may be viewed as a construct does not establish that all content domains can or ought to be considered to be constructs.

The fact that a measurement correlates to an outcome variable may support a hypothesis that the measurement reflects a personality construct. A good criterion-related predictor, nevertheless, may perform poorly in an appropriate multi-trait multi-method matrix; conversely, a construct scale may be completely consistent with theoretical expectations, yet be inferior to an alternative criterion-related predictor. The fact that a well-functioning construct scale may provide satisfactory prediction in some situations does not mean that other criterion-related tests will necessarily benefit from construct validation.

By definition, construct validation supports the credibility of the construct. Support for other purposes such as domain-referenced testing or criterion-related testing will not be optimal. The biblical proverb is appropriate, a servant cannot serve two masters.

Construct validation cannot be said to include all tests for the simple reason that not all tests use interval scores: some tests only test and do not measure. A domain-referenced test with multiple content areas and a constant cut-score may be used for selecting certification candidates. Candidates are selected based on technical knowledge and not on any construct of ability. Although number-right scores are used for equating to the original cut-score, the test is pass/fail thus there is no measurement scale. Many certification and licensure tests fit this model.

The following are disadvantages to the notion of a unitary test validity:

(1) Measurement validity is confounded with test validity

(2) Tests which have different purposes are not distinguished by appropriate differences in types of test validity

(3) Any information relevant to the test is said to be validity evidence; irrelevant and trivial evidence will be collected

(4) Validity evidence is "accumulated" indirectly rather than fashioning a study to target the validity question directly

(5) A pointless collection of validity evidence constitutes an overall lowering of validity standards which will further erode confidence in testing

(6) Since the need to identify various types of validity will continue, there will be confusion over the meaning of construct validity

WHICH SHOULD BE PRIMARY, VALIDITY OF CONSTRUCT OR TEST PURPOSE?

The validity of a test depends on whether its purpose has been met. A valid construct scale does not insure test validity, nor does an invalidated construct scale prevent test validity. A test with an altered purpose is a different test with different validity requirements. A test may sustain modifications in constructs, conditions of administration, content specifications, or passing standards, and retain its validity on the same basis as the previous test format. Furthermore, legal and ethical constraints require particular standards of evidence for tests according to proposed use. Clearly, the test purpose and not a construct scale should be acknowledged as the primary validity issue.

SHOULD CONSTRUCT VALIDITY BE REQUIRED OF ALL TESTS?

It is argued that measurement error and test bias are construct validity problems. Because all tests are subject to the error and bias, the argument goes, all tests must be concerned with construct validity. In a general sense, the statement is true; with or without a construct scale, validity requires that the item be correctly interpreted. Nevertheless, in most situations the most reasonable and effective solution is not to implement a construct analysis and validation study but to upgrade item reliability and to reduce the cultural loading of items. I would suggest that the label "structural validity" be used to convey the notion that the intended meaning of the test item is correctly interpreted by the examinee. If the test item is part of a construct scale, structural validity would also include that a construct relevant aspect of the item is correctly interpreted by the examinee. Structural validity could be assessed by a combination of item reliability analysis and the appropriate differential item functioning analysis. Structural validity would differ from construct validity by pertaining to individual test items only, and it would not be required to reflect any construct except in the case where the item contributes to a construct scale.

CAN CONSTRUCTS BE FOUND TO BE VALID?

