Consequential Validity: Right Concern—Wrong Concept

W. James Popham
University of California, Los Angeles
and
IOX Assessment Associates

Why is the 1985 view of validity in the Standards *clear and useful? Does the issue of social consequences confuse the meaning of validity? Can we address test use consequences without making them a facet of validity?*

In recent years, those who frequent the literature of educational measurement have encountered, with ever-increasing frequency it seems, the expression *consequential validity*. Acceptance of this concept has been fostered chiefly by the attention given to the consequences of test use in Samuel Messick's influential validity chapter in the third edition of *Educational Measurement* (1989).

Although every right-thinking measurement person ought to be concerned about the consequences ensuing from a test's use, it does not follow that test-use consequences need to be linked to the now widely held view that *validity* is rooted in the accuracy of inferences we derive from examinees' test performances. I believe, therefore, that the reification of consequential validity is apt to be counterproductive. It will deflect us from the clarity we need when judging tests *and* the consequences of test-use.

A 3-Point Argument

In the following paragraphs, I intend to argue against the concept of consequential validity. My argument will be based on three contentions:

1. The current *Standards of Validity and the Validity of Standards in Performance Assessment* (Messick, 1995) view of validity, because it is both clear and useful, and should not only be endorsed but widely promulgated. The concept of validity embodied in the current AERA-APA-NCME *Standards for Educational and Psychological Testing* (1985) focuses educators' attention properly on the accuracy of test-based inferences. It should be better understood and more widely used by educational practitioners.
2. Cluttering the concept of validity with social consequences will lead to confusion, not clarity. Currently, many educational practitioners do not understand that validity refers to inferences, not tests. If the concept of validity becomes confounded with the notion that it must also address the social consequences of a test's use, most educators will be unnecessarily confused.
3. Test-use consequences should be systematically addressed by those who develop and utilize tests, but not as an aspect of validity. Because the social consequences of a test's use are so important, they must be carefully considered. The measurement community should insist, therefore, that developers and distributors of important educational tests assemble evidence regarding the intended and potential consequences of a test's use.

Now I'd like to support each of these three contentions. I recognize that I'm taking a position not embraced by some of our field's heaviest hitters. They are measurement people whom, without exception, I respect. Indeed, one is opposing the heart of the batting order when tussling with folks such as Messick (1989, 1995), Linn (1993), Shepard (1993), and Moss (1995). But even heavy hitters occasionally strike out. And I'm convinced that those who have, with laudable intent, hopped aboard the consequential validity bandwagon are, unfortunately, headed in a dysfunctional direction.

The advocates of consequential validity are operating on the basis of well-warranted concerns about the past and potential misuses of educational tests. These concerns have led them to cram social consequences where they don't go—namely, in determining whether a test-based inference about an examinee's status is valid.

The 1985 *Standards'* Conception of Validity

Because I've been involved with educational measurement since the time the earliest versions of the *Standards* were available in the mid-1950s, and have observed the way that different versions of the *Standards* dealt with validity, I regard the 1985 *Standards* treatment of va-

W. James Popham is Director of IOX Assessment Associates, 5301 Beethoven St., Suite 208, Los Angeles, CA 90066-7061. He specializes in measurement and evaluation.

lidity to be a striking advance over earlier conceptualizations of validity. For one thing, the authors of the 1985 *Standards* (Messick, 1985) don't equivocate regarding the paramount significance of validity. Their treatment of validity starts off with the assertion that "validity is the most important consideration in test evaluation." The 1985 *Standards* then go on to make clear that validity refers to test-based inferences, not the test itself:

> A variety of inferences may be made from scores produced by a given test, and there are many ways of accumulating evidence to support any inference. Validity, however, is a unitary concept. Although evidence may be accumulated in many ways, *validity always refers to the degree to which that evidence supports the inferences that are made from the scores.* (AERA, APA, NCME, 1985, p. 9, italics added)

The 1985 *Standards* assert that various types of evidence—namely, construct-related evidence, content-related evidence, and criterion-related evidence—can be brought to bear when someone is making a judgment about the validity of a score-based inference. Because, in the field of education, we are typically trying to get an accurate estimate of a student's status with respect to desired instructional aims such as the student's mastery of certain skills or domains of knowledge, an educator's instructional decisions are obviously apt to be unsound if that educator's score-based inferences about student status are invalid.

