The Centrality of Test Use and Consequences for Test Validity

Lorrie A. Shepard
University of Colorado at Boulder

What are the origins of consequential validity? What is the role of intended test use in validation? Is the study of unintended effects part of validation? What practical problems does this pose?

In his landmark treatise on validity theory in the third edition of *Educational Measurement*, Messick (1989) addressed the consequential basis of test use. As a short hand, *consequential validity* is now used by many to refer to the incorporation of testing consequences into validity investigations. Most measurement specialists acknowledge that issues of social justice and testing effects are useful ideas, but some dispute whether such issues should be addressed as part of test validity. They worry that addressing consequences will overburden the concept of validity or overwork test makers. Wiley (1991), for example, argued that understanding of "use errors is conceptually and socially important, but involves social and moral analyses beyond the scope of test validation as defined here and would needlessly complicate the conception and definition of test validity" (p. 88). Maguire, Hattie, and Haig (1994) also say that the consequences of test use are important, but they complain that Messick favors test use over test development and that his large-scale, systematic strategies for examining social consequences are too onerous. These demands only make sense "if testing is viewed as an industry like the pharmaceutical industry" (p. 113). Thus the more pointed version of this debate is not whether consideration of consequences is worthwhile but whether it should be an integral part of validity theory and practice.

By coining a new term, antagonists and advocates have created a false impression that a new kind of validity was invented in 1989 and appended to extant and more rigorous definitions of validity. In this article, I argue instead that consequences are a logical part of the evaluation of test use, which has been an accepted focus of validity for several decades. While it is true that our understandings of validity theory have evolved, consequences—especially those implied by construct definition and testing purpose—are not outside the underlying network of relationships that frame a validity investigation. My contention then is that examination of effects following from test use is essential in evaluating test validity.

An Old Idea: Test Validity Depends on Test Use

In the first validity chapter in the first edition of *Educational Measurement*, Cureton (1951) began:

> The essential question of test validity is how well a test does the job it is employed to do. The same test may be used for several different purposes, and its validity may be high for one, moderate for another, and low for a third. (p. 621)

Cureton went on to analyze how the meaning of what was being measured by the same test could vary depending on the prior experiences of groups taking the test. In the second edition of *Educational Measurement*, Cronbach (1971) wrote the chapter on validity. He too emphasized that validity is not in a test but must be evaluated anew for each testing application. Cronbach also included within the scope of validity studies evaluation of *decisions and actions* based on test scores as well as descriptive interpretations. His analysis of the types of evidence needed to support test-based decisions anticipated Messick by almost 2 decades:

> A decision is a choice between courses of action. The college admits or rejects a prospective student. The high school allocates an algebra student to a fast, average, or slow section. The primary school decides that one child should be taught to read immediately, and another should first practice on auditory and visual discriminations. The justification for any such decision is a prediction that the outcome will be more satisfactory under one course of action than another. Testing is intended to reduce the number of incorrect predictions and hence the number of decisions that will be regretted later. When validating a decision-making process, the concern is with the question: What is the payoff when decisions are made in the proposed way, and how does this compare with the

Lorrie A. Shepard is a Professor at the University of Colorado at Boulder, Campus Box 249, Boulder, CO 80309-0249. She specializes in educational measurement.

> payoff resulting when decisions are made without these data? (p. 448)

Once you include the soundness of test-based decisions as part of validity—not just descriptions or interpretations without actions—you are obliged to think about effects or consequences. Cronbach (1971) went on to elaborate the types of validation needed to support decision-oriented test use. In particular, he distinguished *validity for selection* from *validity for placement.* In the 1974 test standards, the concept of validity was expanded to refer to "the validity of particular interpretations or of particular types of decisions" (American Psychological Association [APA], American Educational Research Association [AERA], & National Council on Measurement in Education [NCME], 1974, p. 31). And by 1985, the test standards incorporated Cronbach's idea that evidence of aptitude-by-treatment interactions is needed to support placement decisions:

> In that context evidence is needed to judge the suitability of using a test for classifying or assigning a person to one job versus another or to one treatment versus another. It is possible for tests to be highly predictive of performance for different education programs or jobs without providing the information necessary to make a comparative judgment of the *efficacy* of assignment or treatment. (Emphasis added, APA, AERA, & NCME, 1985, p. 13)

A subsequent section of the 1985 standards appears to excuse test makers from verifying what relationships exist between test results and intervention effects so long as they make no claims about effects:

> Standard 8.11. Test users should not imply that empirical evidence exists for a relationship among particular test results, prescribed educational plans, and desired student outcomes unless such evidence is available. (APA, AERA, & NCME, 1985, p. 54)

It is clear from the ensuing commentary, however, that evidence verifying these linkages would be needed if test maker or test user wished to claim that the test was valid for planning interventions.

