PHILOSOPHY OF EDUCATION 1995

Validity, Bias, and Justice in Educational Testing: The Limits of the Consequentialist Conception

Ken Howe
University of Colorado at Boulder

Educational measurement has been historically dominated by *technicists*. Technicists abstract questions of test validity and bias from social conditions, and maintain that everyone should play by the ground rules that they, the technical experts, set. This requires first, defining the primary problem as obtaining accurate measurements, and second, embracing a sharp distinction between scientific investigation and value judgments. When the game is played by these rules, investigating the validity of testing is put squarely in the category of scientific investigation, and is thereby insulated from broader questions about social conditions and their implications for testing.

Recently, *consequentialists* have come to the fore in educational measurement.[1] Consequentialists deny the sharp distinction between scientific investigation and value judgments embraced by technicists, and see questions of test validity and bias as unavoidably embedded in social conditions. Accordingly, they hold that educational testing practices must be evaluated in terms of the broad social consequences that result from their use.

Whatever its advantages over the technicist conception (some of which I discuss as my arguments unfold), the consequentialist conception contains a crucial ambiguity regarding the relationship between test validity and value judgments. In particular, it is open to a strong (or categorical) interpretation and a weak (or hypothetical) interpretation. Given a strong interpretation, value judgments are internal to validity, such that "X use of testing is valid" entails "X use of testing ought to be employed." Given a weak interpretation, value judgments are external to validity, such the "X use of testing is valid" entails "X use of testing ought to be employed" only relative to this or that value judgment.

In this paper I argue that because a consequentialist conception is committed to the weak interpretation, it does not distance itself as much from a technicist conception as its advocates seem to believe. In a bit of a digression, I also make a few observations about performance assessment, the latest measurement rage, in connection with its putative promise to foster equity in educational testing.

Some Observations on Validity and Bias

Whatever one's position on educational testing, the elimination of bias is a minimal requirement of its just use. Questions of test bias are closely related to questions of test validity. A test (test use) possesses validity if it measures what it purports to and invalidity if it does not. Bias is a kind of invalidity that arises relative to groups. In general, a test is biased against a particular group if it under-predicts their performance on the criterion of interest relative to some other group(s). Charges of bias typically arise when a given identifiable group scores low on a test relative to some other group(s).[2] For example, the SAT is charged with bias against women on the grounds that although women generally score lower than men, their scores correlate with higher college performance.[3]

The kind of bias described is conventionally termed *predictive*. (Later I distinguish this variety of bias from *criterion* bias.) Historically, two conceptions of predictive bias have predominated: "external" and "internal."[4] External conceptions construe the problem of bias in terms of how well a

test predicts the real world performance it is designed to measure. (The above example of college entrance examinations and women provides an illustration of this conception.) Internal conceptions construe the problem of bias in terms of characteristics internal to the test, particularly differences among items. For example, suppose a mathematics achievement test contains an item that makes use of knowledge about the layout of a football field and that women score poorly on this item both relative to men and relative to their overall performance on the test. If so, the item is biased. If, on the other hand, women score low on this item relative to men but not relative to their overall performance on the test (which entails that they score lower overall than men), then the item is not biased.

The proposal that justice may be achieved by employing internally unbiased tests may be dismissed as just so much technical hocus-pocus. Put simply, rendering tests internally unbiased fails to insure that performance of interest is being measured or can be accurately predicted. To take the previous example of college entrance examinations, it is altogether possible that women could consistently score low relative to men such that no particular item correlated poorly with the remainder of items. Given an internal conception, the examination is not biased against women. But it is quite sensible to ask whether the examination as a whole is biased against women relative to some relevant external criterion of performance, for example, college grades. If women did achieve grades at least as high as men (and they in fact do),[5] then, provided grades are an appropriate criterion of college performance, the examination would be clearly biased against women and would, accordingly, be an unjust means of making admissions decisions. This general problem with internal conceptions leads some measurement experts to conclude that internal bias analyses should be limited to serving as flags for use in test development and, in particular, should never supplant external analyses.[6]

The proposal that justice may be achieved by eliminating external bias is considerably more defensible than the proposal that it may be achieved by eliminating internal bias. For, unlike internally unbiased tests, which provide no guarantee of being appropriately related to performance criteria, externally unbiased tests are so related by definition.

