DOCUMENT RESUME

ED 418 129 TM 028 223

AUTHOR Humphries-Wadsworth, Terresa M.
TITLE Emerging/Evolving Views of the Meaning of Score Validity.
PUB DATE 1998-04-11
NOTE 14p.; Paper presented at the Annual Meeting of the Southwestern Psychological Association (New Orleans, LA, April 11, 1998).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Codes of Ethics; Definitions; Evaluation Methods; Psychological Testing; *Scores; Test Construction; *Test Use; *Validity
IDENTIFIERS *Consequential Evaluation

ABSTRACT

The American Psychological Association, in the late 1940s, began work to establish a code of ethics to include and address the needs of members in scientific and applied fields. Out of the ethics work emerged a set of standards for evaluating psychological tests. Four categories, or types of validity, were identified: content, predictive, concurrent, and construct. In subsequent years, predictive and concurrent were combined in a single category labeled criterion validity. The resulting three categories of validity, sometimes called the holy trinity, having survived nearly 40 years of use, are now entrenched concepts in test construction and evaluation. Current trends in the conceptualization of test validity dismiss these three categories as separate entities, but old habits die hard, as apparently do old ideas. This paper reviews the emergence of validity as a unitarian concept, and discusses current views with particular attention to consequential validity. The current theory that delineates the superordinate nature of construct validity and dictates that all lines of validity evidence ultimately point to the construct is reviewed. Consequential validity refers to actual and potential outcomes of test use. It is a controversial concept in that where and when attention to the social and political ramifications of test use should be addressed is arguable. (Contains 24 references.) (SLD)

Running head: EMERGING/EVOLVING VIEWS OF VALIDITY

Emerging/Evolving Views of the Meaning of Score Validity

Terresa M. Humphries-Wadsworth

Texas A&M University

Paper presented at the annual meeting of the Southwestern Psychological Association, New Orleans, LA, April 11, 1998.

Abstract

The American Psychological Association [APA], in the late 1940's, began work to establish a code of ethics to include and address the needs of both scientific and applied members. Out of the ethics work emerged a set of standards for evaluating psychological tests. Four categories (or types) of validity were identified: content, predictive, concurrent, and construct. In subsequent years, predictive and concurrent were combined into a single category labeled criterion validity. The resulting three categories of validity, sometimes called the holy trinity, having survived nearly 40 years of use, are now entrenched concepts in test construction and evaluation. Current trends in the conceptualization of test validity dismiss these three categories as separate entities, but old habits die hard, as apparently do old ideas. The present paper reviews the emergence of validity as a unitarian concept, and discusses current views with particular attention to consequential validity.

Emerging/Evolving Views of the Meaning of Score Validity

After nearly 40 years of attempts to define validity, the definition remains elusive. Early efforts to define validity, were fumbling attempts to make an abstract concept concrete. Nevertheless, the early guidelines provided a foundation upon which the current structure of validity is built. The present paper reviews the emergence of validity concepts and discusses current views of a new addition to validity conceptualization, namely consequential validity.

During the 1930's the discipline of psychology was fraught with dissention and political conflict. Psychologist practitioners were dissatisfied with their voice and representation in the American Psychological Association [APA], which at that time was dominated by academicians, and so formed their own Association for Applied Psychology.

Efforts to reunite the psychological community, beginning around 1943, ultimately resulted in a reorganized APA whose clearly stated objective was to advance psychology as a profession as well as a science (Benjamin, 1996). In order to solidify connections between the two divisions, establishing a single code of ethics was a crucial first task. Within the code of ethics, the focus upon testing adequacy was particularly challenging. Both camps wanted to reduce the misuse of tests, but had no coherent guidelines for defining acceptable practices. It took four years for a team of measurement specialists to produce the first set of guidelines (Cronbach, 1989).

In 1954, APA produced the first set of specifications for psychological tests, Technical Recommendations for Psychological Tests and Diagnostic Techniques (APA, 1954). This manual identified four types of validity: content, predictive, concurrent, and construct. Each of these

types of validity was proposed to answer different questions about how a test performs.