Yet, even though the 1985 *Standards* set forth a lucid definition of validity, most of our teachers and school administrators have, at best, only a murky notion of what validity really is. Having spent the past decade or so in trying to sharpen teachers' classroom assessment skills, I have encountered relatively few American educators who understand that validity depends on the accuracy of score-based inferences and is not a property of a test itself.

When teachers finally grasp the significance of validity-focused-on-inferences rather than validity-focused-on-tests, they really begin to understand that assessment instruments do not invariably, or even frequently, yield results leading to incontestably accurate inferences. A test is only a measuring instrument—an instrument far less precise than most practitioners believe. Such instruments should be used to arrive at inferences about examinees' status with respect to the domain of knowledge, skills, or affect represented by the test. And, because educational tests do not represent with unflawed perfection those domains, the resultant score-based inference will often be less than completely accurate. But because educational tests typically yield *numerical* results, and because human beings usually ascribe excessive accuracy to numbers, many educators regard the results of educational tests with unwarranted deference. If educators think that the measuring instrument is valid, they'll also tend to regard its numerical results as "valid"—that is, as accurate. The 1985 *Standards*' emphasis on the validity of inferences makes it clear that the validation of inferences depends on the *inferer's judgment* rather than on numbers. That view sends a solid signal to those many educational practitioners who currently kowtow to the numerical precision of assessment instruments.

Cluttering Instead of Clarifying

My second contention is, I believe, particularly important. I think that when the advocates of consequential validity attempt to make social consequences an integral part of validity, they do a disservice to clarity. They are asking a crisp concept—that is, *validity as the accuracy of score-based inferences*—to take on additional complexities. As a result of this excessive complicating, few educators will really understand what the essence of validity truly is. Note, for example, the multiple foci that Messick wants validity to embrace:

> Validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *interpretations* and *actions* based on test scores and other modes of assessment. (Messick, 1995, p. 5)

Messick's conception of validity, the cornerpiece of his 1989 chapter about validity in *Educational Measurement*, goes way beyond the notion that validity refers to the accuracy of inferences based on the results of testing. Messick, of course, agrees with most writers, and with the 1985 *Standards*, that validity is not a property of the test itself. Rather, he says, it is a property of the meaning of the test scores. But such meaning, he contends, must be derived not only from test items and stimulus conditions but also from the examinees and the context of the assessment. Where Messick's conceptualization of validity opens the door for social consequences is when he asserts that "in particular, what needs to be valid is the meaning or interpretation of the scores as well as any implications for action that this meaning entails" (Messick, 1995, p. 5).

Messick wants the social consequences of a test's use to become an important part of his validity framework. As he says,

> what is needed is a way of cutting and combining 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 in testing applications, and *that formally brings consideration of value implications and social consequences into the validity framework.* (Messick, 1989, p. 20, emphasis added)

I think Messick's 1989 validity framework did, indeed, cut and combine evidence so that social consequences became a key facet of validity. It's just that the price to be paid for doing so is far too high. We will need to give up the fundamental clarity that the architects of the 1985 *Standards* carved out for us when they asserted that validity refers to score-based inferences. More than 25 years ago, Cronbach pointed out that "one validates, not a test, but an interpretation of data arising from a specified procedure" (Cronbach, 1971, p. 447). The 1985 *Standards* echoes his view.

There is, for most of us, solid satisfaction in encountering a simply stated, compelling truth. When the educational measurement community arrived at its 1985 *Standards*

consensus that validity refers to the accuracy of score-based inferences, many of us were gratified. Here was a potent notion that could be communicated to our field's practitioners. Here was a way to get teachers and administrators to focus their attention on the adequacy of the *evidence* that was used to support a score-based inference about a student's status. But the clarity of validity-as-score-based-inferences is being threatened by those who would require the concept of validity to shoulder theoretical baggage that, though focused on an issue of importance, does not really bear on the accuracy of score-based inferences.

Let me use a simple example to illustrate this problem. Suppose that we want to make educational placement decisions about middle-school students in an affluent school district. Let's assume that a mathematics achievement test has been carefully developed over a 3-year period by district educators and external consultants. The mathematical knowledge and skills to be promoted by the district during Grades 6–8 are carefully delineated and then reviewed rigorously by a committee of district educators, parents, and independent mathematics consultants. Varied types of selected-response and constructed-response test items are employed in the test. The items are reviewed for content representativeness and potential bias, field-tested, revised, then reviewed and revised, again and again. *Everyone* who scrutinizes the test and the definitions of the content domains it was created to represent reaches the same conclusion. This is a test from which valid inferences can be drawn regarding the levels of a middle-school student's mathematics achievement.