But achieving a strong relationship between test performance and performance on external criteria is also insufficient to insure justice. In addition to the kind of predictive bias I have described so far, tests may also encounter the problem of criterion bias. In contrast to the question of predictive bias, which is whether groups members' promise is accurately assessed *given* some criterion of performance, the question of criterion bias is whether the criterion of performance is itself biased against groups, independent of how well it correlates with test scores.

Criterion bias may take two forms: across and within. Across-criteria bias occurs when test performance is too heavily emphasized relative to other qualifications. Over reliance on testing inflates the importance of criteria that it can accurately measure -- various academic talents and accomplishments, in particular -- to the point of viewing these criteria as all-purpose qualifications that should be negotiable currency in virtually any arena. In general, qualifications cannot be viewed apart from the scheme of social cooperation from which they gain their meaning. For example, assuming that more women should be in positions of authority in elementary education in order to help eliminate the sexist attitudes produced by the present disproportionate number of men who occupy these positions, women have a qualification that men lack for admission to graduate programs in educational administration and ultimately for administrative posts.[7] Assuming that minorities should be admitted to universities in order to help overcome racism and stereotyping, and to make available to university communities a diversity of life experiences and perspectives, they have a qualification that non-minorities lack. The failure to appreciate the full spectrum of qualifications associated with a given educational or employment opportunity often fuels the complaint that it is unjust to prefer women over men and minorities over non-minorities when the latter are "more qualified" (read: score higher on tests). The problem is exacerbated when positions become more scarce, for the solution is often the convenient (and mindless) one of simply "raising

standards" (for example, raising the cut scores on college admissions tests used to determine which applicants will be given further consideration).

Within-criterion bias is bias in its most insidious form. Performance criteria may be perfectly matched to the demands of a given domain of performance, including being appropriately weighted, but be defined so that advantages and disadvantages arise associated with various characteristics such as race, social class, and gender. To take a fanciful example, imagine a test for Grand Wizard of the KKK (the GWT). The GWT could be unbiased in the sense of being a perfectly accurate predictor of poor on-the-job performance by African Americans. Of course it is the performance itself, carrying out the duties of the Grand Wizard, that renders the GWT biased against African Americans. To take a more realistic example, if the curricula and pedagogy of the U.S. educational system are indeed heavily biased in favor of white males, as they are so frequently charged with being, then criteria of performance are *ipso facto* biased in favor of white males as well. Under these kinds of conditions, eliminating predictive bias from educational tests can do little to eliminate injustice. All it can do is improve predictions of who will perform well given the criteria of performance associated with those who have historically enjoyed advantages within unjust institutional arrangements.

The Consequentialist Conception of Validity

Proponents of a consequentialist theory of test validity embrace a much more expansive conception that avoids many of the problems that plague the technicist conception. Consequentialist theorists chide the reliance on a single correlation as evidence for validity, a practice that remains pervasive,[8] requiring instead that multiple sources of evidence be employed. In general, they require that both the intended uses of tests and their associated social consequences be included in the evaluation of a test's validity.[9]

Shepard, for instance, suggestively likens the difference between technicist and consequentialist conceptions of validity to the difference between settling for "truth in labeling" and demanding that testing also be "safe and effective."[10] As an illustration, she appeals to the practice of readiness testing. According to her, the validity of readiness testing cannot be established solely on the basis of its ability to predict who will do well or poorly. Rather, to be valid ("safe and effective"), its use should not serve to "hurt" students to whom it is applied, especially those who are deemed not ready for given educational experiences. This requires investigating the broad range of consequences that attend the use of readiness testing.