Content validity referred to how well the test items (or procedures) represent the universe of items (or procedures) for the subject matter in question. Content validity attempted to answer the question, "Are the test items truly representative of the pool of items that define the construct?"

Predictive validity endeavored to forecast future outcomes based upon current performance. This was accomplished through correlations between the current performance, and some future outcome measure (or criterion). The question here was simply, "Based on a person's performance on this test, what can we predict about his or her performance on X in the future?"

Concurrent validity proposed a closely related question, in that it related current performance on the specified measure with performance on another measure (or criterion). The question here was more clearly stated, "How well does performance on this test relate to performance on another test or measure completed at roughly the same time?" The primary procedure for evaluating concurrent validity was to administer both tests and measure the correlation between them.

Construct validity was considered to be useful when one "wishes to infer the degree to which the individual possesses some trait or quality... presumed to be reflected in the test performance" (APA, 1954, p. 13). The evaluation of construct validity focuses upon the theory that underlies the trait. The Technical Recommendations stated there were two steps in the construct validation procedure. The first was to specify what predictions could be made based on the construct theory regarding differences in scores. The second was to gather data to confirm (or deny?) these predictions.

Cronbach and Meehl (1955) further developed the concept of construct validity. They explained that "it is naive to inquire 'Is the test valid?' One does not validate a test, but only a principle for making inferences" (p. 297). Furthermore, as one moves further away from objective, directly observable measures, one walks into the gray, misty area of theory. It is in this theoretical area that construct validity is most crucial.

Cronbach and Meehl (1955) suggested that a 'nomological net' is necessary to direct the research and evaluation of the theoretical construct about which the inferences are being made. This net is, basically, the theory and includes extensive delineations of "laws" which are supposed to support the researcher's assertions regarding the construct. Cronbach (1989) acknowledged that in social sciences, this level of specification, or substantive proof, may be impossible. Nevertheless, the need for theory-grounded evaluation of the construct has remained an important component of construct validation.

The revised recommendations in the Standards for Educational and Psychological Test and Manuals (APA, 1966) established what has been described as the trinitarian view of validity, i.e., there are three overarching types of validity: content, criterion, and construct. Criterion validity combined predictive and concurrent validity into one category.

The third edition of the Standards, produced in 1974, upheld the trinitarian view of validity but changed from calling them "types" of validity to calling them "aspects" of validity, thus indicating that the three are all related and part of a larger entity, whose makeup can only be understood by examining its varying dimensions. Additionally, the questions to be answered regarding validity were reduced to two: "a) what can be inferred about what is being measured by the test? b) what can be inferred about other behavior?" (APA, 1974, p. 25).

These seemingly minor alterations actually signalled fundamental changes in the ways in which validity was conceptualized. No longer was validation an attempt to calibrate a test, but now was seen as a ongoing process of evaluation of the test scores, and the interpretations that can be made from them. Furthermore, the 1974 Standards stressed that validity is not a property that can be directly measured, but rather is inferred from the three "aspects" of validity. Finally, the Standards (1974) articulated the emerging contemporary conceptualization of validity by stating, "These aspects of validity can be discussed independently, but only for convenience. They are interrelated..." (p. 26). In 1980 Cronbach clarified the matter even further, noting that "All validation is one, and in a sense all is construct validation" (p. 99).

The 'all for one and one for all' battle cry among contemporary measurement theorists is appealing. This view clearly delineates the superordinate nature of construct validy and dictates that all lines of validity evidence ultimately point to the construct (and expand the understanding of that construct).

The 1985 version of the Standards (AERA/APA/NCME, 1985) attempted to underscore the unity of validity by changing the holy trinity names from content, criterion, and construct aspects of validity to content-, criterion-, and construct-related evidence. This attempt to emphasize the relatedness of the three lines of evidence was poorly handled. The decision to retain the familiar titles and move toward the unified view of validity only served to cloud the centrality issue, allowing some to dismiss (or miss altogether) the changed meanings of these validity concepts (Shepard, 1993). As Moss (1995) pointed out, "when construct validity is viewed as the basis for all validity research, it makes little sense to use the same term as one category of evidence" (p. 7). Many authors (e.g., Shepard, 1993; Messick, 1989a, 1989b) have

argued for discarding the trinitarian titles and establishing a new system, but no single scheme for evaluating validity has yet emerged to dominate our language and thinking.