Thus the importance of attending to testing effects as a part of validity is not a new invention and was only made to seem a major departure because of Messick's use of a new term and a new set of conceptual categories.

The Mistake of Messick's Matrix

In his chapter, Messick (1989) argued that validity is a unitary concept. Although there are many sources of evidence, requiring both logical analysis and empirical data, such evidence must be brought together to form an integrated evaluative judgment. Messick presented his unified but *faceted* validity framework by means of the two-by-two matrix shown in Figure 1.

> A unified validity framework . . . may be constructed by distinguishing two interconnected facets of the unitary validity concept. One facet is the source of justification of the testing, being based on appraisal of either evidence or consequence. The other facet is the function or outcome of the testing, being either interpretation or use. (p. 20)

In my view, the matrix was a mistake. Although Messick goes on at some length to emphasize that the facets cannot be pulled apart and considered independently and to remind us that construct validity resides in all the cells, the temptation is too great to think that the traditional, scientific version of construct validity resides in the upper, left-hand cell and that consequences in the lower, right-hand cell are the business of moral philosophers and the politically correct. As stated previously (Shepard, 1993), my quarrel is not with Messick's ideas, only with the segmented presentation. Messick himself is clear that questions of test score meaning are necessarily value laden. So choices in the first cell—what to name a construct, how to represent it, what indicators and relations to include in a validity design—all involve value judgments, reflected on in the second cell but not carried out as a separate enterprise. The left-right division of the matrix is the more familiar distinction between test interpretation and test use. It is possible to appraise the construct validity of a test interpretation without considering test use so long as no use is intended. As soon as a use is specified, then the validity investigation, including analysis of effects, must be tailored to the particular application.

What Messick has contributed is a more thorough analysis of the kinds of thinking and data gathering required to support a validity conclusion. He has not invented a new kind of validity tacked on in the fourth cell.

	TEST INTERPRETATION	TEST USE
EVIDENTIAL BASIS	Construct validity	Construct validity + Relevance/utility
CONSEQUENTIAL BASIS	Value implications	Social consequences

FIGURE 1. *Messick's facets of validity framework*
(Note. From "Validity," by S. Messick, in Educational Measurement (3rd ed., p. 20) edited by R. L. Linn, 1989, New York: American Council on Education and Macmillan. Copyright 1989 by the American Council on Education and Macmillan. Reprinted by permission.)

Intended Consequences Are a Part of Construct Meaning

Originally, the conduct of construct validity studies was framed by a nomological net, which located the construct to be measured in a conceptual space showing its hypothesized connections to other constructs and observed behaviors. This understanding of construct validity, now taken to be the organizing framework for all of validity theory, is quintessentially the model of scientific theory testing. Although the term *nomological* is problematic for post-positivist science—"by some other name the organizing and interpretive power of something like a nomological net is still central to the conduct of validity investigations. Perhaps it should be called a conceptual network or a validity framework" (Shepard, 1993, p. 417).

As formulated by Cronbach and Meehl in 1955,[1] construct validation involved evaluating the construct validity of so-called criterion measures in the hypothesized network as well as the target measure, and it involved experimental studies as well as correlational evidence. Experiments were to test the stability of the construct across contexts and *whether experimental manipulations induced changes consistent with the theory*. Cronbach and Meehl also emphasized that for the test developer "both the test and the theory are under scrutiny"(p. 296) in these investigations. The test developer "is free to say *to himself privately*, 'If my test disagrees with the theory, so much the worse for the theory.' This way lies delusion, unless he continues his research using a better theory" (p. 296). Thus test developers were responsible for checking on the theorized relationships between test results and outcomes and were accountable for the validity of both the test and the explanatory theory.

Depending on test use, intended consequences of testing may or may not be a part of the relationships represented in the nomological net. Some uses are purely descriptive; they don't have consequences except that they help or hinder our understanding of phenomena (although even this could be called a consequence). But many test uses have explicit cause and effect relations as a part of the test rationale. This is true for the placement decisions already discussed, for prerequisite or minimum competency uses of tests, and for any test used to guide an intervention or make a differential diagnosis. There is a theory underlying these test uses, which connects test scores and outcomes, that must be investigated.

For example, construct validation of an IQ measure entails different requirements depending on whether it will be used solely for research—let's say to trace the heretability of learning disabilities—or to determine which children can benefit from instruction in the regular classroom. To support special education placements, tests must do a better job of distinguishing normal from abnormal, but there's more to it than locating the discriminating power of the test at the cut point. It is also essential that the intervention following from the test categorization have the theorized effect.