Consequentialist theorists thus dismiss the technicist notion that the evaluation of educational testing can be divorced from social consequences and implicit value commitments. So far, so good. Having come to this point, however, they demur, leaving open the question of precisely what criteria are to be used to determine when educational testing is just. What is more, and despite avowals to the contrary, they seem to embrace, however unwittingly, the technicist notion that scientific questions can be disentangled from value questions and pursued independently.

Shepard, who provides a history of the emerging dominance of the consequentialist perspective in the measurement community and who wholeheartedly endorses this development, is a case in point. The following paragraph is illustrative:

> In my view, validity investigations cannot resolve questions that are purely value choices (e.g., should students be given an academic curriculum versus being tracked into vocational and college preparation programs?). However, to the extent that contending constituencies make competing claims about what a test measures, about the nature of its relations to subsequent performance in school or on the job, or about the effects of testing, these value-laden questions are integral to a validity evaluation. For example, the question as to whether students are helped or hurt as a result of test-based remedial placement is amenable to scientific investigation.[11]

It would seem that Shepard owes us a further explanation of what decisions are to count as "purely value choices." Take the question of tracking students into vocational education, and consider the following empirical questions that are pertinent to answering it. Can tests accurately predict who should be in which track? Do those in the vocational track wind up with jobs? Are their jobs satisfying? Do they receive adequate salaries in these jobs? Do they forego aspects of the curriculum relevant to effective citizenship? And so forth. Presumably, and as she herself indeed suggests, these and similar questions are ones she would want to include in evaluating testing for the purposes of placement in vocational education. But, then, how does the original question about tracking involve "purely value choices?"

Shepard's subsequent example of test-based remedial placement does little to clarify matters. On the contrary, it is supposed to be an example of a value-laden question that is "amenable to scientific investigation." But it is difficult to see how it is any different from the "purely" value-laden issue of tracking. Indeed, there is no way to adequately evaluate test-based remedial placement without addressing the questions of what it is to "hurt" a student and when this might indeed be justified, questions that cannot be disentangled from (purely value-laden?) questions like: What is fair to other students? What builds character? When must students suffer deserved consequences? What practices are most cost-effective? and so forth.

Shepard's difficulty in breaking with the technicist way of separating science and values is directly linked to the question of social justice in her discussion of the use of a multiple criteria approach for admission to selective colleges. She defends the use of the SAT as an adjunct to other criteria, such as music or athletic talent, minority group status, and geographical distribution. (Here she is concerned with avoiding what I referred to earlier as across-criteria bias.) Contrasting scientific questions with value choices, she claims:

> At one level, examination of these selection practices might provoke a debate between different philosophical positions. Should decisions be guided by meritocratic or other theories of social justice….At a more technical level, [multiple criteria] can be defended "scientifically" given that academic predictors are both incomplete and fallible predictors of success….Therefore, [the] value perspective [associated with multiple criteria] holds that colleges can reasonably select among qualified applicants using criteria aimed at other goals such as increasing the diversity of perspectives represented among their students. This value choice cannot be resolved within the validity framework but should be made explicit and examined for consequences as part of the validity investigation.[12]

This passage is relevant to both forms of criterion bias, and it may be used to illustrate the limitations of Shepard's consequentialist theory in addressing them. Regarding across-criteria bias, Shepard holds that criteria other than test scores may be "reasonably" employed. However, in the absence of some more substantive value commitment (for example, increasing diversity), there is no way to determine whether, more than just being "reasonable," the criterion of race indeed ought to be included in selection decisions. Regarding within-criterion bias, in the absence of some more substantive value commitment (for example, to a multicultural curriculum and form of instruction) there is no way to determine whether the criteria of success are indeed biased against certain groups. It seems that we can eliminate neither form of criterion bias by relying solely on Shepard's "validity framework."

The Red Herring of Performance Assessment

Although not strictly within the logical flow of the argument to this point, performance assessment is well worth a slight digression, particularly the second of the following two claims made on its behalf: that it will (1) improve curricula and instruction and (2) foster greater equity in educational testing.

