SAMUEL MESSICK

TEST VALIDITY: A MATTER OF CONSEQUENCE[1]

ABSTRACT. In this note I comment briefly on Keith Markus's illuminating article on "Science, measurement, and validity: Is completion of Samuel Messick's synthesis possible?" Markus's analysis bears directly on the controversial status of the consequential basis of test validity in relation to the more traditional evidential basis. After addressing some key points in his argument, I then comment more generally on sources of the controversy over the claim that empirical consequences of test interpretation and use constitute validity evidence.

SUBJECTIVE AND OBJECTIVE VALUES IN CONSTRUCTIONS OF REALITY

Markus aptly illuminates two syntheses outlined in my chapter on validity (Messick, 1989). Both syntheses were necessitated by the relativity of facts to values: One, called critical realism, is a synthesis of correspondence and coherence theories of truth; the other, called critical rationalism, is a synthesis of realist and constructivist interpretations. Markus emphasized the positive contribution of constructivism to the latter synthesis, arguing that an integration of contrary positive features is more valuable than a synthesis of positive and negative influences. Because constructivism is central to the whole enterprise of construct validity, I ultimately preferred to call this synthesis "constructive realism" (Messick, 1989: pp. 29–30).

In this constructive-realist view of psychological measurement, constructs represent our best, albeit imperfect and fallible, efforts to capture the essence of traits that have a reality in behavior independent of our attempt to characterize them. Just as on the realist side there may be traits operative in behavior for which no construct has yet been formulated, on the constructive side there are useful constructs having no counterpart in reality. These latter instrumental constructs are usually inductive summaries of data that serve as heuristic devices for organizing observed relationships with

Social Indicators Research **45:** 35–44, 1998.

no necessary presumption of real entities underlying them. Examples include higher-order constructs such as "ego" or "self", as well as useful classifications such as "working class" and "middle class" or "childhood" and "adolescence."

In this synthesis of realism and constructivism, theories can no longer be directly tested against facts because value-neutral data are problematic in the post-modern world. However, conjectures can still be tested against observations relative to specific social practices of science in the hope that this will lead to rational decisions about the theories. As one strategy for accomplishing this , I invoked Churchman's (1971) treatment of interacting systems of inquiry and Singer's (1959) insistence that the value implications of scientific models should be explicitly probed (Messick, 1989). It was hoped that the multiple perspectives of interacting inquiry systems and exposure of the value bases of scientific models would facilitate convergent and discriminant arguments penetrating the theory-laden and value-laden character of particular data.

In commenting on this strategy, Markus focuses on the possible meanings of Singer's point that "the reality of an inquiring system depends on its being 'observed' by another inquiring system" (Messick, 1989: p. 32). Markus belabors Singer's possible meanings to the point of bemusement, reminding me of one reason that I labeled the corresponding section of my validity chapter "Philosophical Conceits." Such deliberations aside, Singer's intention is clarified two sentences later: "Hence, a central feature of Singerian inquiry is that each type of inquiring system can be observed or *processed by the methodologies of the other types* in an effort to elucidate or disrupt the distinctive properties of the target inquiry process" (Messick, 1989: p. 32, italics added). For Singer, any inquiry system can be applied recursively to another system, including itself, thereby generating a complete matrix of inquiry systems processed by each other system, which Markus terms a "scientific panopticon."

Markus cogently conjectures that because a given processing system contributes to the manner in which the observed system is manifested, different processing systems will realize the observed system differently, suggesting that any single realization will generally be incomplete. He concludes that the incomplete nature of such one-on-one system processing means that it is not possible

for the scientific panopticon to guarantee rational decisions, which is not surprising since there are no guarantees in science. However, the incomplete nature of one-on-one system processing also suggests the potential value of recursively applying the alternative perspectives of multiple inquiry systems in an effort to flesh out the realization of the observed system.

Markus's major contribution is his insightful probing of the relation between the evidential basis and the consequential basis of test validity, which he rightly perceives as having been left as an incomplete synthesis in my validity chapter (Messick, 1989) or, rather, left as a tension to be negotiated in validation practice. Markus argues that a unified theory of validity implies validity in the singular, which he takes to mean a single validity for a test interpretation or test use, the unified emphasis being on combining various lines of evidence for this validity. In contrast to the singularity and unity of the evidential basis of test validity, the consequential basis recognizes a diversity of values, with the possibility that validity may vary with different value perspectives. If the consequential basis of test validity allows multiple validities as a function of a plurality of social values, how can this multiplicity be reconciled with the presumed singularity of evidential validity?

