PHILOSOPHY OF EDUCATION 1995

Measurement by Tests and Consequences of Test Use

Stephen P. Norris
Memorial University of Newfoundland

Professor Howe challenges the consequentialists who currently have a dominant voice in educational testing and measurement. Consequentialists maintain that questions of measurement validity must take into account the social consequences of test use in judgments of validity. Howe agrees with the consequentialists on this broad issue, but disagrees with them on detail. He charges that the consequentialists are merely sophisticated technicists who desire a sharp separation between questions of validity and value judgments about the consequences of testing. Howe argues that, for the consequentialist, claims of the form "X use of testing is valid" are disguised hypotheticals of the form "X use of testing is valid if you endorse y," where y is some value commitment.

Howe's concern with such hypothetical claims of validity is that they are incapable of supplying "the answer to the pressing practical question of which among competing testing schemes ought to be endorsed and put into practice." He concludes that if validity judgments are to support testing practices, then the judgments must be categorical. By this he means that the judgments would incorporate substantive value judgments. In this way, value judgments would be "internal" rather than "external" to validity.

As Howe maintains, "whether value judgments are internal to or external to a conception of validity, there is no way to escape them." This claim is surely correct. However, I shall argue with the technicists that it is better to keep separate questions of test validity from questions of the social consequences of test use. Adopting the technicist approach leaves room for all of the judgments that Professor Howe wishes to make, while keeping separate issues that are handled better separately.

I shall proceed by outlining briefly the historical rise of consequentialism within testing theory. I shall then frame some challenges to this view, and to Professor Howe's, and invite his reply.

The Rise of Consequentialism

Earlier than about 50 years ago, the validity of educational and psychological tests was portrayed in terms of measurement: a test was valid if it measured what it purported to measure. The validity notion started to become more nuanced in the decade starting about 40 years ago. In the 1954 American Philosophical Association (APA) guidelines[1] (that "P" is for "Psychological"), for instance, the question was raised of what it is that is actually validated -- a test or the theory underlying it? A year later, Cronbach and Meehl announced that the inquiry, "Is this test valid?"[2] was naive, meaning by this injunction that tests were not validated, but rather "principles for making inferences." By 1960, Cronbach moved even further from the original notion that spoke of valid tests as those that measured properly. He declared that "it is incorrect to ask 'Is this test valid?' since any test is valid for certain purposes and not for others."[3] By 1966, the official APA stance was that "It is incorrect to use the unqualified phrase "the validity of the test,"[4] and three years later Cronbach concluded that "it is illogical to speak of test validity."[5]

This movement away from the idea that a test is valid when it measures what it purports to measure paved the way for the consequentialist view to emerge. By 1975, the timing was right for Messick to

draw endorsement from the measurement field when he quoted Cronbach approvingly: "One validates, not a test, but an interpretation of data arising from a specified procedure."[6]

Over the two decades leading up to 1975, validity claims shifted from being general endorsements of tests to endorsements of interpretations of data gathered in specific ways. This shift created an ambiguity in what was being judged. Was the judgment a *general* interpretation, "in general data gathered in this way means this"; or a *specific* interpretation, "this data that we have gathered means this." To the extent that the focus had shifted to interpretations of specific data, then concern had shifted from answering the general question, "Does this test measure what it is purported to measure?" Also, to the extent that the focus shifted to interpretations of specific data, validation questions were turned from issues of the quality of tests to issues of *testing*, of tests in use. Having turned to issues of testing, it was a small step to focus on issues of the consequences of testing. This is not to say that issues of the consequences of testing cannot deal with generalities; it is to conjecture that the ambiguity over whether validity judgments were general or specific interpretations opened the door to focus on the specific instead of the general, and to focus on consequences, since it is in specific circumstances that consequences are best seen and most felt.

A Plea for the "Old Way"

I shall urge that it is important to maintain a distinction between what a test measures and the consequences of its use. Furthermore, I shall urge that questions of what a test measures should be considered as questions of the validity of the test; questions of the consequences of the use of a test should not be considered questions of test validity. I shall make several points.

