DOCUMENT RESUME

ED 438 309	TM 030 620
AUTHOR	Marsh, S. Neil
TITLE	A Brief History of the Evolution of Views of the Nature of Score Validity.
PUB DATE	2000-01-27
NOTE	18p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Dallas, TX, January 27-29, 2000).
PUB TYPE	Information Analyses (070) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Evaluation Methods; *Inferences; Psychometrics; *Scores; *Validity

ABSTRACT

This paper explores the historical evolution of views of score validity beginning with J. Guilford (1946) and ending with the writings of L. Cronbach (1989), S. Messick (1989), M. Kane (1992), P. Moss (1992), and L. Shepard (1993). Current treatments of validity have emphasized that it is the inferences made from scores, and not tests, that are valid (L. Wilkinson and the American Psychological Association Task Force on Statistical Inference, 1999). Recent treatments have increasingly emphasized the importance of considering falsification in evaluating validity. (Contains 15 references.) (Author/SLD)

A Brief History of The Evolution of Views of The Nature of Score Validity

S. Neil Marsh

Texas A&M University

77843-4225

Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, January 27, 2000.

Abstract

The paper explores historical evolution of views of score validity beginning with Guilford (1946) and ending with the writings of Cronbach (1989), Messick (1989), Kane (1992), Moss (1992), and Shepard (1993). More recent treatments of validity have emphasized that it is the inferences made form scores, and <u>not</u> tests, which are valid (Wilkinson & The APA Task Force on Statistical Inference, 1999). Furthermore, recent treatments have increasingly emphasized the importance of falsification as important in evaluating validity.

A Brief History of The Evolution

of Views of The Nature of Score Validity

Recent thinking about score validity has been marked by a movement to reject the so-called "holy trinity" of content, criterion-related, and construct validity (Guion, 1980) -- and to replace the old view with a unified view of validity. Although controversy as to what this unified view might entail continues, the unified view has already permeated some professional standards (cf. Joint Committee, 1994). The Joint Committee (1994) reported that the validation process entails the gathering of information that supports the interpretations and the uses of specified instruments, proceduures, and/or measurements. Furthermore, members of the Joint Committee posited that in validating research the following guidelines are recommended:

> A detailed description of the constructs and behaviors about which information will be acquired; an analysis of what type of information a particular data collection procedure purports to acquire; a detailed description of how the procedure was implemented, how responses and observations were judged or scored, and how interpretations were made; a presentation of evidence -- both qualitative and quantitative -- that justifies the use of the particular procedure; and an overall assessment of the validity of the interpretation and use of the information provided by the procedure, with reference to the evaluation questions and processes (p. 145).

To facilitate understanding of current changes in thinking related to score validity, this paper examines both historical and contemporary views that express the evolution of such a salient concept. In doing so, the writings of Guilford (1946), Campbell (1957), Campbell and Fiske (1959), in addition to more recent researchers (e.g., Cronbach, 1989; Messick, 1989; Kane, 1992; Moss, 1992; 1995; and Shepard, 1993). The purpose of this paper is simply to provide a brief historical overview of views of score validity.

Score Validity: The Beginning

Guilford's (1946) seminal article proposed that the concept of validity is often misconstrued and is best defined in two manners. "Factorial validity" refers to the question, "Does the test measure what it is supposed to measure?" and is affected by variables' pattern and structure coefficients on meaningful, common, reference factors. On the other hand, "practical validity" is evaluated in terms of the correlation of a test with a given criterion of adjustment, vocational, or personal. Furthermore, Guilford went on to mention numerous opinions that had unfortunately become endemic in research scholarship. Among these hypotheses, Guilford confronted the following:

> (1) validities of .50 to .60 are the practical upper limits of correlation between test scores and criteria of success; (2) validities of .10 and .20 are so inconsequential that tests with such small predictive values are not worth using, even in test batteries; (3) each test in a battery should have a maximum correlation with the practical criterion; (4) after combining four or five tests in a battery, the validity of the composite cannot be materially increased by adding more tests; (5) there would be no question concerning the utility of tests with validities of .60 to .80 and; (6)tests are valid if by inspection they obviously look valid (1946).

Guilford posited that many researchers operate under the disillusion that by increasing the reliability of test scores, score validity is necessarily enhanced. Guilford (1946) disputes this idea by reporting that, only when greater score reliability engenders an increase in variance among valid factors, is validity actually increased.