First, let us consider whether the unified theory of validity implies a single validity for a test interpretation or use. What is singular in the unified theory is the kind of validity: All validity is of one kind, namely, construct validity. Other so-called separate types of validity – whether labeled content validity, criterion-related validity, consequential validity, or whatever – cannot stand alone in validity arguments. Rather, these so-called validity types refer to complementary forms of evidence to be integrated into an overall judgment of construct validity. What needs to be valid are the inferences made about score meaning, namely, the score interpretation and its action implications for test use. Because value implications both derive from and contribute to score meaning, different value perspectives may lead to different score implications and hence to different validities of interpretation and use for the same scores. This suggests that validity might be indexed to values and perhaps contingent on different facts (or interpretations of facts) surrounding the conditions of use. Markus explored the possibility of elaborating

score meaning by means of a matrix of validities indexed to values and contingent on facts, but found this resolution to be less than wholly satisfactory from his perspective.

To facilitate the argument, Markus presents the incomplete synthesis of evidential and consequential bases of validity in the form of a Hegelian conflict: The thesis is that in the evidential basis, validity is value independent; the antithesis is that in the consequential basis, validity is value dependent. Things are not quite as neat as this, of course. In the evidential basis, value-neutral as well as theory-neutral facts are problematic, so validity is not really independent of values. Furthermore, in the consequential basis, validity is not only dependent on values but also on evidence of consequences and of antecedent conditions.

In any event, Markus argues, and I largely agree, that a complete synthesis is unlikely unless we relax the claim that values are prerational and subjective; instead, we should try to represent values as objective and justifiable. Indeed, I once heard Churchman remark that if ethical judgments can be justified in terms of situational conditions and the consequences of action, then ethics becomes a social science. If values are justifiable, they are akin to facts, and the tension between the evidential and consequential bases of validity is resolved.

The justification of values is no simple matter but nor is it out of the question. Theories of moral development from Piaget (1932) to Kohlberg (1964) and Loevinger (1966) have held that not all values are relative, that some values are more justifiable than others in adult behavior. The problem is that many important values are likely to remain relative to their community of believers or stakeholders. This combination of justified and relative values is reminiscent of the synthesis of realist and constructivist interpretations, wherein many constructs represent real entities while others are useful summaries of data relationships with no presumption of a real entity underlying them. Given a mixture of justified and relative values, Markus's valiant attempt at a synthesis is only partially complete. Under many circumstances, the relation between the evidential and consequential bases of validity remains a tension that must be carefully negotiated in the validation of test interpretation and use.

THE CONTROVERSY OVER CONSEQUENCES

Some measurement theorists and many practitioners seek to resolve the tension between the evidential and consequential basis of validity by simply eliminating consequences as a legitimate aspect of validity. Indeed, the elevation of construct validity to an overall validity framework for evaluating test interpretation and use is currently highly controversial, especially with respect to the role of consequences as validity evidence contributing to score meaning. The nub of the controversy centers on the superordinate role of construct validity as a generalized validity framework for evaluating score properties, even reliability, in relation to score meaning (Messick, 1989). Thus elaborated, construct validity becomes the unifying force that makes validity a coherent unitary concept and validation a unified process for evaluating evidence of the adequacy and appropriateness of interpretations and actions based on test scores. Or in the words of the 1985 test standards, validation addresses "the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores" (p. 9).

In order to talk about some of the nuances and intricacies inherent in a complex unified concept such as construct validity, I also articulated six features with complementary forms of evidence, namely, content, substantive, structural, generalizability, external, and consequential aspects of construct validity (Messick, 1995). All of these aspects need to be addressed in evaluating unified validity or else a compelling rationale is needed for why any one of them is forgone. Critics argue intensely that consequences should be forgone completely, mainly because their inclusion in the validity framework was ill-conceived in the first place.

Such intense controversy usually masks conflicts in values and ideologies. Although ideological disputes are too profound to explore here, I may be able to clarify some of the other sources of controversy, such as misconstruals, misinterpretations, and misattributions by one or another of the parties to the debate.

As an instance, some opponents of the use of consequences as validity evidence believe that advocates of such use are side-tracked by a misplaced concern over test misuse. These opponents argue that the consequences of procedural errors and unsound interpretations should not detract from the validity of legitimate uses of

a test. As Wiley (1991) states it: "The understanding of these *use errors* is conceptually and socially important, but involves social and moral analyses beyond the scope of test validation . . . and would needlessly complicate the conception and definition of test validity" (p. 89). Or as Popham (1997) argues:

> 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 (p. 9). . . . I believe one of their motives was surely to draw our attention to the unsound uses of test results (p. 12).