Schematically, the mathematics test is depicted in Figure 1 by the rectangle at the right, while the content domain of mathematics skills and knowledge it represents is seen at the left. Based on efforts to produce varied forms of evidence regarding the validity of inferences about student mathematical status based on the students' test results, everyone involved concludes that the test allows the district's educators to make accurate inferences about students' mathematical achievement levels.

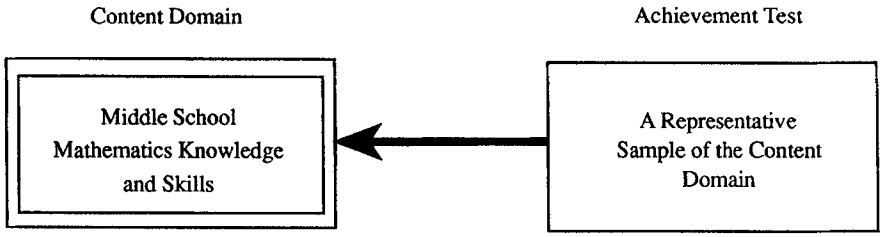

FIGURE 1. *An achievement test as a representative sample of a content domain in middle school mathematics*

Now, having established that the new test adequately represents the mathematics content domain, and assuming that there are no major contextual factors (such as insufficient testing time) that confound our inferences, it is reasonable to assume that the district's educators can use students' test scores to make valid inferences about the mathematics achievement levels of the district's middle-school students. With few exceptions, students who score well on the test will possess most of the mathematics skills and knowledge circumscribed by the content domain. With few exceptions, low-scoring students won't possess the knowledge and skills set forth in the content domain. Two such *valid* score-based inferences (Valid Inference A and Valid Inference B) are depicted in Figure 2. Clearly, there may be exceptions to the validity of our score-based inferences in the case of particular students. Joselyn Jones may have had the flu on the day of the test and, therefore, might have scored lower than she otherwise would have. Billy Barton may have, during the evening prior to the test's administration, used his mail-order, pick-a-lock kit to gain access to the teacher's locked desk, memorized most of the answer key before the test was administered, and hence scored higher than he really should have. But, because of the care with which the test was developed, valid score-based inferences will generally be made from almost all students' performances. And, because of those valid inferences, students will be assigned to the appropriate computer-delivered mathematics programs.

Now, let's suppose in our imaginary affluent school district that a new, strong-willed school board is elected. For the sake of this illustration, imagine that the new board members enact some genuinely bizarre policies regarding the use of students' scores on the district's new mathematics test. For example, the board requires that any girl who scores below the overall median will be (a) prohibited from engaging in any extracurricular activities, (b) prevented from taking any art or music courses, and (3) required to take at least two mathematics courses per semester. All boys who scored below the median would be expelled. If the boys scored really low, they would first be publicly beaten. Such absurd decisions, while

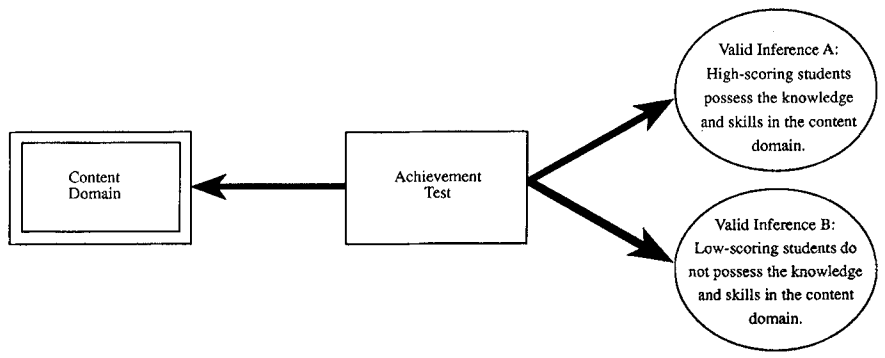

FIGURE 2. *Two valid inferences based on students' test results*

deplorable, do not alter one whit the validity of the test-based inferences about students' mathematics achievement. Those inferences are just as accurate as they were before the board made its ludicrous test-use decisions.