The same argument holds for Kane's (1992) analysis of the evidence needed to establish the validity of an algebra test used as a prerequisite for calculus. In addition to the usual conceptual and empirical verification that the test is a valid measure of the intended algebra skills, when used for differential placement, it is also critical to demonstrate that students with low scores on the placement test "do substantially better in the calculus course if they take the remedial course before taking the calculus course" (Kane, 1992, p. 531). The relationship in the nomological net between prerequisite algebra skills and later success in calculus is central to the meaning of the construct, not an add-on social nicety. The hypothesized consequences could be checked by means of a randomized experiment, comparing the calculus performance of low scorers with and without remediation. Of Figure 2 in his 1996 article, Popham is correct to say that a test carefully tied by logical and empirical evidence to the intended content domain is *valid* for reporting on the status or level of student achievement. But if he also claims that it can be used for placement decisions in middle school, more evidence is needed to establish the appropriateness of cut scores, predictive validity for subsequent performance, and verification of the assumed skill hierarchy.

Similar arguments can be made for the use of tests to target instruction in the classroom or to evaluate programs and monitor improvement. Achievement tests are just status measures, you say. True, only if they are not used to guide subsequent decisions. When school districts are promised increased funding if they raise their test scores, then the real learning consequences of instructional efforts that follow are at the very center of the validity inquiry. Cronbach and Meehl (1955) made this point 40 years ago regarding the coachability of mental tests, citing Gulliksen (1950) before them:

> When the coaching is of a sort that improves the pupil's intellectual functioning in schools, the test which is affected by the coaching has validity as a measure of intellectual functioning; if the coaching improves test taking but not school performance, the test which responds to the coaching has poor validity as a measure of this construct. (pp. 288–289)

In the case of coaching or teaching to the test, the threat to validity is not just that use of test results does not have the intended effect on learning; it is also that a flaw in the conceptualization of the test made it susceptible to invalid score gains that then render its use invalid. Often when we examine why the intended relationship between test and outcome did not hold up, we find that some narrowness in the content framework or limitations in item format implicitly narrowed representation

of the construct. In a validity investigation, we don't just express a personal preference for consequences that we like or dislike. Consequences are evaluated in terms of the intended construct meaning.

Even Side Effects Bring Us Back to Construct Meaning and Test Purpose

OK, so intended effects are clearly in the nomological net and the focus of validity studies. What about unintended effects? It is likely that some side effects can be found that are outside the confines of test score meaning. For example, several years ago medical college admissions officers worried that the MCAT was prompting pre-med students to take an excess of science courses to the detriment of their communication skills and broader knowledge in the humanities and social sciences. The consequence, as seen by deans and admissions officers, was that students being admitted to medical school were too narrowly educated. These consequences could hardly be brought under the tent of validity for judging the test battery as a measure of science knowledge, but, if the construct were redefined as preparation for professional practice, then representation of additional desired skills would clearly become part of the validity framework.

An important example of testing side effects that Messick (1989) cites is adverse impact. More boys are identified for special education than girls, more boys than girls are awarded college scholarships, more Whites than Blacks are selected for jobs. But neither Messick's consideration of consequences nor the literature on test bias automatically assumes that adverse impact is evidence of invalidity. The only claim that is made is that such questions deserve to be a central part of the validity investigation. According to Messick, if adverse impact stems from construct over- or underrepresentation, then it is evidence of test invalidity. If no such evidence is found, then the meaning of the construct itself (and the underlying theory) still warrant investigation.

Studies of side effects may not always be a part of initial test development, although consistently encountered side effects such as adverse impact should be routine. Once unanticipated effects are encountered, however, they should inform both subsequent test development efforts and validity investigations. For example, Popham (1994) has recently used evidence of consequences to reformulate how he believes that test specifications should be written. In high-stakes environments, the *single-definition* approach to operationalize the skill or knowledge domain being measured by a criterion-referenced test has, according to Popham, tended "to foster instructional methods that lead to nongeneralizable mastery on the part of students" (p. 30). This means they are invalid (my word, not his) but only in this context where they have this effect. Popham does not suggest that these single-definition specifications of basic skills items were intrinsically invalid, only that they became so when they led to bad instructional decisions that did not enhance student learning. So effects, unintended as well as intended, as they reflect back on test score meaning are a part of validity.

A Word About Misuse

Side effects, which are the unintended consequences of a test used for its intended purpose, should not be mistaken for the effects of test misuse. If a test were designed to be used to measure likelihood of success in college and is used instead as an outcome measure to evaluate the effectiveness of college teaching, this is an example of test misuse. Test makers are not responsible for negative consequences following from test misuse, nor should such effects be folded into the validity evaluation. The lively electronic exchange between Michael Scriven and others on the American Educational Research Association D list on the topic of consequential validity has sometimes conflated the question of who is responsible for the consequences of improper use with the question of what to include when evaluating validity for the test's intended use.