I, for one, had no such motive. On the contrary, I wanted to draw attention to unanticipated side-effects of *legitimate* test use, especially if unanticipated adverse effects are traceable to sources of test invalidity such as construct underrepresentation and construct-irrelevant difficulty. Of course, procedural errors and unsound interpretations would invalidate the particular local use, to be sure, but such *misuse* should not be incorporated into the validation process. Moreover, test makers are not responsible for the consequences of misuse; the responsibility in this regard clearly lies with the (mis)user (Cronbach, 1971; Messick, 1989; Shepard, 1997). However, the unanticipated consequences of legitimate score interpretation and use bear not only on the justification of the use but also on the soundness of score meaning and, hence, are an integral part of the validation process.

Another way to make this point is to invoke the concept of a nomological network from the Cronbach and Meehl (1955) manifesto on construct validity. Anticipated consequences of score interpretation and use, as well as predicted relationships with other measures and behaviors, form strands in the nomological network. Such a network provides a framework for deriving empirically testable consequences of construct theory and a foil for framing plausible rival hypotheses to challenge construct meaning. Its positivist label may seem outmoded these days, but by any other name the nomological network remains a fundamental aspect of all scientific inquiry. The point is that just as the anticipated and predicted consequences of test interpretation and use are strands in the construct's nomological network, so are any unanticipated con-

sequences. As strands of the nomological network, both anticipated and unanticipated consequences are clearly qualified as contributors to score meaning and as sources of evidence for construct validity. In contrast, the consequences of test misuse are irrelevant to the nomological network, to score meaning, and to the validation process.

By restricting their understanding of the consequential aspect of construct validity to the consequences of test misuse, the opponents of consequences as validity evidence find it easy to argue that evidence supporting the accuracy of score inferences about a person's current status on a construct is separable and orthogonal to the consequences of test misuse (Mehrens, 1997; Popham, 1997; Tenopyr, 1996). The argument is easy because it is basically true but immaterial. Furthermore, these critics of consequences want to restrict the meaning of construct validity to the accuracy of score inferences, acknowledging that the adverse consequences of misuse constitute a problem that needs to be addressed but not as part of the validation process.

Even on its own terms, this argument is deficient because validity refers not just to the accuracy of score inferences but also to evaluation of the appropriateness, meaningfulness, and usefulness of score inferences, which involves judgments not only of truth but of worth (Cronbach, 1988; Linn, 1997). Intrinsically, judgments of worth need to take into account the consequences of test interpretation and use. Note the emphasis here on the consequences of test interpretation as well as of test use, the former rarely being addressed by the critics of consequences with their major focus on the consequences of test use, indeed, of misuse. However, score interpretations have social consequences, too, which are fundamentally contributory to score meaning and hence to construct validity. These social consequences of score interpretation include the value implications of the construct label, of the broader theory in which the construct is embedded, and of the still broader ideologies that constrain and color theory construction (Messick, 1989). These value implications of the construct may or may not be commensurate with the construct's trait implications and need to be addressed in appraising score meaning.

The Achilles' heel of the opponents' position is that, by virtue of focusing on the consequences of test misuse, it completely misconstrues the meaning of the consequential aspect of construct validity, which refers to the unanticipated consequences of legitimate score interpretation and use. It is possible to separate score meaning as construct validity from the consequences of test misuse, as the critics contend, because the two are indeed orthogonal. However, it is not possible to separate score meaning from the consequences of legitimate interpretation and use because these consequences are an inherent part of score meaning.

To simplify matters a bit, I will now only consider unanticipated adverse side-effects of test interpretation and use. I address beneficial by-products, or what applied linguists call positive washback, elsewhere (Messick, 1996). All educational and psychological tests underrepresent their intended construct to some degree and all contain sources of irrelevant variance. The details of this underrepresentation and irrelevancy are typically unknown to the test maker or are minimized in test interpretation and use because they are deemed to be inconsequential. If noteworthy adverse consequences occur that are traceable to these two major sources of invalidity, however, then both score meaning and intended uses need to be modified to accommodate these findings. If these sources of underrepresentation and irrelevancy were known in advance, they would be taken into account in test development. There is no way that such unanticipated consequences of legitimate test interpretation and use can be considered to be irrelevant to the validation process.

