sidered part of validation. Removing considerations of consequences from the domain of validity, however, would relegate them to lower priority. Validity is, after all, as stated in the 1985 *Standards*, the most important consideration in test evaluation.

Consequences of the uses and interpretations of test scores are central to an evaluation of those uses and interpretations. The evaluation of consequences rightly belongs in the domain of validity. The *Standards* is intended to serve as "a technical guide that can be used as the basis for evaluating testing practices" (AERA, APA, & NCME, 1985, p. 2). An evaluation that ignores the consequences of those practices would surely be inadequate. The best way of encouraging adequate consideration of major intended positive effects and plausible unintended negative effects of test use is to recognize the evaluation of such effects as a central aspect of test validation.

## References


American Educational Research Association. (1955). *Technical recommendations for achievement tests*. New York: National Education Association.

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

American Psychological Association. (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, and the National Council on Measurement in Education. (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

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

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

Linn, R. L. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher, 23*(9), 4–14.

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

Popham, W. J. (1997). Consequential validity: Right concern—wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9–13.

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

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5–8, 13, 24.


# Commentary

# The Consequences of Consequential Validity


William A. Mehrens
Michigan State University


In the good old days, when one wanted to indicate the adequacy with which behaviors elicited (or emitted) on an assessment were a reasonable sample of a domain of behaviors, there was a term which could be used—*content validity*. This term pertained to the adequacy of the sample as representative of a domain. Evidence regarding sample adequacy was called *content-validity evidence*. When the behaviors assessed were indirect measures of an attribute, and one wished to infer to an attribute not operationally defined, then the term *construct validity* was used to indicate the appropriateness of the inference. This term pertained to the adequacy of the behavior as a sign (an indicant) of some hypothetical construct. Evidence supporting the appropriateness of the inference was called *construct-validity evidence*. If an assessment, perhaps fortuitously, correlated with some other measure of interest, one could speak of the *criterion-*


*William A. Mehrens is a Professor at Michigan State University, 462 Erickson Hall, East Lansing, MI 48824. He specializes in educational measurement.*

*related validity evidence*. The correlation implied either a concurrent relationship or predictability, but not necessarily understanding or explanation in a scientific sense.

Thus, there were different words used for different kinds of evidence relating to different kinds of inferences. One could use these words to differentiate among the adequacy of the sample, the legitimacy of the sign, and/or the predictability of a different behavior. For example, a test of advanced algebra could be evaluated on whether (a) the behaviors to be elicited were a representative sample of the behaviors desired from someone who had mastered advanced algebra (content validity evidence), (b) the score from the test, for whatever reason, predicted success in calculus (criterion-related validity evidence), and/or (c) the test measured some hypothetical construct—perhaps important as a prerequisite for success in calculus (construct-validity evidence). The terms differentiated sample from sign inferences and predictability from scientific explanation inferences. Some individuals have promoted combining all such terms into one—suggesting that all validity is construct validity and that all evidence is evidence for (or against) construct validity. Such reductionist labeling blurs distinctions among types of inferences. Some people consider that progress.

Some individuals are promoting even more "progress" by suggesting that the validity of an assessment should be evaluated based on the consequences. That is, one should confound the results of using data in a decision-making process (which is what I think such individuals mean by consequential validity) with the accuracy of the inference about the amount of the characteristic an individual has (construct validity). *This confounding of inferences about measurement quality with treatment efficacy (or decision-making wisdom) seems unwise to me.*

Suppose a physician takes the oral temperature of an individual and finds it to be 105 degrees. The accuracy of the inference that the person has a fever is separable from the consequences of any treatment decision that may follow it. Or, suppose an adult male has an elevated PSA reading. The issue of whether this is an accurate indicant of prostate cancer is separable from the consequences of whatever treatment may follow. To confound them seems unwise.

The same reasoning is true in education. The accuracy of an inference about the amount of some characteristic an individual has is separable from the efficacy of any treatment (or the wisdom of any action). While one can call them both validity, it seems unwise to do so.

In the articles that stimulated this response, both Popham and Shepard make cogent and articulate points with respect to the issue of consequential validity. Shepard makes a well-reasoned plea for the centrality of test use and consequences for validity. As she correctly notes, the "debate is not whether consideration of consequences is worthwhile but whether it should be an integral part of validity theory and practice" (1997). She contends "that examination of effects following from test use is essential in evaluating test validity" (1997). Popham argues that "lumping our attention to the social consequences of test-use with the concept of validity will . . . muddy the validity waters for most educators" (1997). Both authors quote the 1985 *Standards for Educational and Psychological Testing* (American Psychological Association [APA], American Educational Research Association [AERA], & National Council on Measurement in Education [NCME]) to support their views.