Improvement in curricula and instruction will allegedly follow as a result of exploiting the practice of teaching to the test.[13] The reasoning is roughly as follows: Teaching to the test is a given. ("You get what you assess." "You do not get what you do not assess."[14]) Thus, rendering the performance

required on tests as close as possible to the actual desired performance (for example, writing essays) will automatically drive curricula and instruction in the right directions. I shall not criticize this defense of performance assessment in any detail here, since it is tangential to questions of validity and bias. I simply observe that when uniform standards and assessments are imposed top-down, by the measurement experts, they smack of coercion and betray a condescending attitude toward teachers. They are a circuitous way of addressing the shortcomings of curricula and instruction that encourages the assessment tail to continue to wag the instructional dog.[15]

The second claim for performance assessment -- that it fosters greater equity -- is my primary interest here, and it is with respect to this claim that performance assessment is most clearly a red herring.

That performance assessment is potentially more predictively valid than traditional measures is unassailable. Having people run a 100 meter race is certainly a better way to determine their sprinting ability than having them take a multiple-choice test on the principles of running. And having students write essays is certainly a better way to determine their writing ability than a multiple-choice test on the principles of writing. But suppose, having developed such performance assessments to our satisfaction, we now have to weight them to determine which applicants to state university should be admitted. The problem is that no matter how predictively valid these assessments are individually, they provide no help with the weighting problem, the problem of across-criteria bias, for this requires a judgment of the relative value of each kind of performance. For a very similar reason, performance assessment also fails to provide any help in eliminating within-criterion bias. If the desired performance is itself biased against certain groups, then having more direct measures of it is beside the point. Worse, if, as is likely, the same patterns of differential performance among groups persist or get worse,[16] performance assessment may well exacerbate the problem of injustice by further hiding it behind a layer of the latest brand of technical veneer,[17] and by providing ammunition for those who would "blame the victim."

CONCLUSION

In the end, performance assessment provides no help whatsoever with the problem of criterion bias, and the consequentialist view winds up being simply a more sophisticated version of the technicist view. Although the issue of what constitutes an acceptable "scientific" evaluation is clearly more complex on the consequentialist view, it is nonetheless conceived in a way that conditions the scientific question of what uses of testing are valid on "value choices." Thus, the claim that "X use of testing is valid" is always a disguised hypothetical -- always of the form "X use of testing is valid if you endorse y" (where y is some value commitment, for example, meritocracy, utilitarianism, Rawlsianism, and so forth). So long as validity judgments remain hypothetical in this way, they will be incapable of determining the answer to the pressing practical question of which among competing testing schemes ought to be endorsed and put into practice. In this way, the consequentialist conception retains the fundamental bifurcation between test validity and value judgments -- and the "truth in labeling" criterion as well.

If validity findings are to justify testing practices by themselves, they ultimately must be categorical, must be of the form "X use of educational testing is valid, period."[18] A categorical conception of test validity would incorporate substantive value commitments into the very meaning of "validity," and such a conception has much to recommend it. Among other things, it can serve to direct validity research in morally defensible directions and to prevent measurement experts from hiding their implicit value commitments behind a cloak of scientific objectivity.[19] In any case, whether value judgments are internal to or external to a conception of validity, there is no way to escape them. Limited as it otherwise might be, this is the welcomed insight of the consequentialist's conception of validity.

1. See, for example, Samuel Messick, "Validity," in *Educational Measurement*, 3d ed., ed. Robert L. Linn (New York: American Council on Education and Macmillan, 1989), 13-103; and Lorrie Shepard, "Evaluating Test Validity," in *Review of Research in Education*, vol. 19, ed. Linda Darling-Hammond (Washington, D.C.: American Educational Research Association, 1993), 405-50.

2. A test could conceivably be biased against a group in the absence of differential test performance between the group in question and other groups, though I am not familiar with such cases.

3. See American Association of University Women Educational Foundation, *How Schools Shortchange Girls* (Annapolis Junction: Md., 1992).