Traditionally, construct validity (the evidence category) was an attempt to establish the boundaries of the construct and to gain a clearer understanding of the mechanisms at work. This could be accomplished through several means. Convergent and discriminant validity are two sources (or perhaps two views of one source) of establishing the boundaries of what defines the construct.

Convergent validity confirms what is known about the content of the construct through correlations with other tests that purport to measure the same construct. Moderately positive correlations yield evidence that the test measures approximately the same behavior or trait (Anastasi, 1988). Discriminant validity confirms what is not the content of the construct through correlations with other tests that are thought to measure differing constructs. Therefore, one would expect very low correlations or no correlation between the measures.

Campbell and Fiske (1959) combined these two methods into one approach called the multitrait-multimethod technique. This technique combines convergent and discriminant validity procedures and applies them in one fell swoop. The procedure is fairly straightforward. Take at least two differing traits (constructs) and measure them both through at least two differing methods. The highest squared correlations should be between scores measuring the same traits; the smallest squared correlations should be between scores involving both different traits and different methods. While some have voiced objections to this technique as "a rather rote exercise" (Gray, 1996), it continues to be widely utilized.

Factor analysis is yet another technique for examining construct validity (Thompson &

Daniel, 1996). For example, historically "construct validity has [even] been spoken of as... 'factorial validity'" (Nunnally, 1978, p. 111). Joy Guilford's article some 50 years ago is illustrative:

> Validity, in my opinion is of two kinds... The *factorial validity* of a test is given by its loadings in meaningful, common, reference factors. This is the kind of validity that is really meant when the question is asked "Does this test measure what it is supposed to measure?" A more pertinent question should be "What does this test measure?" The answer then should be in terms of factors and their loadings... I predict a time when any test author will be expected to present information regarding the factor composition of his [sic] tests. (Guilford, 1946, pp. 428, 437-438, emphasis added)

Similarly, Gorsuch (1983, p. 350) has noted that, "A prime use of factor analysis has been in the development of both the operational constructs for an area and the operational representatives for the theoretical constructs." In short, "factor analysis is intimately involved with questions of validity.... Factor analysis is at the heart of the measurement of psychological constructs" (Nunnally, 1978, pp. 112-113).

Factor analysis reduces a multitude of variables to a few common factors. Examination of the factors reveals a pattern (or patterns) of responses. Examining the pattern helps the researcher to better understand what elements "hang together" and reveals the nature of the construct (or constructs) at work within the test.

As one gains a better understanding of the construct, one may need to re-evaluate

previous notions about how the construct operates, and/or the composition of the construct. Clearly, the construct does not remain untouched by the validity evidence (Cronbach & Meehl, 1955). On the contrary, as more evidence accumulates, the understanding of the construct changes; views evolve regarding the ways in which the construct operates, under what conditions.

The changing malleable nature of the construct and the interpretations made from test scores has some authors to carefully consider the consequences of test use (e.g., Cronbach, 1988; Messick, 1975, 1989a, 1989b). Messick (1989a) in particular is a strong advocate for considering social consequences of test use, what he calls consequential validity.

Consequential validity refers to actual and potential outcomes of test use (Messick, 1989b). These outcomes may include value judgements, social implications, and political consequences. Messick (1975) stressed that the appropriateness of test interpretation and use is an ethical question and as such is laden with value judgements. Shepard (1993) agreed that to ignore the connections between value judgments and the ultimate interpretations of the test is to ignore potential sources of invalidity or plausible rival hypotheses which should be investigated.

However, some authors (e.g., Maguire, Hattie, & Haig, 1994; Wiley, 1991) argue that although social consequences of test use are extremely important, they are an issue to be examined separately from the construct-centered validation process. These authors argue that involving social and/or political value analysis within the test validation process would weigh down the concept of validity beyond its ability to stretch.

Consequential validity is a very controversial concept, not in the sense that one should or should not be cognizant of the social ramifications of test misinterpretation or misuse, but rather where and when this attention to social and political ramifications should be addressed. Moss

(1992) articulated a mediating position. She stated the consequences of test use should be included in the definition of validity, because the omission of this important reference may lead some to discount the potential dangers of misuse.