Other critics (e.g., Brandon, 1996) maintain that the consequential basis of test validity is either circular or redundant, or both. They agree that adverse consequences signal invalidity only if traceable to sources of invalidity such as construct underrepresentation or construct-irrelevant variance. However, they argue that these sources of invalidity are identified and taken into account using the methods associated with Messick's other five validity aspects (content, substantive, structural, generalizability, external). Therefore, for these critics, consequences are unnecessary for examining validity.

However, even after diligently applying the methods of the other five validity aspects, every test still underrepresents its construct to some degree and contains sources of irrelevant variance, if for

no other reason than it is a test and not a criterion performance (Loevinger, 1957). Test makers usually maintain that this remaining invalidity is inconsequential until confronted with evidence to the contrary. This is precisely why unanticipated consequences constitute an important form of validity evidence. Unanticipated consequences signal that we may have been incomplete or off-target in test development and, hence, in test interpretation and test use.

To see that this is no minor point, just count the number of years that adverse consequences of multiple-choice testing were deemed benign until the possibility of deleterious effects on teaching and learning was highlighted (Frederiksen, 1984). Growing concern over construct underrepresentation and construct irrelevancy in multiple-choice tests then fueled the performance testing movement in education (Messick, 1994). If consequences are not part of the validation process, many sources of invalidity will remain unexposed to the detriment of the science and practice of educational and psychological measurement.

NOTE

[1] Portions of this paper were presented at the annual meeting of the American Psychological Association, Chicago, August, 1997 as part of an invited symposium of Division 5 Past Presidents: Looking Back, Looking Forward. I am pleased to thank Keith Markus for stimulating me to think in new ways about issues that have been with me for years. This stimulation was especially satisfying because it comes form a former student.

REFERENCES

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education: 1985, Standards for Educational and Psychological Testing (American Psychological Association, Washington, DC).

Brandon, P. R.: 1996, Discussion of the Consequential Aspect of Validity. AERA-D list on the Internet.

Cronbach, L. J.: 1971, 'Test validity', in R. L. Thorndike (ed.), Educational Measurement, 2nd ed. (American Council on Education, Washington, DC), pp. 430-507.

Cronbach, L. J. and P. E. Meehl: 1955, 'Construct validity in psychological tests', Psychological Bulletin 52, pp. 281–302.

Frederiksen, N.: 1984, 'The real test bias. Influences of testing on teaching and learning', American Psychologist 39(3), pp. 193–202.
Kohlberg, L.: 1964, 'Development of moral character and moral ideology', in M. Hoffman and L. W. Hoffman (eds.), Review of Child Development Research 1, pp. 383–431.
Linn, R. L.: 1997, 'Evaluating the validity of assessments: The consequences of use', Educational Measurement: Issues and Practice 16(2), pp. 14–16.
Loevinger, J.: 1957, 'Objective tests as instruments of psychological theory', Psychological Reports 3, pp. 635–694 (Monograph Supp. 9).
Loevinger, J.: 1966, 'The meaning and measurement of ego development', American Psychologist 21, pp. 195–206.
Mehrens, W. A.: 1997, 'The consequences of consequential validity', Educational Measurement: Issues and Practice 16(2), pp. 16–18.
Messick, S.: 1989, 'Validity', in R. L. Linn (ed.), Educational Measurement (Macmillan, New York), pp. 13–103.
Messick, S.: 1994, 'The interplay of evidence and consequences in the validation of performance assessments', Educational Researcher 23(2), pp. 13–23.
Messick, S.: 1996, 'Validity and washback in language testing', Language Testing 13(3), pp. 241–256.
Piaget, J.: 1932, 'The Moral Judgment of the Child (Kegan Paul, London).
Popham, W. J.: 1997, 'Consequential validity: Right concern – wrong concept', Educational Measurement: Issues and Practice 16(2), pp. 9–13.
Shepard, L. A.: 1997, 'The centrality of test use and consequences for test validity', Educational Measurement: Issues and Practice 16(2), pp. 5–8, 13, 24.
Tenopyr, M. L.: 1996, April, Construct-Consequences Confusion. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, San Diego.
Wiley, D. E.: 1991, 'Test validity and invalidity reconsidered', in R. E. Snow and D. E. Wiley (eds.), Improving Inquiry in the Social Sciences: A Volume in Honor of Lee J. Cronbach (Lawrence Erlbaum, Hillsdale, NJ), (pp. 75–107).

Educational Testing Service
Princeton, NJ 08541
USA