Score Validity: Recent Thought

The concept of validity has become a perplexing ideal that has created a great deal of dissonance among professionals who have invested themselves in finding concrete and static defining attributes. In looking at types of validity, if such a notion remains reasonable, Cronbach (1989) recognized that the field of psychology often finds construct validity "confusing" (p. 147). In fact, Moss (1992) pointed to the Standards (AERA, APA,, & NCME, 1985) as a prototype of the inconsistent representation of score validity. Moss pointed out that although the most recent version of guidelines related to validity revolve around a "...unitary concept requiring multiple types of evidence to support specific inferences made from test scores...it retained the traditional three-part framework..." (p. 232). In fact, Cronbach (1989) reported that with all the translating of theory into action and literature, professionals in measurement have created a spectrum of inconsistencies ranging from "...utopian doctrine to vapid permissiveness" (p. 147).

Cronbach (1989) noted that although historical steps were initiated to further emphasize objectivity and a thorough understanding of score validity, professionals in strategic positions allowed the large portion of research reconstruction to focus on more acceptable issues such as methods to support interpretations. However, recent efforts to focus on validity of interpretations rather than actual tests and measurements, has placed more importance on the ability to generalize test performance. Cronbach (1989) argued that saying "The evidence to date is consistent with the interpretation" is more applicable than stating that, "The test has construct validity" (p. 151). Furthermore, in assessing testing events, the researcher is only able to form interpretations based on the test results and the researcher's values, which involved often forgotten aspects of validity, namely, the social consequences of test interpretations.

As stated earlier, Moss (1992) recognized the incongruency between practice and theory regarding validity issues. Not only did Moss identify the problems of relying on the historical, "holy trinity" of validity, she proposed that only "...construct validity...provided a scientifically useful basis for establishing the validity of a test" (p. 232). Furthermore, Moss (1992) valued the need to assess social consequences of testing, when evaluating the level of validity of interpretations of a specific testing event. Although Moss recognized that concerns related to negative and positive consequences of assessment are not a novel area of interest, she stressed that evaluation of these consequences is vital to any examination of validity. In doing so, Moss (1992) posited that although traditional standards of scientific testing are helpful, more recent changes are attempting to create more practical and realistic applications in testing, in which the interests of the tested individual remain the primary focus.

In addition to Cronbach (1989) and Moss (1992), Messick's seminal (1989) writing looked at other aspects of validity. For example, Messick attempted to account for the historically forgotten aspects of the social repercussions of testing and the related interpretations. Messick's (1989) conceptualization of validity contained the following description: "...an integrated evaulative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 13). Again, Messick's view of validity invoked ideals that require adjusting the researcher's focus from test validity toward a focus on the validity of test interpretations. In fact, Messick's "Facets of Validity" (1980; as cited in Messick, 1989) table provided a visual depiction of a method that separates and combines "...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..." (p. 20).

In addition, Messick (1989) confronted the commonly taught and trusted thinking of "static" validity coefficients for a specific test. Messick discussed the temporal nature of validity findings, and stated that "...validity is an evolving property and validation is a continuing process. Because evidence is always incomplete, validation is essentially a matter of making the most reasonable case to guide both current use of the test and current research to advance understanding of what the test scores mean" (p. 13). Therefore, as research in a specific area engenders more understanding, the concept of validity will also itself unfold or evolve.

Kane (1992) added to Messick's view and reported that in making inferences from test scores to non-test behavior the researcher is forced to make assumptions about the interaction between test performance and non-test behavior. In addition, "...inferences from test scores to theoretical constructs depend on assumptions included in the theory defining the construct" (p. 527). These assumptions require stringent scrutiny before presuming that scores are valid in a given application.

Kane's "argument-based" approach to validity proposed that the researcher must decide on the statements and decisions to be based on the test scores. Kane posited that by operationally defining the details of the argument, conclusions regarding score validity are more clear. Second, Kane recommended that the researcher specify the inferences and assumptions leading from the test scores to these statements and decisions. In doing so, the researcher applies logical and scientific reasoning to determine the degree of validity or invalidity of the proposed inferences. Third, Kane specified that the individual interpreting test result must identify potential competing interpretations. This validation step stemmed from theory based on ideas of falsifiability and rival hypotheses (Popper, 1962; Campbell, 1957; and Campbell and Fiske, 1959), which will be briefly discussed later in the writing. Similarly, Kane's final suggestion emphasized the importance of seeking evidence to support the inferences and assumptions in the proposed interpretations and possible rival arguments. Kane argued that "...it is importance to identify the assumptions being made and to provide supporting evidence for the most questionable of these assumptions" (p. 528). Similar to the idea of statistical significance testing (Thompson, 1994), information supporting a highly likely result is less important than evidence that supports a less likely occurrence (Kane, 1992).