Through the forty years of development, the conceptualization of validity has continued to evolve. While no clear, singular definition is yet acceptable, there appears to be a consensus among measurement theorists that the construct is the center of all validity questions. The process of examining the construct continues to develop, as does the ethical concern for the use of tests (and the resulting interpretations), largely because in the social sciences a construct is a difficult thing to define.

References

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

American Psychological Association (1954). Technical recommendations for psychological tests and diagnostic techniques [Supplement issue]. Psychological Bulletin, 51(2, Pt. 2).

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

American Psychological Association (1974). Standards for educational and psychological tests and manuals (revised). Washington, DC: Author.

Anastasi, A. (1988). Psychological testing (6th ed). New York: Macmillan.

Benjamin, L. T. (1996). The founding of the American Psychologist: The professional journal that wasn't. American Psychologist, 51, 8-12.

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

Cronbach, L. J. (1980). Validity on parole: How we go straight. In W. B. Schrader (Ed.), New directions for testing and measurement: Measuring achievement, progress over a decade: No. 5 (pp. 99-108). San Francisco: Jossey-Bass.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer (Ed.), Test validity (3-17). Hillsdale, NJ: Erlbaum.

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

Cronbach, L. J., & Meehl, P. E. (1954). Construct validity in psychological tests. Psychological Bulletin, 52, 281-302.

Gorsuch, R. L. (1983). Factor analysis (2nd ed.). Hillsdale, NJ: Erlbaum.

Gray, B. T. (1996, January). Controversies regarding the nature of score validity: Still crazy after all these years. Paper presented at the meeting of the Southwest Educational Research Association, Austin, TX.

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

Maguire, T., Hattie, J., & Haig, B. (1994). Construct validity and achievement assessment. Alberta Journal of Educational Research, 40, 109-126.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 35, 1012-1027.

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

Messick, S. (1989b). Meaning and values in test validation: The science and ethics of assessment. Educational Researcher, 18(2), 5-11.

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

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

Nunnally, J. C. (1978). Psychometric theory (2nd ed.). New York: McGraw-Hill.

Shepard, L. A. (1993). Evaluating test validity. In Review of Research in Education, 19, 405-450.

Thompson, B., & Daniel, L. G. (1996). Factor analytic evidence for the construct validity of scores: An historical overview and some guidelines. Educational and Psychological Measurement, 56, 213-224.

Wiley, D. E. (1991). Test validity and invalidity reconsidered. In R. E. Snow & D. E. Wiley (Eds.), Improving inquiry in the social sciences: A volume in honor of Lee J. Cronbach. Hillsdale, NJ: Erlbaum.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: EMERGING/EVOLVING VIEWS OF THE MEANING OF SCORE VALIDITY	
Author(s): TERRESA M. HUMPHRIES-WADSWORTH	
Corporate Source:	Publication Date: 4/11/98

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/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. 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 the identified document, please CHECK ONE of the following options and sign the release below.

☒ ⬅ Sample sticker to be affixed to document

Check here
Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

TERRESA M. HUMPHRIES-WADSWORTH

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document ➡ ☐

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

or here
Permitting reproduction in other than paper copy.

Level 2

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical 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."

Signature: Terresa M. Humphries-Wadsworth	Position: RES ASSOCIATE
Printed Name: TERRESA M. HUMPHRIES-WADSWORTH	Organization: TEXAS A&M UNIVERSITY
Address: TAMU DEPT EDUC PSYC COLLEGE STATION, TX 77843-4225	Telephone Number: (409) 845-1831
	Date: 3/12/98

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 this 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 which cannot be made available through EDRS).

Publisher/Distributor:	
Address:	
Price Per Copy:	Quantity Price:

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

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

Name and address of current copyright/reproduction rights holder:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

If you are making an unsolicited contribution to ERIC, you may return this form (and the document being contributed) to:

ERIC Facility
1301 Piccard Drive, Suite 300
Rockville, Maryland 20850-4305
Telephone: (301) 258-5500