In fact, the *Standards* do support both positions: "validity always refers to the degree to which . . . evidence supports the inferences that are made from the scores. The inferences regarding specific uses of a test are validated, not the test itself." (APA, AERA, & NCME, 1985, p. 9). The first sentence supports Popham's view, and the second supports Shepard's view. The problem with the second sentence is that an inference may not always be about a specific use. I may make a general inference—for example, that a test measures general intelligence or that it measures knowledge of the English alphabet. The inference that a test measures intelligence (recall that Snow, 1984, has said the concept of intelligence has as much scientific status as gravity) implies no specific use. It is just an inference about whether the assessment measures the construct. The meaning and usefulness of the construct of intelligence—and certainly the wisdom of any action based on inferences about level of intelligence—are separable from the issue of whether any particular test measures the construct—although to be sure, if the nomological net had no connections, the construct would not be very meaningful or useful. One can investigate the validity of the inference that a score is a reasonable indicator of the amount of a construct possessed independent of any specific use of the score. As Shepard has admitted, "It is possible to appraise the construct validity of a test interpretation without considering test use so long as no use is intended." (I submit the quote is also true eliminating the last seven words.) But Shepard argues that "As soon as a use is specified, then the validity investigation, including analysis of effects, must be tailored to the particular application." The issue is not whether to analyze effects of a particular application but whether to call that a validity investigation. The problem with tying consequences of a specific use to the notion of validity is that a test score therefore may be considered *invalid* for making an inference about a construct when, in fact, the inference would be accurate.

However, the issue gets muddied for two, somewhat related, reasons. First, it is true that a meaningful construct fits within a nomological net (or conceptual

network). If a hypothetical characteristic of individuals is totally independent of other characteristics or behaviors of those people, the characteristic has limited meaning. However, I can investigate the relationships of the construct within the network without calling it consequential validity. The consequences of a *particular use* do not necessarily inform us regarding either the meaning of a construct or the adequacy of a particular assessment process in measuring that construct. Indeed, the meaning of the construct and evidence that the test measures that construct may be well established prior to some specific use.

A second point that muddies the issue is that, at times, the test name (and indeed the specific purpose for constructing the test) may imply a particular purpose. Shepard's point (referencing Kane, 1992) about an algebra test needing additional evidence if it is to be used for differential placement is a good case in point. If the inference to be made from the algebra test is that it is a valid measure of algebra skills, then differential placement evidence is not needed—although given a hypothesis that such skill is necessary to succeed in a particular course, positive findings would be consonant both with the placement hypothesis and the inference that the test measures algebra skills. Only when the inference is made (or hypothesis is presented) that a certain level of algebra skills is a necessary prerequisite to success in calculus is the placement evidence needed. The argument is not about whether such evidence is desirable but whether the term validity (or consequential validity) should be used. Somehow, we need to use terms so that it is obvious that placement data are needed for placement inferences (uses) but that the accuracy of an inference made from an assessment result about the amount of a construct possessed does not necessarily depend on any particular use.

Words are powerful. How we use them is important. I am becoming convinced that the term *validity* and its accompanying terms of *construct validity* and *consequential validity* are being used with terminological inexactitude and that their meanings are overgeneralized. If validity is everything, then validity is nothing. Another oxymoron?

I suggest that the psychometric community *narrow* the use of the term validity rather than expand it. Let us reserve the term for determining the accuracy of inferences about (and understanding of) the characteristic being assessed, not the efficacy of actions following assessment. Of course, the consequences of decisions made (actions taken) based on test score data need to be examined. The bottom line in examining the results from an action is whether the positive consequences outweigh the negative consequences. But this examination of consequences may tell us more about the adequacy of the treatment or the general wisdom or social acceptability of the action than the validity of the inference about the construct. Further, consequences are often political value judgments, and these judgments may tell us nothing about the accuracy of the inferences about whether the assessment is a good measure of a construct.

Whatever the eventual outcome of the debate about the term consequential validity, it is evident that currently there is no agreement about the wisdom of its usage. For example, Maguire, Hattie, and Haig (1994) have suggested that "a concern with consequences should be moved out from the umbrella of construct validity and into the arena of informed social debate and formulated into ethical guidelines . . ."(p. 115). Tenopyr (1996) has stated that

> Conceptions of construct validity have changed somewhat over the years, but, vagaries of measurement aside, most psychologists have agreed that constructs basically pertain to living things. To expand construct validity to cover actions by the test user or others appears to be a misinterpretation of the common consensus of measurement experts. To carry this further and speak of "consequential validity" is a perversion of the scientific underpinnings of measurement. (p. 14)

I would hope that the authors of the revision to the *Standards* recognize the debate and relegate any discussion of consequences to a context separate from the validity chapter.

## Note

The author would like to thank Robert L. Brennan, Harvey Clarizio, Christine DeMars, Michael Kane, and Irvin Lehmann for their editorial comments on a previous draft of this article. Opinions expressed are those of the author and not necessarily those of the reviewers.

## References

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

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

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

Popham, W. J. (1997). Consequential validity: Right concern—wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9–13.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5–8, 13, 24.

Snow, R. E. (1984). Placing children in special education classes. Some comments. *Educational Researcher, 13*(3), 12–14.

Tenopyr, M. L. (1996, April). *Construct-consequences confusion*. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, San Diego.