For each test use, there is a validity framework or conceptual network that includes intended effects and likely side effects or side effects after they have been discovered. When an existing test is adopted for a new purpose, a fresh validity evaluation is required, although some existing data may be relevant. For example, it is a new use when a college selection test, normally used in conjunction with high school grades and other sources of information, is used instead to screen applicants for scholarships. When users appropriate tests for purposes not sanctioned and studied by the test developers, users become responsible for conducting the needed validity investigation.

Concluding Remarks

My emphasis in this article has been on the continuity between this notion of consequential validity and long-standing principles of validity theory. This does not mean I don't think that validity theory has changed and is changing—otherwise, we would still think that "a test is valid for anything with which it correlates" (Guilford, 1946, p. 429) or, as was said at one time, "Intelligence is what IQ tests measure." Rather my purpose is to keep consequences of testing on the table and to direct our professional debates toward more thoughtful analyses of the implications and limitations of these ideas. When are consequences a part of the nomological net, and when are they the purview only of policymakers and politicians? What did Messick mean when he distinguished test validity and ethics but said they were entwined? When are we asking questions that are outside of the

(Continued on page 13)

to regard the intended and unintended effects of test-use.

In a recent essay on validity in the context of performance assessment, Messick argues for attending to "evidence and rationales for evaluating the intended and unintended consequences of score interpretation and use in both the short- and long-term, especially those associated with bias in scoring and interpretation or with unfairness in test use" (Messick, 1995, p. 7). Messick is identifying the key points to consider when looking at the social consequences of a test's use. I agree with those points, except where he would wish us to refer to such considerations as an aspect of validity. I want to keep them separate.

Lumping our attention to the social consequences of test use with the concept of validity will not only muddy the validity waters for most educators, it may actually lead to less attention to the intended and unintended consequences of test use. Those consequences will be so masked by the subsuming and confusing framework of validity that they are likely to be overlooked. Such an unintended social consequence of Messick's reformulated validity framework would, of course, be genuinely unfortunate.

A Rapid Reprise

In review, I've attempted to repudiate the idea that the social consequences of test use should be regarded as a "facet," "aspect," or "dimension" of measurement validity. The social consequences of test use are vitally important. Indeed, that's why most of us began working with tests in the first place—to bring about worthwhile social consequences. But social consequences of test use should not be confused with the validity of interpretations based on examinees' performances.

My argument was based on the following three points:

1. The current 1985 *Standards* view of validity (as the accuracy of score-based inferences), because it is both clear and useful, needs to be more widely adopted by educational practitioners.
2. The attempt to make the social consequences of test use an aspect of validity introduces unnecessary confusion into what is meant by measurement validity.
3. The social consequences of test use should be addressed by test developers and test users, but the assembly of evidence regarding test-use consequences can be accomplished without considering such evidence to be a facet of validity.

As the title of this analysis suggests, I believe that the proponents of consequential validity are striving for something good. Their concern is on target. Their mistake, I believe, is in trying to tie social consequences into a validity framework. Such a wedding of related but distinctive concepts will not be symbiotic, it will be septic.

Note

This article is based on a presentation at the Annual Meeting of the American Educational Research Association in New York, April 8–12, 1996. I wish to thank William A. Mehrens for reacting to an early version of this analysis.

References

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

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

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

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–694). Washington, DC: American Council on Education.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15*, 1–16.

Maguire, T., Hattie, J., & Brian, H. (1994). Construct validity and achievement assessment. *The Alberta Journal of Educational Research, XL*(2), 109–126.

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

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5–8.

Moss, P. A. (1995). Themes and variations in validity theory. *Educational Measurement: Issues and Practice, 14*(2), 5–13.

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

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* (pp. 75–107). Hillsdale, NJ: Erlbaum.

Centrality of Test Use

(Continued from page 8)

traditional validity paradigm, and when are we merely seeing with new eyes the implications of validity principles that have guided our science for decades? Are there avenues of inquiry that we agree are outside historical definitions of validity but that now should be included within the scope of validity theory?

In addition to these conceptual arguments, attending to consequences also presents practical problems. Questions such as Which uses must be investigated? How soon must unintended consequences be folded into the appraisal of test effects? Who is responsible: the test maker or the test user? are difficult issues but should not be confused with the warrant for including consequences in validity studies. Kane (1992) and Shepard (1993) have suggested ways to use an argument-based approach to prioritize validity questions and thereby reduce the burden of validity studies. I, for one, would not hold test publishers responsible for all possible test uses. Makers of standardized tests are not responsible for the effect of scores on the real-estate market, for example. But they are responsible for the uses that they advertise and that are closely implied by

(Continued on page 24)