First, sometimes a test can be used for a purpose other than to measure what it is designed to measure; the consequences of such a use typically have no bearing on what the test measures, and thus no bearing on test validity as traditionally conceived. Suppose I am conducting an experiment to determine the effect of an instructional approach on the improvement of children's reading. I cannot randomly assign children to treatment and control groups. Instead, I match the groups as best I can and then give all the children an IQ test, and use scores on that test to control for any differences between the groups that arise from the inability to randomize. In using the IQ test, I need not assume that it measures what it purports to measure. I use it because I know it correlates better than just about any other test with school reading achievement. Suppose I find in my experiment that the instructional approach works better for girls than for boys. What significance does this consequence have on what the IQ test measures? If reading theory predicted the difference between girls and boys, would the experimental result support the claim that the IQ test measures intelligence? If my reading theory predicted no difference, would the result bear against the claim that the IQ test measures intelligence? I suggest that the correct answers to both questions are negative. Whether or not the IQ test measures intelligence is not tested by this experiment.

Second, sometimes the consequences of a test use are not problematic, but the test does not measure what it purports to measure. Suppose there is a test that purports to measure students' ability to think critically about arguments. They are presented with an argument and asked to write an evaluative response. Based on responses to the test from her class, a teacher concludes that the students' ability in this area is low and makes a concerted effort to teach them what the test purports to measure. The consequence is that the students subsequently perform much better on similar tests.

However, suppose studies show that the test seriously underestimates students' critical thinking ability in the area. What it measures, in addition to ability, is students' dispositions to think critically about arguments. The teacher's instruction, which she designed to increase the students' ability, had the effect of increasing their dispositions to think critically, and it was the increased dispositions that led to better performance. The salutary use of the test does not change the fact that the test does not measure what it purports to measure, and hence is not valid in the traditional sense. However, it is important to know what the test measures because it helps to explain why the teacher's instruction had the effects it did.

Third, sometimes the consequences of test use are problematic, but the test measures what it is purported to measure. Suppose there is a test that is purported to measure sight vocabulary, that is, words that individuals can recognize immediately upon seeing them. Suppose, in addition, that based on some children's scores on the test, they are assigned to different reading groups. The consequence, let us hypothesize, is that both children who are much better readers than their sight vocabulary would indicate, and children who are much worse readers are assigned to inappropriate groups and receive inappropriate instruction. The inappropriate instruction occurs from improper interpretation of the test scores. Sight vocabulary is but a small part of most children's reading vocabulary. Many children can read words in context that they do not recognize by sight; so sight vocabulary underestimates their reading vocabulary. The inappropriate assignment to reading groups as a consequence of the test use nevertheless does not alter the claim that the test measures what it purports to measure. The problem is not with the validity of the test, but with how scores on the test were used. It is important to identify the source of the problem and to keep it in focus.

CONCLUSIONS

Consequentialism in testing arose through an evolutionary process that shifted questions of validity from what tests measure to how tests are used. However, we need to be able to endorse tests under standard conditions of use. If we cannot do this, then a test never can be endorsed at all, because its use cannot be predicted or controlled. We also want to preserve the notions of measuring a construct and of how well that is done. I urge that we preserve the notion of validity to endorse tests for measuring what they are purported to measure, and consider separately from measurement issues questions of the consequences of test use.

None of this is to claim that the consequences of a test's use cannot be brought to bear on claims of what a test measures. However, the two are not necessarily connected, as my examples were intended to show. Sometimes a test can be used for a purpose other than to measure what it is designed to measure, while its use has no bearing on claims of what the test measures. Sometimes the consequences of a test's use are problematic, but the test measures what it purports to measure. Sometimes the consequences of a test's use are not problematic, but the test does not measure what it purports to measure. Measurement by tests and consequences of test use are not unrelated, as Professor Howe urges. However, contrary to Professor Howe, I believe there is more to be lost than to be gained by using the concept of validity to cover judgments in both areas.

---

1. American Psychological Association, *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (Washington, DC: American Psychological Association, 1954).

2. Lee J. Cronbach and Paul E. Meehl, "Construct Validity in Psychological Tests," *Psychological Bulletin* 52, no. 4 (1955).

3. Lee J. Cronbach, "Validity," in *Encyclopedia of Educational Research*, ed. C.W. Harris (New York: Macmillan, 1960), 1551-55.

4. American Psychological Association, *Standards for Educational and Psychological Tests and Manuals* (Washington, DC: American Psychological Association, 1966).

5. Lee J. Cronbach, "Validation of Educational Measures," in *Proceedings of the 1969 Invitational Conference on Testing Problems* (Princeton, N.J.: Educational Testing Service, 1969).

6. Samuel Messick, "The Standard Problem," *The American Psychologist* 30 (1975): 955-66.