Messick and the other proponents of consequential validity will, I fear, make it impossible for run-of-the-mill educators to understand what measurement validity is really about. By asking the concept of validity to carry more conceptual freight than it should, advocates of consequential validity will have eroded the lucidity of what's meant by assessment validity. Wiley sums up this point nicely. He contends that:

> A valid set of measurements—defined in terms of realized intent—may be badly used . . . . The understanding of these 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. (1991, p. 88)

Shepard (1993) argues that, although considering the consequences of test use as an aspect of validity may seem new to many measurement specialists, it was implied by traditional conceptions of validity. She cites Cureton's (1951) comments in the first edition of *Educational Measurement* in which that author contended "the essential question of test validity is how well a test does the job it was employed to do" (p. 621). Yet, Shepard begins her own recent review of measurement validity with the indisputable assertion that "validity theory has evolved over time" (1993, p. 405). Messick's attention to the consequences of validity may, indeed, have been foreshadowed by Cureton and others, but in the evolving validity arena, some of those earlier foreshadows may have led to distortions. We need to defend our notions about validity with conceptual clarity, not historical precedent.

Why is it, then, that first-rate measurement thinkers such as Messick and others have urged us to add these new consequential trappings to what, in the 1985 *Standards*, was the straightforward idea that measurement validity referred to the accuracy of score-based inferences? As indicated earlier, I believe one of their motives was surely to draw our attention to the unsound uses of test results.

But they've added a social consequence attribute that fundamentally distorts the clarity of the notion of validity linked to score-based inferences. Social consequences are not *opposed* to the concept that validity revolves around the accuracy of score-based inferences. Rather, social consequences represent an important consideration, but one that is *orthogonal* to the 1985 *Standards* conception of validity. Almost all educational practitioners are well served by a more basic idea of validity—that is, the notion that validity refers to the accuracy of score-based inferences.

Addressing Test-Use Consequences

Without exception, advocates of consequential validity believe the social consequences of test use should be considered when judging whether a test's use is appropriate. So do I. But, whereas the proponents of consequential validity would intertwine that consideration with validity by using Messick's 1989 validity framework, I would keep social consequences separate.

Reliability, after all, is regarded by measurement specialists as an important consideration when evaluating tests. But that doesn't mean we need to make reliability a facet of validity. Attending to the social consequences of test-use should be an important task of test developers, test distributors, and test users. But such attention should be given *separately* from the assembly of evidence regarding the accuracy of score-based inferences.

Moss sets forth the choices regarding social consequences for the individuals currently revising the 1985 *Standards*:

> While there is little dispute about the significance of consequences, there are at least three distinguishable issues to face in revising the *Standards*: (a) whether or not the *Standards* should encourage or require assessment developers and users to consider evidence about consequences; (b) to what extent the consideration should address *actual* consequences, thus requiring evidence about the outcomes of assessment use, or *potential* consequences, thus requiring careful hypothesizing and use of existing research; and (c) whether that consideration of consequences should be viewed as an aspect of validity or as a distinct concept. (Moss, 1995, p. 10)

I'm not sure that I'd *require* assessment developers and users to collect evidence about consequences. Perhaps such evidence should be mandated. At a minimum, I'd like to see the assembly of social-consequence evidence strongly encouraged. And, as long as we're doing some encouraging, I think it appropriate to push for evidence regarding both actual and potential consequences. With respect to the last issue Moss raises—that is, whether a consideration of consequences should be viewed "as an aspect of validity or as a distinct concept"—it's pretty obvious where my sentiments lie.

It is increasingly apparent that Messick's conception of consequential validity is becoming a topic of debate among measurement professionals. See, for example, the essay by Maguire, Hattie, and Brian (1994) in which the authors argue that the stress on test consequences stems from substantially increased test-related litigation. They fear that "although following Messick's advice might reduce liability for commercial test developers in a litigious society, such efforts to avoid suits would probably direct energy away from the developers' primary responsibility to collect and report evidence on how the test and its scoring system relate to the underlying construct" (Maguire et al., 1994, p. 113). The new revisions of the *Standards* should reflect established views of the measurement field, not positions that are the foci of professional debate.

I want test developers and users to assemble as much evidence as is feasible about the likely or potential consequences of test use. And, after a test has been in use for some time, I'd like to see solid scrutiny of any unintended effects, positive or negative, of the test's use. Shepard (1993) offers some useful advice about how

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)