In addition to Kane (1992), Shepard's (1993) article noted that establishing validity is required for each testing administration. In doing so, Shepard reported that the validation for each testing event should be marked by a combination of logical argument and empirically-grounded inferences. Furthermore, Shepard noted reasons to reject the "trinitarian doctrine" of validity, reporting that the unified concept of construct validity more appropriately meets current standards of inference validation.

In the rejection of the trinitarian doctrine of validity, Shepard reported that individuals supporting the use of content-related validity, which was best described as the evaluative property that examines an individual's performance on a defined universe, were losing "...ground in their battle with construct validity because invariably test-based interpretations assume the generality of broader concept labels..." (p. 412). Furthermore, Shepard's examination of criterion-related validity, which was historically subdivided into predictive validity and concurrent validity, indicated that more recent thought in testing demanded more thorough evidence in the validation process. In fact, Shepard supported thinking that "Empirical relations are necessary but not sufficient to establish the validity of test use." (p. 411). Although, criterion-related and content-related validity enable the researcher to gain information related to construct validity (Anastasi & Urbina, 1997), neither has the ability to successfully assess overall validity of an individual testing event. Shepard (1993) also said that " It is ironic that a field so attuned to the fallacy of mistaking correlation for causation in experimental contexts would be willing to accept correlations in the measurement sphere as immediate proof of test validity" (p. 411).

On a larger scale, a great deal of effort has been placed into professional "test result interpretation." More recent treatments of validity by the APA Task Force on Statistical Inference have emphasized that it is the inferences made form scores, and <u>not</u> tests, which are valid (Wilkinson & The APA Task Force on Statistical Inference, 1999). Furthermore, when examining a test's ability to provide reliable estimates of an individual's standing in relation to specific constructs, the test must lack characteristics that are similar to unrelated constructs. The Task Force cited Messick (1989) as a needed source for further explanation regarding ethical and effective guidelines, as traditional researchers have misconstrued the importance and meaning of "valid" findings.

During the historical evolution of views of validity, researchers have moved from requiring just "professional judgement" (Messick, 1989) in determining the validity of a specific testing event. More recently, the need for more "reliable response consistencies as well as construct evidence that test and domain behaviors are similar or from the same response class" (p. 18) has encouraged educational researchers to adhere to more stringent guidelines when interpreting test results.

Falsification and Rival Hypotheses

Various scholarships have increasingly emphasized the importance of falsification as important in evaluating validity. The notion of time- and situation-bound validity of inferences implies an interest in exploring the boundaries of valid score use. Historically, "validity" was considered a static phenomenon, a test was either valid or invalid. However, as Messick (1989) discussed, validity is not a dichotomous value, but rather a concept accounted for in degrees. In addition, the validity of inferences related to a specific testing event are only as accurate as the researcher's examination and consideration of external (generalizability) and internal (effectiveness of experimental stimulus) factors surrounding the specific testing experience. Campbell (1957) discussed the importance of considering the following variables when evaluating possible variables that interfere with an individual's ability to provide an accurate picture of their performance ability: "...history, maturation, testing, instrument decay, regression, selection, and mortality" (p. 311).

This interest, in turn, implied the utility of logics such as "plausible rival hypotheses" (Campbell, 1957). In examining the role of all factors and their influences on the specific testing event or occurrence, the researcher is more able to make valid conclusions based on the test performance. For example, a researcher may need to consider possible effects of a non-randomized sample, when reaching conclusions about existing differences between two sets of test performances. In examining all possible reasons for the performance the researcher is able to confidently "stand behind" their individually-specific inferences.

In addition, the multitrait-multimethod evaluation of convergent and discriminant validity (Campbell & Fiske, 1959) is important in evaluating score validity. Campbell and Fiske defined convergent validation as "...a confirmation by independent measurement procedures" (p. 81), and discriminative validation as the test's ability to discriminate itself from other tests. Therefore, Campbell and Fiske's (1959) development of the multitrait-multimethod matrix assisted in examining both discriminative and convergent validity in research. Campbell and Fiske's research was based on the idea that validity of specific results should be able to be assessed in more than one experimental condition. Campbell and Fiske intended to exemplify the inability of an atheoretical approach to test construction to provide score validity evidence.

Popper's (1962) concept of falsification also beckoned researchers to examine their thinking in validation issues. Popper's idea falsifiability, refutability, or testability (1962) was the backbone for determining the objective worth of a scientific measurement. In determining the testability of a specific testing event Popper proposed the following guidelines:

1. It is easy to obtain confirmations, or verifications, for nearly every theory--if we look for confirmations.

2. Confirmations should count only if they are the result of risky predictions; that is to say, if, unenlightened by the theory in question, we should have expected an event which was incompatible with the theory -- an event which would have refuted the theory.