4. Gregory Camilli and Lorrie Shepard, *Methods for Identifying Biased Test Items* (Thousand Oaks, Calif.: Sage, 1994)

5. American Association of University Women, *How Schools Shortchange Girls*.

6. Camilli and Shepard, *Methods for Identifying Biased Test Items*.

7. See, for example, Amy Gutmann, *Democratic Education* (Princeton, N.J.: Princeton University Press, 1987) and Ronald Dworkin, *Taking Rights Seriously* (Cambridge, Mass.: Harvard University Press, 1978).

8. See Shepard, "Evaluating Test Validity."

9. See, for example, Messick, "Validity," and Shepard, "Evaluating Test Validity."

10. Shepard, "Evaluating Test Validity."

11. Ibid., 429.

12. Ibid., 435.

13. See Lauren B. Resnick and Daniel P. Resnick, "Assessing the Thinking Curriculum: New Tools for Educational Reform," in *Changing Assessments: Alternative Views of Aptitude, Achievement and Instruction*, ed. Bernard R. Gifford and Mary Catherine O'Connor (Boston: Kluwer Academic Publishers, 1992), 37-75.

14. Resnick and Resnick, "Assessing the Thinking Curriculum: New Tools for Educational Reform," 59.

15. I am skeptical that assessment can drive curriculum and instruction without also shaping them to serve assessment. George Madaus, shares my skepticism in his "A Technological and Historical Consideration of Equity Issues Associated with Proposals to Change the Nation's Testing Policy," *Harvard Educational Review* 64, no. 1 (1994): 76-95. He employs the delightful example of the manner in which the invention of tomato picking machines led to the development of tough-skinned, less tasty, tomatoes.

16. See, for example, Michael Apple, "The Politics of Knowledge: Does a National Curriculum Make Sense? *Teachers College Record* 95, no. 2 (1993): 222-41; Georgia Earnest Garcia and P. David Pearson, "Assessment and Diversity," in *Review of Research in Education*, vol. 20, ed. Linda Darling-Hammond (Washington, D.C.: American Educational Research Association, 1994), 337-93; and Madaus, "A Technological and Historical Consideration of Equity Issues Associated with Proposals to Change the Nation's Testing Policy."

17. I find the suggestion quite baffling that providing students with different ways of demonstrating the same desired performance helps with the problem of equity, a suggestion made by, for example, Linda Darling-Hammond, "Performance Based Assessment and Educational Equity," *Harvard Educational Review* 64, no. 1 (1994): 5-30; and Robert Rothman, "Assessment Questions: Equity Answers, in *Evaluation comment: Proceedings of the 1993 CRESST Conference* (Los Angeles: UCLA's Center for the Study of Evaluation & The National Center for Research on Evaluation, Standards, and Student Testing, 1994), 1-12. If the performance (construct) is itself biased, then the means of measuring it are moot. Compare a multiple-choice version of the Grand Wizard Test with a portfolio version. Furthermore, it is odd to suggest that different kinds of performance assessments can zero in on the same performance (construct), or it is at least odd that we would want them to. Isn't one of the virtues of performance assessment its relatively direct tie to the richness and peculiarities of given kinds of activities?

18. The distinction between hypothetical and categorical validity judgments is inspired by Kant's distinction between hypothetical and categorical imperatives, which correspond, respectively, to prudential, means-ends ought statements (e.g., "If you want to be rich, then you ought to x" versus moral ought statements (e.g., You ought not x.). I do not mean to suggest by this usage that statements of the form "X use of testing is categorically valid" may not involve a good deal of

controversy. (It is perhaps worth noting here that the claim "The earth revolves about the sun," although once very controversial, was consistently advanced categorically.)

19. For a discussion of similar issues regarding the validity of educational research, see Kenneth R. Howe, "Two Dogmas of Educational Research," *Educational Researcher* 14, no. 8 (1985): 10-18; and Kenneth R. Howe and Margaret Eisenhart, "Standards for Qualitative (and Quantitative) Research: A Prologomenon," *Educational Researcher* 9, no. 4 (1990): 2-9.