

Fairness in Employment Testing

Constance Holden (News & Comment, 2 June, p. 1036) has written a useful summary of the issues addressed in the National Research Council's recent report on the General Aptitude Test Battery (GATB). However, she has seriously misrepresented the committee's position on adjusting the scores of black and Hispanic test-takers so that *able* nonwhite workers have the same chance of referral to jobs as *able* white workers.

As the figure reproduced in Holden's article illustrates, the direct use of test scores, without adjustments, will result in the false rejection of a larger proportion of able black and Hispanic workers than of able white workers (although some able workers in all groups will be erroneously rejected). This is not because the test is biased, as Holden says the report implies. The "false negative" effect is not a function of race or ethnicity. Rather, the disproportionate rejection of able minority workers is due to the interplay of two factors: the modest predictive accuracy of the test and the lower average test scores of these applicants. The combination of the two means that proportionately more black and Hispanic test-takers who could perform well on the job will be falsely predicted to be unsatisfactory.

Moreover, the committee does not believe, as a reader of the article might surmise, that the within-group percentile scoring system currently being used by the Labor Department's Employment Service is the only way—or in all circumstances an appropriate way—to ensure equal referral chances for *able* minority and white applicants. Our endorsement of the within-group percentile method is clearly linked to the current predictive power of the GATB. As long as the GATB predicts job performance with only modest accuracy (correlation, 0.3), scores based on group norms will achieve approximately equal referral rates for able white, black, and Hispanic workers.

A crucial point is that the size of the adjustment needed to effect the recommended outcome will necessarily depend on the accuracy with which job performance is predicted by the test. The attraction of the second scoring strategy specifically endorsed in the report (a so-called performance-based method in which test scores are adjusted by group so that the distribution of test scores at a given level of job performance is the same for all groups) is that it is responsive to changes in the predictive accuracy of a test. Highly accurate prediction would mean small score adjustments; at current levels of

accuracy the adjustments would be just about the same as those produced by the within-group percentile system.

Finally, Holden quotes an official from the Office of Personnel Management as saying that the vast bulk of research shows that pure rank-ordering of scores "is the only scientifically justified position." We disagree. It is indeed true that selection on "pure rank-ordering" will generate a work force with the highest expected productivity. But it is also true that able black and Hispanic workers will be rejected far more frequently by such a referral policy than whites at the same level of job performance. This is a scientific fact, demonstrated theoretically and empirically in the report.

JOHN HARTIGAN
Committee on the General Aptitude Test Battery, National Research Council, Washington, DC 20418 and Department of Statistics, Yale University, New Haven, CT 06520

ALEXANDRA WIGDOR
Committee on the General Aptitude Test Battery, National Research Council

Response: Wigdor and Hartigan basically raise two objections about my article.

1) They say that I say that they imply the GATB is racially biased. In fact, I made it clear that the committee did not find racial bias in the test. However, some statements could lead a rational person to infer that the test is unfair to minorities—for example, the authors assert that able blacks and Hispanics will be rejected by test scores "far more frequently" than whites "at the same level of job performance."

2) The authors disagree with an Office of Personnel Management official that pure rank-ordering of scores is the only "scientifically justified" position. But, as I indicated in my article, while within-group scoring may well be justified socially and politically, the scientific basis is questionable, for at least two reasons.

Since the purpose of the test is to maximize productivity by predicting worker performance, and since, as the authors acknowledge, pure rank-ordering produces a work-force of the "highest expected productivity," it is arguable that there is no scientific justification for tinkering with test scores that does not improve the validity of prediction.

Moreover, the committee has adopted a race-based solution for a problem that it says is not race-based. It might be argued that a more "scientific" solution to the fact that marginal scorers get more false negatives would be to adjust the scores of all low scorers as a group.

My article certainly may be construed as

being critical of the committee's reasoning, but I do not see where I have misrepresented their position.—CONSTANCE HOLDEN

Holden writes that the National Research Council committee concerned with job test scores and job performance states that it has "scientifically justified" ways of reporting scores to replace "pure rank-ordering." The "system for reinterpreting the ability test scores of blacks and Hispanics" depends in large part on the committee's distinction between "predictive fairness" and "performance fairness." The former entails predicting performance from test scores. The latter entails predicting test scores from performance, but it is used by the committee to support affirmative action hiring.

"Performance fairness"—which implies group equality in outcome of the selection procedure—does not represent a scientific basis for that purpose because it is "internally contradictory" (1). It lacks consistency in applications because there is a reversal in its effect when it is applied to a remedial program for low scorers as opposed to job referrals for high scorers. For example, if within-group scoring were used in determining eligibility for a Head Start program, "performance fairness" would favor whites.

Affirmative action programs for certain minorities rest on value judgments, not on educational and psychological data or on statistical finagling with test scores. Value judgments should be made explicitly and openly, not camouflaged by rhetoric or statistical legerdemain.

The very name "performance fairness" is rhetorical camouflage. The name suggests that tinkering with scores will result in equal performance. But in fact it will not. When the decision to select is made, the only information available on performance of either individuals or groups is from the imperfect selection instrument or instruments.

A related issue is that the committee's rationale can be extended without any empirical or technical qualification to tests and grades used in the selection of undergraduate, graduate and professional school students, and the hiring of professionals. A qualification that procedures suitable for working class occupations are not suitable for the learned professions is not acceptable in a democratic society.

LLOYD HUMPHREYS
Department of Psychology, University of Illinois, 603 East Daniel Street, Champaign, IL 61820

REFERENCES

1. N. D. Peterson and M. R. Novick, *J. Educ. Meas.* 10, 3 (1976).