3. Every 'good' scientific theory is a prohibition: it forbids certain things to happen. The more a theory forbids, the better it is.

4. A theory which is not refutable by any conceivable event is nonscientific. Irrefutablity is not a virtue of a theory (as people often think) but a vice.

5. Every genuine test of a theory is an attempt to falsify it, or to refute it. Testability is falsifiablity; but here are degrees of testability: some theories are more testable, more exposed to refutation, than others; they take, as it were, greater risks.

6. Confirming evidence should not count except when it is the result of a genuine test of the theory; and this means that it can be presented as a serious but unsuccessful attempt to falsify the theory. (I now speak in such cases of 'corroborating evidence'.)

7. Some genuinely testable theories, when found to be false, are still upheld by their admirers -- for example by introducing ad hoc some auxiliary assumption, or by re-interpreting the theory ad hoc in such a way that it escapes refutation. Such a procedure is always possible, but it rescues the theory from refutation only at the price of destroying, or at least lowering, the scientific status...(p. 37).

The concept of falsification requires that a theory not be deemed credible until the theory has survived serious disconfirmation efforts. As Moss (1995) explained,

> A "strong" program of construct validation requires an explicit conceptual framework, testable hypotheses deduced from it, and multiple lines of relevant evidence to test the hypotheses. Construct validation is most efficiently guided by the test of "plausible rival hypotheses" which suggests credible alternative explanations or meanings for the test score that are challenged and refuted by the evidence collected… Essentially, test validation examines the fit between the meaning of the test score and the measurement intent, whereas construct validation entails the evaluation of an entire theoretical framework (pp. 6-7).

Discussion

As the concept of score validity has evolved the focus of score validity evaluation has also changed to fit the culture's most recent emphasis. Traditional views of validity have "paved the way" for more contemporary researchers to further explore and examine possible overt and latent effects of valid and invalid inferences drawn from measurements. Furthermore, with current research focusing on issues such as quality of life and ethics, researchers have a responsibility to apply stringent guidelines to both qualitative and quantitative research that may effect individuals participating in their studies. In validity research and general understanding, professionals share the burden of moving into more current understandings of cornerstone issues such as the importance of evaluating score validity, and also the ability of professionals to make valid inferences from test results.

In doing so, professionals in helping professions better meet the needs of those who may benefit from professional assistance. Cronbach (1989) portrayed the researcher's responsibility most accurately by stating: "Unlike the test developer, the evaluator holds no brief for or against the test, but rather is committed to serve all the persons having stakes in affairs the test might influence" (p. 164).

References

Anastasi, A., & Urbina, S. (1997). Psychological testing (7th edition). Upper Saddle River, NJ: Prentice-Hall, Inc.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. Psychological Bulletin, 54, 297-312.

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

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

Guilford, J. P. (1946). New standards for test evaluation. Educational and Psychological Measurement, 6, 427-439.

Guion, R. M. (1980). On trinitarian doctrines of validity. Professional Psychology, 11, 385-398.

Joint Committed for Standards on Educational Evaluation. (1994). The program evaluation standards: How to assess evaluations of educational programs. Newbury Park, CA: Sage.

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

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

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

Moss, P. A. (1995). Themes and variations in validity theory. Educational Measurement: Issues and Practices, 14 (2), 5-12.

Popper, K. R. (1962). Conjectures and refutations: The growth of scientific knowledge. New York: Harper & Row.

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), Review of research in education (Vol. 19, pp. 405-450). Washington, DC: American Educational Research Association.

Thompson, B. (1994). The concept of statistical significance testing. Measurement Update, 4 (1), 5-6.

Wilkinson, L., & The APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54, 594-604. [reprint available through the APA Home Page: http://www.apa.org/journals/amp/amp548594.html]

Table 1: Facets of Validity

	Test Interpretation	**Test Use**
Evidential Basis	Construct Validity	Construct Validity + Relevance/Utility
Consequential Basis	Value Implications	Social Consequences

Note. Adapted from (Messick, 1989).

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM030620

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: A BRIEF HISTORY OF THE EVOLUTION OF VIEWS OF THE NATURE OF SCORE VALIDITY

Author(s): S. NEIL MARSH

Corporate Source:

Publication Date: 1/27/00

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1	Level 2A	Level 2B
☒	☐	☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature:	Printed Name/Position/Title: S. NEIL MARSH	
Organization/Address: TAMU Dept Educ Psyc College Station, TX 77843-4225	Telephone: 409/845-1335	FAX:
	E-Mail Address:	Date: 1/20/00

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com