Messick (1989) and others reiterate a somewhat pessimistic statement about finding validity: "Inevitably, then, validity is an evolving property and validation is a continuing process." (p. 13). Cronbach (1988) states "That question...[validity]...I now regard as shortsighted and unanswerable...**validation is never finished**" (p. 5). The straightforward reason that construct validity can never be realized is that the validation process does not pertain to the issue of whether the measurement is caused by the underlying construct, nor does it pertain to the existence of such a construct. Messick (1989) states "The test score is not equated with the construct..., nor is it considered to define the construct"(p. 17). What is called construct validity is not a method of validation at all. It is known a priori that no construct will ever be validated as a consequence of construct validation. Construct validity is the consistency with which measurement data correspond to expectations drawn from construct theory. As Messick says, "...primary emphasis in construct validation has been placed...on patterns of relationships among item scores or between test scores and other measures" (p. 17). Construct validity is actually a form of reliability pertaining to the consistency of measurement data with theoretical expectations. For the researcher's convenience, the theoretical expectations may even be altered to fit the data, for as Messick points out "constructs are not explicitly defined..."(p.17). The theoretical reliability evidence provided by construct validation can never be compelling proof that the measurement is linked to an ability, nor that such an ability even exists. What is more, even if construct validity could validate construct measurement, predictive test validity would nevertheless be needed to justify claims of predictiveness, and content validity would be still needed to substantiate the appropriateness of the content sample. Predictive test validity can be compelling proof of test validity because it can provide evidence of the consistency of the actual predictive inference which is intended. Likewise, content validation can prove that a set of items is a valid sample of the content domain that needs assessment. Construct validity provides less than proof, and for that reason it never can solve the validity question. Because it is not a method of determining validity it should be properly called "construct feasibility".

SUMMARY

The notion of unitary test validity was primarily advanced by Sam Messick (1989) in his chapter on validity in **Educational Measurement, third edition.** The author disagrees with several assertions included in Messick's reformulated thinking on validity: (1) that validation should be an evaluation, (2) that inference validity and action validity should be, (or can be) drawn from test scores, (3) that social consequences of testing are an immanent concern, or that they need to be a universal concern, (4) that construct validity is the whole of validity, and (5) that validation is a perpetual process. The author contends (1) that validation requires proof that the test has met its purpose, (2) that validity is drawn from experience rather than from test scores and that inference and action validity should be built into the test, (3) that purported "social consequences" are largely political, and that tangential validities like scientific growth and "social enhancement" are not universally required in all test contexts, (4) that construct validation can never determine validity and that predictive and content validity remain the primary test validities, and (5) that validation is established once the appropriate proof is provided. The author suggests that the unitary notion of test validity is a method for invalidating tests. He offers the term "item structural validity" to be used to represent the clarity of the item "test" to the examinee rather than referring to that validity issue solely as construct validity.

Messick's view of validity appears to be that of a construct requiring argumentive evaluation to establish its literal and quantifiable existence. The author contends that validity is merely a human-relevant judgment that has no literal existence; "valid" being similar to evaluations of "good" or "appropriate". He views validity as a selection test applied to a test--a "metatest", if you like--under which the test that meets a carefully determined standard is deemed valid. For criterion-related and domain-referenced tests it is possible to demonstrate satisfaction of predictive and domain sampling metatests.

References

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

Ebel, R.L. (1961) Must all tests be valid? American Psychologist 16:640-647.

Ebel, R.L. (1980) Practical problems in educational measurement. Lexington, Massachusetts: D.C.Heath.

Garret, H.E. (1947). Statistics in psychology and education. New York: Longmans, Green.

Geisinger, K.F. (1992) The metamorphosis of test validation. Educational Psychologist 27(2):197-222.

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

Messick, S. (1989) Validity. In R.L.Linn (Ed.), Educational measurement (3rd. ed., pp. 13-103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Moss, P.A. (1992) Shifting conceptions of validity in educational measurement: implications for performance assessment. Review of Educational Research 62(3):229-258.

Noble, J., Crouse, J., Sawyer, R., Gillespie, M. (1992) Ethnic/gender bias and the differential preparation hypothesis: implications for performance on the ACT assessment. A paper presented at the annual meeting of the American Educational Research Association, San Francisco, April.

Shepard, L.S. (1993) Evaluating Test Validity. In L.Darling-Hammond (Ed.), Review of research in education, 19. Washington, DC: American Educational Research Association.

Stevens, S.S. (1946) On the theory of scales of measurement. Science, 103:677-680.